Essays in the Economics of Local Labour Markets

by

Iain Gordon Snoddy

B.A., Trinity College Dublin, 2012

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Economics)

The University of British Columbia
(Vancouver)

October 2019

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, the dissertation entitled:

Essays in the Economics of Local Labour Markets

submitted by **Iain Gordon Snoddy** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** in **Economics**.

Examining Committee:

David Green, Economics
Supervisor

Thomas Lemieux, Economics
Supervisory Committee Member

Craig Riddell, Economics
Supervisory Committee Member

Paul Schrimpf, Economics
University Examiner

Werner Antweiler, Business
University Examiner

Arthur Sweetman, Economics
External Examiner

Abstract

This dissertation examines various aspects of the geographic segmentation of economic activity across local labour markets. In Chapter 1, I present a methodological improvement to traditional methods of selection bias correction. Selection bias plagues many empirical studies that exploit variation across regions or cities, due to the residential sorting of individuals. Using machine learning tools I develop a semi-parametric method which relaxes the strong assumptions and restrictions imposed by more traditional methods. Through numerical experiments I show that this method tends to outperform more traditional methods and more flexibly corrects for selection bias.

Using this improved methodology in Chapter 2, I correct for selection bias in estimating the returns to education at the State level using the 1990 US Census. I confirm the general pattern of an upward bias in the returns to a college education as found by previous studies, but my results depart significantly from those obtained using more traditional means of selection correction. My findings indicate that traditional methods overstate the upward bias in the returns to a college education, and conversely, understate the upward bias in the returns to an advanced degree. These results confirm the importance of using the improved methodology outlined in Chapter 1 of this dissertation.

In Chapter 3, I analyse the cross-city wage effects resulting from workers having the opportunity to find relatively high paying unionised jobs. I examine, and separately identify, two distinct channels of effects; the wage bargaining channel whereby workers with relatively high alternative employment opportunities can negotiate higher wages, and the wage emulation channel whereby firms pay higher wages to stave off unionisation. I build a novel model of union formation and wage setting which informs the empirical methodology employed. Specifically, I use Bartik style shift-share instruments to deal with issues related to endogeneity. I find evidence in favour of substantial spillover effects: average hourly earnings would be 2 percentage points higher had the composition of union work remained at its 1980 level. Furthermore, I find substantial heterogeneity across sub-populations in the role played by union spillovers in wage fluctuations.

Lay Summary

To talk about the national labour market is somewhat of a misnomer given the geographic fragmentation of economic activity across cities and regions. Each local labour market is distinct in its institutions, opportunities, and the skill profile of its workers. This dissertation examines the degree to which wages vary across local economies, why these earnings vary, and develops a novel methodology to aid researchers overcome biases that plague estimation of these effects. Specifically, using machine learning tools I present a new method to control for the bias introduced by the residential sorting of individuals across regions based on unobservable characteristics. Using this improved procedure I present new evidence of variation across regions in the return to education. Finally, I examine how the existence of higher paying unionised jobs in a city increases wages paid in non-unionised jobs.

Preface

This dissertation is original, unpublished work. Chapters 1 and 2 are the result of independent work by the author, Iain Gordon Snoddy, and chapter 3 was written in close collaboration with Professor David A. Green.

Table of Contents

Abstract iii

Lay Summary iv

Preface v

Table of Contents vi

List of Tables ix

List of Figures xi

Glossary xiii

Acknowledgments xiv

Dedication xvi

Introduction 1

1 Learning about Selection: An Improved Correction Procedure 4
- 1.1 Introduction 4
- 1.2 Roy model 8
  - 1.2.1 Model Overview 9
  - 1.2.2 The Outcome Equation 9
  - 1.2.3 Utility Maximization 10
  - 1.2.4 Selection Bias 11
  - 1.2.5 Issues in Estimation 12
- 1.3 Estimating the Earnings Equation 12
  - 1.3.1 Dahl's Approach 12
  - 1.3.2 An Improvement: Post-Double Lasso 17
  - 1.3.3 Identification 22
  - 1.3.4 Restricting the Control Function 23
- 1.4 Monte Carlo Experiment 25
  - 1.4.1 Model 25
  - 1.4.2 Implementation 26

1.4.3 Results . . . 26
1.5 Estimating Migration Probabilities . . . 39
1.5.1 Discussion of Random Forest . . . 40
1.5.2 Illustrative Example . . . 44
1.5.3 Discussion . . . 51
1.6 Conclusion . . . 51

2 Re-estimating the Returns to Education . . . 53
2.1 Introduction . . . 53
2.2 Data . . . 54
2.3 Estimating Migration Probabilities . . . 56
2.3.1 Transition Matrices . . . 64
2.4 Results . . . 67
2.4.1 Additional Results . . . 79
2.5 Exploring the Selected Control Terms . . . 87
2.6 Wage Inequality . . . 95
2.7 Locational Choice and Returns to Education . . . 100
2.8 Conclusion . . . 101

3 Estimating Union Wage Spillovers: The Role of Bargaining and Emulation Effects 104
3.1 Introduction . . . 104
3.2 The Model . . . 108
3.2.1 Model Set-up . . . 108
3.2.2 Workers . . . 109
3.2.3 Firms . . . 111
3.2.4 Wage Determination . . . 113
3.2.5 Union Determination . . . 117
3.3 Estimation . . . 121
3.3.1 Data . . . 121
3.3.2 The Outside Option Term . . . 122
3.3.3 Dealing with Endogeneity . . . 125
3.3.4 Descriptive Patterns . . . 130
3.3.5 Results . . . 132
3.3.6 Alternative Specifications . . . 141
3.3.7 Controlling for Selectivity . . . 142
3.3.8 Subsample Analysis . . . 146
3.4 The Firm Response to the Union Threat . . . 151
3.4.1 The Wage Response . . . 152
3.4.2 The Amenity Response . . . 155
3.4.3 The Intimidation response . . . 155
3.4.4 Regression Analysis . . . 156
3.5 Conclusion . . . 159

Conclusion . . . 160

Bibliography **162**

A Appendix to Chapter 1 . . . . . **170**
A.1 A Brief Discussion of Artifical Neural Networks 170
A.2 Monte Carlo Figures and Tables 172

B Appendix to Chapter 2 . . . . . **180**
B.1 Data Appendix 180
B.1.1 Data Sources for Roy Model Estimation 185
B.2 Results Tables and Figures 186

C Appendix to Chapter 3 . . . . . **204**
C.1 Mathematical Appendix 204
C.1.1 Derivation of the Firm Surplus 204
C.1.2 Nonunion Wage Equation Derivation 206
C.1.3 Firm Size Derivation 207
C.1.4 Wage Equation Linearisation 208
C.1.5 The Firm Wage Response 209
C.1.6 Decomposition of Mean Wage Movement 211
C.2 Data Appendix 215
C.3 Tables and Figures 220

List of Tables

Table 1.1 Monte Carlo Output: 5 Locations 27
Table 1.2 Monte Carlo Output: 10 Locations 30
Table 1.3 Monte Carlo Output: 3 Locations 32
Table 1.4 Monte Carlo Output: 5 Locations, 10-fold Cross Validated Penalty Level . . . 34
Table 1.5 Monte Carlo Output, 5 Locations: Alternative Methods 36
Table 1.6 Monte Carlo Output: 5 Locations, Multiple Control Functions 38
Table 1.7 Estimated versus Actual Probability Estimates: A Comparison of Approaches for each Data Generating Process 47

Table 2.1 Descriptive Statistics 55
Table 2.2 Summary of Cell Probabilities: Cell Variation 57
Table 2.3 Summary of Cell Probabilities: Individual Variation 57
Table 2.4 Summary of Random Forest Estimated Probabilities 59
Table 2.5 Uncorrected Estimates for Selected States 68
Table 2.6 Corrected Estimates for Selected States using Dahl's Approach: Separate Control Function for Stayers and Movers 69
Table 2.7 Corrected Estimates for Selected States using Dahl's Approach: Single Control Function 71
Table 2.8 Corrected Estimates versus OLS 72
Table 2.9 Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities 75
Table 2.10 Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Controlling for Heteroskedasticity and Weighted Penalty Loadings 81
Table 2.11 Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Separate Control Function for Stayers and Movers 82
Table 2.12 Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Separate Control Function for Each Birth Region 83
Table 2.13 Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Modified Control Function . . . 86
Table 2.14 Summary of Terms Selected 88
Table 2.15 Share of Migrants with a College Education 92
Table 2.16 Factors Affecting Control Term Inclusion 94

Table 2.17 National Level Estimates of the College Earnings Premium 97
Table 2.18 Migration Flows and Differences in Returns 99
Table 2.19 Migration Flows and Differences in Returns: Alternative Return Measure . . 102

Table 3.1 OLS Results 134
Table 3.2 IV Results 136
Table 3.3 Controlling for Selectivity 143
Table 3.4 Subsample Analysis - Coefficient Estimates on Outside Options 145
Table 3.5 Outside Options Contribution to Changing Wages - Subsample Analysis . . . 148
Table 3.6 Outside Options Contribution to Changing Wages - Males Only 150
Table 3.7 Including Wage Emulation Effects 158

Table B.1 Geographic Divisions & Regions 181
Table B.2 Aggregated Industry Definitions 184
Table B.3 Corrected Estimates for Selected States using Dahl's Approach and Random Forest Estimated Probabilities: Separate Control Function for Stayers and Movers 186
Table B.4 Corrected Estimates for Selected States using Dahl's Approach and Random Forest Estimated Probabilities: Single Control Function 187
Table B.5 Corrected Estimates for Selected States using post-double-Lasso and Cell Estimates of Migration Probabilities 188
Table B.6 Corrected Estimates versus OLS 189
Table B.7 Lasso Estimates versus Dahl 190
Table B.8 PDL Estimates versus Uncorrected Estimates for all States: Less than High School Premia 191
Table B.9 PDL Estimates versus Uncorrected Estimates for all States: Some College Premia 192
Table B.10 PDL Estimates versus Uncorrected Estimates for all States: College Premia . 193
Table B.11 PDL Estimates versus Uncorrected Estimates for all States: Advanced Degree Premia 194
Table B.12 PDL Estimates versus Dahl Estimates for all States: Less than High School Premia 195
Table B.13 PDL Estimates versus Dahl Estimates for all States: Some College Premia . . 196
Table B.14 PDL Estimates versus Dahl Estimates for all States: College Premia 197
Table B.15 PDL Estimates versus Dahl Estimates for all States: Advanced Degree Premia 198

Table C.1 SMSA Rankings 215
Table C.2 Changes to SMSA Definitions 1973-2010 216
Table C.3 Aggregated Industry Definitions 219
Table C.4 Alternative Specifications: Alternative Transition Measure 222
Table C.5 Alternative Specifications - Public Sector 223

List of Figures

Figure 1.1 Classification Decision Tree with 5 nodes and 6 leaves 41
Figure 1.2 Model 1: Actual versus Estimated Probabilities 45
Figure 1.3 Model 2: Actual versus Estimated Probabilities 48
Figure 1.4 Model 3: Actual versus Estimated Probabilities 50

Figure 2.1 Probability Estimates: Cells versus Random Forest 61
Figure 2.2 Random Forest Reliability Curve 62
Figure 2.3 Transition Matrix using RF Probabilities 66
Figure 2.4 PDL Corrected Estimates Versus Uncorrected Estimates: All States 77
Figure 2.5 PDL Corrected Versus Dahl Corrected Estimates: All States 78
Figure 2.6 California: Selected Terms and Migration Patterns 91

Figure 3.1 Declining Unionisation Across Selected Cities 129
Figure 3.2 Time Trends in Union Outside Option Premium 131
Figure 3.3 Average Wage Decomposition 139

Figure A.1 Feedforward Neural Network (FNN) with 4 inputs, 1 hidden layer with 5 nodes, and a single output 171
Figure A.2 Distribution of Monte Carlo Estimates: 5 Locations, SISA Holds Violated . 172
Figure A.3 Distribution of Monte Carlo Estimates: 5 Locations, SISA Strongly Violated 173
Figure A.4 Distribution of Monte Carlo Estimates: 5 Locations, SISA Weakly Violated 174
Figure A.5 Distribution of Monte Carlo Estimates: 10 Locations, SISA Holds 175
Figure A.6 Distribution of Monte Carlo Estimates: 10 Locations, SISA Strongly Violated 176
Figure A.7 Distribution of Monte Carlo Estimates: 10 Locations, SISA Weakly Violated 177
Figure A.8 Distribution of Monte Carlo Estimates: 3 Locations, SISA Holds 178
Figure A.9 Distribution of Monte Carlo Estimates: 3 Locations, SISA Strongly Violated 179

Figure B.1 Transition Matrix between Regions 182
Figure B.2 Transition Matrix between Selected States and Regions 183
Figure B.3 Florida: Selected Terms and Migration Patterns 199
Figure B.4 Illinois: Selected Terms and Migration Patterns 200
Figure B.5 Kansas: Selected Terms and Migration Patterns 201
Figure B.6 New York: Selected Terms and Migration Patterns 202
Figure B.7 Texas: Selected Terms and Migration Patterns 203

Figure C.1 Declining Unionisation Across Selected States 220

Figure C.2 Percentage Decline in Unionisation and Transitions into Union Jobs 221

Glossary

AIC	Akaike Information Criterion
ANN	Artificial Neural Network
CV	Cross Validation
IIA	Independence of Irrelevant Alternatives
ML	Machine Learning
MSE	Mean Square Error
PCA	Principal Components Analysis
PLSR	Partial Least Squares Regression
RF	Random Forest
RMSE	Root Mean Square Error
SISA	Single Index Sufficiency Assumption

Acknowledgments

This dissertation would not have been possible were it not for the guidance, support, and encouragement of countless friends, colleagues, and family.

I am deeply grateful to David Green, my supervisor, for his ever thoughtful comments, and suggestions, and for his sage advice. This thesis would not have been possible without his guidance, and I am especially grateful for his generosity with his time. Thomas Lemieux and Craig Riddell, my committee members provided great insight and helpful advice, and were excessively patient when my ideas were less than fully formed.

Beyond my committee, numerous faculty at the VSE provided comments and critiques on this work, improving it beyond measure. Of particular note, Vadim Marmer and Paul Schimpf advised me on the more technical aspects of the project. I am also greatly appreciative of the excellent instruction I received in my early years in the PhD program. In particular, Hiro Kasahara and Vadim Marmer instilled in me a deep interest in econometrics. I was extremely fortunate to have Nicole Fortin teach the first course in Labour Economics I ever took, which changed the direction of my studies entirely.

Throughout the program I was surrounded by a remarkable cohort of supportive friends and talented economists in Alastair Fraser, Brad Hackinen, Jose Pulido, Nouri Najjar, and Tom Cornwall. João Fonseca is a dear friend with whom I enjoyed many stimulating and meaningful discussions over just as many beers. Students outside of this small cohort, too many to mention, provided important comments and insight on my work. Of particular note are Gaëlle Simard-Duplain and Tímea Molnár.

The administrative staff at the VSE were excellent throughout my time in the program. Maureen Chin in particular ensured every deadline was met, dealt with every concern and query, and effectively removed much of the stress associated with graduate studies.

The support of friends and family has been instrumental to the completion of this dissertation. Many thanks to my wonderful friends, Joel Boyd and Peter Gardner, for providing encouragement an ocean away. Since beginning this dissertation my family has grown, and I am grateful to my in-laws, the Trieu family (and one Milligan), for their support. Thanks to the McKay family, Fiona, Neil, Josh, and Lauren, for their support, and for their good company

on my all too infrequent trips home.

My deepest thanks to my grandparents, Margaret and Gordon Greenlee for their ever-constant encouragement and belief in me. For being always ready for a quare wee chin-wag, and never getting scundered when listening to this buck-eejit gurn about the PhD doing his head in. Pastie suppers and football specials are on me.

My parents, Matthew and Julie Snoddy, instilled in me a thirst for learning that sparked this undertaking. To my mum, thank you for inspiring me with your work-ethic, and for giving me a passion for cooking and baking that sustained me throughout my studies. To my brother David. Thank you for always reminding me what matters, and for your friendship, which I cherish.

To my wife Jennifer. It is not possible to express how much you helped me throughout this process. Thank you for your love and support. You mean the world to me.

For Jenn.

Introduction

A fundamental question in the economics of labour is how individuals' well-being, opportunities, and labour market success is shaped by where they live and work. The importance of local labour markets stems from a geographic fragmentation of the national economy into distinct, interconnected, but somewhat separate markets with distinct features. These markets have very different industrial compositions, physical infrastructure, labour market institutions, and the demographics of the local population can differ markedly. The choice of where to live and work is then tied inextricably to the type of work people do and ultimately the enumeration they receive. Given that place of birth is the greatest predictor of where an individual will live as an adult, the role played by local economies in driving earnings and employment outcomes is a significant driver of national income inequality.

For empirical economists the importance of local labour markets is important for reasons beyond their impact on employment and earnings. That is, local labour markets represent a significant source of variation that the empiricist can exploit to estimate causal relationships. For instance, policy varies locally within large countries, meaning that impacts can be estimated by comparing outcomes across regions. However, exploiting regional variation presents distinct challenges to researchers: the (im)mobility of individuals across regions representing a fundamental selection problem that must be addressed in many studies. This dissertation contributes to a fuller understanding of the relationship between labour market success and local labour markets, and develops novel empirical methods to overcome biases inherent in comparing outcomes across regions.

Chapters 1 and 2 of this dissertation examine the importance of residential sorting and its potential role in introducing selection bias. In exploiting variation across regions researchers must account for the fact that individuals living and working in each region form a specific, selected subsample, due to the (in)ability of individuals to relocate geographically. Bias arises when individuals making specific migratory choices are of a particular 'type', and when 'type' correlates with labour market outcomes, and the response to the policy under examination. Central here is that dimensions of 'type' are unobserved by the researcher.

In Chapter 1 I present an improved empirical methodology for overcoming selection bias.

Though this method can be adapted to a wide array of empirical settings, it finds a natural application in the local labour market literature. Controlling for selection bias in this setting is inherently difficult given the dimensionality of the problem as individuals can choose from amongst many regions to live and work. Previous corrective methods have relied on strong, hard to justify, assumptions on the process characterising the migratory decision. The improved method presented in this Chapter leverages the comparative advantages of the machine learning toolkit in model selection to relax these strong assumptions and more flexibly capture patterns of selectivity. I present detailed statistical evidence outlining the efficacy of this improved approach to selection correction.

In Chapter 2 I employ this methodology in an empirical setting where residential sorting introduces significant selection bias; estimating the returns to education separately for regions of the United States. This Chapter serves to both outline the importance of implementing reliable bias correction using the methodology introduced in Chapter 1, and to provide more accurate estimates of education returns to further our understanding of earnings variation across regions. I compare my estimates to those obtained using more traditional methods of selection correction, finding that the method used to control for selectivity is of great importance. On average, I find that the cost of having not completed high school is lower than previous evidence suggests for many states, while the return to having some college education or a Bachelor's degree is on average higher.

Chapter 3 examines the role played by labour market institutions and the industrial structure of the local economy in wage setting. Specifically I consider spillover effects operating between the union and non-union sectors in the local economy. There are two distinct channels under examination here, the first being the ability of workers in the nonunion sector to bargain higher wages when they can credibly find higher paying jobs in the union sector. The second channel is that of wage emulation which arises due to the traditionally hypothesised 'union threat' where nonunion firms pay a premium to prevent a vote in favour of union certification.

I build a novel search and bargaining model which makes clear the channel through which spillover effects operate and informs the empirical methodology employed in the chapter. This model endogenises union formation through voting, the union wage premium through alternative methods of wage bargaining, and allows for non-union firms to stave off unionisation through the paying of higher wages. To capture spillover effects I construct terms which capture variation across cities in the outside options available to workers in the union and non-union sectors. The estimation strategy proceeds by using Bartik-style shift-share instrumental variables. My results indicate that, for the population as a whole, wages would be 2 percentage points higher in 2010 had the size, and composition, of the union sector been fixed over the period 1980-2010. I also find significant differences in the impact of wage spillovers on sub-

populations of workers over the sample period.

Chapter 1

Learning about Selection: An Improved Correction Procedure

1.1 Introduction

The sorting of individuals across locations, or occupations, or into treatment, to name but a few examples, is one of the most common threats to identification faced by empirical researchers. Self-selection across alternatives leads to biased coefficient estimates, due to the non-random assignment of individuals to 'treatment'.[1] Following the pioneering work of Heckman (1979), it is well known that selection bias can be controlled for by including estimated values of the error mean term in the regression model. Furthermore, it is possible to approximate this term using a flexible function of probability terms which represent the likelihood of selecting into each choice (Dahl (2002), Lee (1983)). However, when individuals select across many alternatives, implementing a full, flexible specification of these probability terms is not feasible. Traditionally, controlling for selection in high-dimensional settings has therefore relied on making strong distributional assumptions which cannot be tested empirically. In this contribution, I present an improved procedure, which overcomes the dimensionality problem by using machine learning tools to perform variable selection. This method is easily adopted to alternative empirical settings, is easy to implement, and most importantly, does not rely on non-trivial, and untestable, distributional assumptions.

Prior parametric methods that deal with selection bias typically assume that the error draws entering the utility equation are independent and identically Gumbel distributed, and hence choice probabilities are estimated using the multinomial logit model (Dubin and McFadden (1984), Bourguignon et al. (2007)). As is well known in the literature, this imposes the undesir-

[1] Or more generally, to values of the explanatory variable of interest.

able Independence of Irrelevant Alternatives (IIA) Assumption . Additionally, further parametric assumptions are required to incorporate the multinomial logit within the selection correction framework.[2] Identifying selection using these parametric frameworks also relies on assuming a particular functional form for the selection equation (utility function), and appropriately modelling tastes and preferences. Though these parametric frameworks permit an empirical representation that overcomes the dimensionality problem associated with a multi-choice setting, they impose strong distributional assumptions and will perform poorly when these assumptions are violated. Specifically, the functional form of the control function relies on the underlying distributional assumptions, as do the estimates of the choice probabilities. A more appealing approach then is non-parametric or semi-parametric estimation.[3]

In an important contribution, Dahl (2002) outlines a flexible estimation strategy which is implemented semi-parametrically, and does not rely on the imposition of a specific distribution for the random components driving selectivity. This method is easy to implement and can be applied to a wide range of applications.[4] Specifically, Dahl estimates the probabilities associated with each choice by non-parametrically grouping individuals into discrete cells based on observables. A flexible control function in these terms would fully characterise the selection problem but with many alternatives is necessarily a high dimensional object. To deal with this dimensionality, Dahl makes a strong assumption on the covariance of the joint distribution characterising selectivity. He refers to this assumption as the 'Single Index Sufficiency Assumption (SISA)' which assumes that selection bias in multi-choice settings is fully characterised by variation in just a single probability term: the first-best migration choice.[5] This assumption is unlikely to hold in a wide range of empirical settings. Intuitively, it imposes a restriction on the returns to factors unobserved by researchers across locations, namely, this return must be identical across choices.

Taking Dahl (2002) as my starting point, I make two important methodological contributions to introduce a new selection correction procedure. This method retains the desirable features of Dahl's framework while relaxing the overly strong SISA. Both of these contributions use tools from the machine learning literature which are well-suited to this setting. Firstly, using

[2]See the survey paper by Bourguignon et al. (2007) for a good overview of these methods and a numerical comparison.

[3]See Pinkse (1993) for an earlier contribution in a binary choice setting. See also Vella (1998) for a review of parametric and semi-parametric methods, especially with regards to two-step estimation.

[4]See Ransom (2016) for a recent contribution using Dahl's method to control for selection into occupations. See also Bayer et al. (2009) for an application with regards to air pollution, Bombardini et al. (2012) who control for selection across industries, Bertoli et al. (2013) for an application to international migration, and Carneiro and Lee (2011) who control for selection bias across states and estimates of schooling in a manner similar to Dahl's original application.

[5]As is discussed below, Dahl himself acknowledges the overly strong restriction this assumption imposes and goes some way to relaxing it in empirical implementation.

a two-step selection procedure which relies on Lasso as a selection tool following Belloni et al. (2014), I show how appropriate selection of control terms constructed using choice probabilities can be achieved using machine learning methods. In doing so, I relax the SISA by choosing terms relevant to model fit and selectivity, rather than restricting the variation to that contained in the first-best probability. This allows for a richer pattern of selectivity, and importantly, does not ex-ante restrict the covariance of the underlying joint distribution characterising the selection problem. Instead, the selection of terms allows for more flexible characterisation of the covariance structure driving selectivity in the underlying data. The restriction imposed by index sufficiency is then replaced by a weaker assumption which requires the control function permit an approximately sparse representation. This sparsity assumption does not impose any specific distributional assumptions ex-ante, and instead requires only that some subset of control terms generated using the migration probabilities approximate the true control function. Intuitively, this would be the case if the return to unobserved factors were approximately equal across a subset of locations, for instance. In this case, selectivity is potentially well captured using a single probability term from this subset. This correction method represents a significant departure from all previous methods which have imposed strong restrictions on the underlying distribution.

Given that selection is identified using variation across migration probabilities, estimation of these terms is also of great importance. Dahl uses non-parametric estimates, grouping individuals into bins by observable characteristics and averaging over migration decisions. Due to data limitations, these bins represent a coarse measure and rely on choices by researchers which may not accurately capture the underlying patterns in migration. The second main contribution of this chapter is in exploring the use of machine learning tools in estimating these migration probabilities. Specifically, I use the Random Forest (RF) algorithm developed by Breiman (2001) to obtain estimates of these key terms. Ransom (2016) uses a decision tree framework (Hothorn et al. (2006)) to estimate migration probabilities in a similar context. The RF algorithm was developed in part to deal with issues related to the estimation of a single decision tree. Namely, the random forest uses bootstrap re-sampling and averages estimates across many trees, to avoid over-fitting. As shown by Niculescu-Mizil and Caruana (2005) the RF algorithm performs well in probability estimation relative to other Machine Learning (ML) methods. Using a numerical example I illustrate the benefits to using the RF algorithm over traditional non-parametric methods.

I present Monte Carlo evidence for the efficacy of the two-step selection procedure. In this numerical experiment I replicate the generalised Roy model presented in Section 2. I consider three alternative specifications in which the SISA holds, or is either weakly, or strongly violated. I consider the performance of each method of selection correction when the sample is small,

or large, and when there are many locations, or few. The Lasso selection procedure performs well in these tests, removing mean bias by 75-99% when the sample size is reasonably large. Most importantly, its performance is not dependent on whether the SISA holds. In contrast, I show that Dahl's method can perform poorly when this assumption is violated. However, since it cannot be determined that this assumption holds ex-ante, using this alternative method yields estimates that are robust to assumptions regarding the joint distribution characterising the selection process. Furthermore, standard inference tests are generally not valid when the SISA is violated and estimates are obtained using Dahl's correction. In contrast, my method generates appropriate confidence intervals regardless of the underlying distribution.

A further contribution of this chapter is to the rapidly growing literature appending traditional econometric tools with the tools of machine learning. As noted by Mullainathan and Spiess (2017), machine learning tools are in general difficult to apply to standard economic problems, as they are often designed to accurately predict, out-of-sample, the outcome variable $\hat{y}$, while econometricians typically focus on estimates of coefficients $\hat{\beta}$. However, in dealing with the selection problem, the machine learning toolbox can be applied in a compelling way. Firstly, machine learning tools present a flexible, model-free way to estimate migration probabilities $\hat{P}$ which are central for identification. Importantly, researchers still choose which variables are included in the prediction of these probability terms, and hence the exclusion restriction necessary for identification is still embedded in the framework. Secondly, the variable selection tools used in this chapter aid casual identification through appropriate selection of terms driving selection bias. The goal of the researcher is still the causal estimation of β; machine learning tools, as applied in this contribution, simply aid causal estimation by preventing the need for strong distributional assumptions, and by preventing researchers from making ad-hoc decisions not supported by the data.

This chapter then contrasts with more recent papers in the economics literature which utilise ML tools to aid in prediction tasks, thus re-framing the focus of the economist away from causal inference. In this vein, Kleinberg et al. (2015) discuss the role of 'prediction policy problems' in economics, and as an example consider the use of ML tools to determine whether it is advisable to perform hip surgery on individuals, given predictions of their lifespan. Björkegren and Grissen (2017) predict loan repayments using cell phone data. Specifically, they use Random Forests and logistic regression with a model-selection component. Glaeser et al. (2018) use Google Street View to predict income in New York. This contribution is more closely related to the literature which uses ML tools to aid causal inference. Economists working in this area are making rapid advancements in combining the tools of ML with the traditional focus of econometricians. Athey (2017) provides a detailed summary of the literature in this area and an overview of how ML tools are used in economics more broadly. This chapter relates most

closely to the literature using ML tools in estimating treatment effects under unconfoundedness where treatment is assumed randomly assigned after conditioning on covariates.[6] The double selection procedure of Belloni et al. (2014) used in this chapter is formulated as a means of estimating treatment effects with possibly many control variables. Using these tools, this chapter presents an effective method for dealing with selection bias in a semi-parametric way, that is easy to implement and free of distributional assumptions.

The remainder of the chapter is organised as follows: In section 1.2, I introduce a generalized Roy model which provides the theoretical basis for the selection correction method used in this chapter. In section 1.3, I discuss the important contribution of Dahl (2002) and outline an improvement which uses ML tools to select control terms. In section 1.4, I present a numerical Monte Carlo experiment which makes clear the benefits of the selection procedure suggested in this contribution. Section 1.5 discusses the estimation of migration probabilities which are central to identification and provides an illustrative example of the benefits to estimation using the RF algorithm and section 1.6 concludes.

1.2 Roy model

To make clear the problem under consideration and the assumptions required for identification, I consider selectivity bias in a generalized Roy model using the framework of Dahl (2002), which allows for choices across many locations, and where non-pecuniary factors affect residential choices.[7] Specifically, Dahl considers empirical estimation of the returns to schooling separately for states in the United States. The inclusion of non-pecuniary factors is important given the importance of amenities in driving cross-state migration (Dahl (2002), Kennan and Walker (2011) and Zabek (2018)). Individuals differ both in their level of schooling and in preferences over where to live/work and migrate to the state which maximises their utility. This utility maximisation decision and resulting patterns of migration are what bias the return to education at the state level.

The Roy model formalises the idea that selection bias is driven by individuals' migration decisions resulting, at least in part, from income maximisation. Individuals will move to states where the return to their education is particularly high, and where they have relatively high

[6] See Athey and Imbens (2015) and Athey et al. (2018) for examples in this literature.

[7] The original version of the model was formulated by Roy (1951) and considers just a choice between two alternatives based solely on income maximisation. Heckman and Taber (2008) and Heckman and Vytlacil (2007) provide a thorough review of the Roy model and extensions. Heckman et al. (1990), and French and Taber (2011) present detailed discussions of identification in the Roy model. The Roy model has been applied to a wide range of applications in the literature, including redistributive taxation (Rothschild and Scheuer (2013)), immigration (Borjas (1987), Borjas et al. (1992)), schooling (Willis and Rosen (1979), dHaultfoeuille and Maurel (2013)), field of study (Ransom (2016), Kirkeboen et al. (2016)), social programs (Eisenhauer et al. (2015)) and occupational choice (Ransom (2016), Dolton et al. (1989)).

earnings potential. Selection bias is therefore present due to variation in earnings potential across states. The importance of location for earnings in the United States is well established by Diamond (2016) and Moretti (2013), amongst others, who document increases in the college-high school wage gap between 1980-2000 and the increased sorting of college workers across metropolitan areas over the same period. In the second chapter of this dissertation I mirror the empirical exercise of Dahl to illustrate the efficacy of the selection correction method outlined here. To keep things somewhat general however, I will refer to individuals choice over 'locations' in the remainder of this chapter. I will however, for simplicity continue to refer to education as the core explanatory variable under examination.

1.2.1 Model Overview

The Roy model presented here assumes that individuals have a fixed level of education, are 'born' into one of the C locations, and subsequently make a once-and-for-all decision to stay in their 'birth' location, or move into another location. As such, the model can be viewed as a two-period model, where individuals are born and make their migration decision in the first period, and in the second period they receive earnings, the level of which is determined by their first period choice. The main components of the model are the earnings equation and the utility equation which defines the pattern of selection. In the absence of selection, the unobserved skill distribution would be the same across locations as productivity would be randomly assigned. I will restrict attention to estimation of the earnings equation which is my focus in Chapter 2.

1.2.2 The Outcome Equation

First consider the case of N individuals (i) for whom I observe an outcome y_{ic} which is location dependent. Individuals choose in which location c to locate and are 'born' into location j. There are a finite number of locations C in which individuals can reside. Individuals' birth location is observed and assumed randomly assigned. As individuals are randomly assigned, the distribution of observable and unobservable skills is approximately equal across locations. However, in the model I allow the outcome equation to differ across locations, and hence the skill distribution will be unequal across areas due to self-selection. The earnings equation defines the earnings of individuals in the second period once they begin working, but is observable to individuals in the first period when considering where to live.

The outcome equation for an individual born in j, who resides in c is given by:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + u_{ic}, \quad c = 1, \ldots, C \tag{1.1}$$

β_{1c} and β_{2c} capture the location-specific return to observed characteristics s_i and x_i while α_c

captures a location specific premium. s_i is the level of schooling and x_i is a set of other characteristics which affect earnings. My goal is to correctly identify the coefficient β_{1c}. Both coefficient estimates on s_i and x_i will be biased due to self-selection across locations whereby the distribution of the error terms is not conditional mean zero in all locations and they are correlated with s_i and x_i. In this current set-up, as in the set-up of Dahl, birth location does not enter the outcome equation. Instead, variation across birth locations of individuals with similar personal characteristics is used to identify selection effects here. As noted by Dahl however, it is possible to relax this restriction. If birth location enters the outcome equation, then identification of selection effects will exploit variation within birth locations of individuals with different characteristics. That is, introducing control dummies for birth location in the outcome equation restricts the variation used to identify selection bias, but the same method outlined herein can be applied without modification.

1.2.3 Utility Maximization

In choosing where to live and work individuals make a utility maximising choice, where the utility associated with living in a location is a joint function of individual earnings and preferences. The utility function is assumed to take the following form:

$$V_{ijc} = y_{ic} + \pi_{ijc}, \quad c = 1, \ldots, C \tag{1.2}$$

where π_{ijc} captures all non-wage factors driving preferences for location c for an individual i, given their location of birth j. These will include factors both observable to the researcher such as distance between birth location and residence location, climate, and unobserved factors such as inter-personal relationships and cultural distance.

Tastes are a linear function of observable characteristics z_i and an unobserved component modelled as an error draw ϵ_{ijc} such that:

$$\pi_{ijc} = \gamma_{jc} z_i + \epsilon_{ijc}, \quad c = 1, \ldots, C \tag{1.3}$$

and hence the utility function can be expressed as:

$$\begin{aligned} V_{ijc} &= \mathbb{E}[y_{ic}|s_i, x_i] + \mathbb{E}\Big[\pi_{ijc}|z_i\Big], + \epsilon_{ijc} + u_{ic}, \quad c = 1, \ldots, C \\ &= \vartheta_{jc} + \omega_{ijc}, \quad c = 1, \ldots, C \end{aligned}$$

where ϑ_{jc} is referred to as the subutility function and captures the mean utility for observably similar individuals born in the same location. ω_{ijk} captures the individual specific utility shock

for location c, which is the sum of the preference and the productivity shocks. It captures individual deviation from average utility for location c for observably identical individuals born in the same location.

1.2.4 Selection Bias

Individuals choose to reside in the location that yields the greatest utility given their earnings potential and personal tastes. The utility-maximizing choice can be summarised by the dummy variable D_{ijc}:

$$\begin{aligned} D_{ijc} = 1 \quad &\iff \vartheta_{jc} - \vartheta_{jk} \geq \omega_{ijk} - \omega_{ijc} \quad \forall k \neq c \\ = 0 \quad &\text{otherwise} \end{aligned} \tag{1.4}$$

where individuals choose to live in the location that maximizes utility across all C locations. That is, an individual is observed in a location if it yields the greatest utility across all choices. The choice over alternatives depends on both deterministic and random components of both earnings and preferences across all locations. Furthermore, individuals select over specific migration paths as utility is birth-location dependent given variability in preferences for locations conditional on originating location.

Outcomes are only observed for individuals in the location that yields their greatest utility and hence, I can write the selection rule:

$$y_{ic} \text{ observed} \quad \iff \quad D_{ijc} = 1 \tag{1.5}$$

This equation summarizes the selection rule whereby individuals' outcome is only observed if $C-1$ selection equations are simultaneously satisfied such that c maximizes utility across all C locations. Even if birth locations are randomly assigned, the observed sample will not be randomly distributed across locations. In general:

$$E[u_{ic} | y_{ic} \text{ observed}] = E[u_{ic} | \vartheta_{jc} - \vartheta_{jk} \geq \omega_{ijk} - \omega_{ijc}, \forall k \neq c] \neq 0 \tag{1.6}$$

and estimates will be biased if this conditional mean term is correlated with the explanatory variables in the outcome equation, which is highly likely given that these variables drive differences in the sub-utility functions. The magnitude and direction of the bias will depend on this correlation and the relationship between the outcome equation error term and the error components of the utility function.

1.2.5 Issues in Estimation

Prior to the important contribution of Dahl (2002) estimation of the Roy model typically proceeded with the imposition of parametric and distributional assumptions on the joint distribution of shocks in the earnings and utility equations. With few choices it is possible to estimate the model assuming the error terms are jointly normal, however this becomes difficult to estimate as the choice-set becomes large. With a larger choice-set estimation is still possible using a multinomial logit to estimate migration probabilities and imposing additional parametric assumptions to incorporate this estimation within a selection framework. Therefore, researchers must impose more rigid distributional assumptions when individuals' face choices over a large set of alternatives. In general, parametric estimation will perform poorly when the assumed distribution is notably different from the true underlying distribution.

A further complication is that estimation using these methods requires the researcher to model tastes appropriately, and account for location specific factors which, as mentioned by Dahl (2002) and Ransom (2016), are often unobserved by researchers, or must be proxied by poorly measured alternatives. Dahl deals with this problem by side-stepping the estimation of the utility model, and instead uses a non-parametric approach to estimate migration probabilities which are then used as controls in the outcome equation. Though this overcomes the issues inherent with modelling the utility function, the issue of dimensionality remains. To deal with the curse of dimensionality, Dahl imposes an identifying restriction which constrains the covariance structure of the model error draws. The main contribution of this chapter is to show that using ML tools designed for targeted variable selection, I can relax this overly strong distributional assumption and impose a weaker restriction based on sparsity. The next section discusses Dahl's methodology and my proposed improvement at length.

1.3 Estimating the Earnings Equation

The key difficulty in estimating the problem outlined in the previous section is in overcoming the curse of dimensionality as individuals face a choice over multiple (potentially many) alternatives. In this section, I present the method pioneered by Dahl to overcome the dimensionality problem in a non-parametric setting and the restriction it imposes on the model errors. I then present an improved method which uses a two-step Lasso procedure to select relevant control variables and relies on a weaker identifying assumption.

1.3.1 Dahl's Approach

In correcting for selectivity bias, Dahl uses a reformulation of Lee (1983), to express the joint distribution characterizing the selection problem in terms of a maximum order statistic. This

approach is necessary to reduce the dimensionality of the problem. Specifically in the case of many locations, estimation would require a full specification of the joint distribution of the error term in the outcome equation and differenced error draws in the utility function. This would require the estimation of a (C-1) - fold integral.

Using the insight of Lee (1983), this multidimensional problem can be reduced to a two-dimensional problem. The selection problem can be expressed in terms of the maximum difference of the utility function across locations. That is, though individuals choose among C locations, the choice can be summarized using the observed location, and the maximum utility difference between this choice and the best alternative. Intuitively, the likelihood of observing an individual in a given location can be expressed as a function of the probability that their subutility in this location is the maximum across all alternatives. Specifically, the selection rule in (5) can be expressed as:

$$\begin{aligned} y_{ic} \text{ observed} \quad &\Longleftrightarrow \quad D_{ijc} = 1 \\ &\Longleftrightarrow \quad \vartheta_{jc} - \vartheta_{jk} \geq \omega_{ijk} - \omega_{ijc} \quad \forall k \neq c \\ &\Longleftrightarrow \quad \max_{k} \left(\vartheta_{jk} - \vartheta_{jc} + \omega_{ijk} - \omega_{ijc} \right) \leq 0 \end{aligned} \tag{1.7}$$

which makes clear the relationship between the probability an individual is observed and the maximum order statistic of the differenced subutility functions.

The selection problem is summarised by the joint distribution between the earnings equation error term, and the differenced subutility errors. Following Lee (1983), and re-framing the problem in terms of maximum order statistics allows the problem to be re-expressed in terms of a bivariate joint distribution over the earnings equation error and the maximum order statistic. Lee shows that there is a one-to-one mapping between these distributions:

$$\begin{aligned} &f_{jc}\left(u_{ic}, \omega_{ij1} - \omega_{ijc}, ..., \omega_{ijC} - \omega_{ijc} | \vartheta_{j1} - \vartheta_{jc}, ..., \vartheta_{jC} - \vartheta_{jc} \right) \\ &= g_{jc}\left(u_{ic}, \max_{m} \left(\vartheta_{jm} - \vartheta_{jc} + \omega_{ijm} - \omega_{ijc} \right) | \vartheta_{j1} - \vartheta_{jc}, ..., \vartheta_{jC} - \vartheta_{jc} \right) \end{aligned} \tag{1.8}$$

where f_{jc}, the multi-dimensional distribution maps to the two-dimensional distribution g_{jc}. Note that re-writing the problem in this manner requires no distributional assumptions. Further note that the joint distribution is expressed conditional on the differenced subutility functions which express relative preferences for areas common across all observably equivalent individuals.

By re-framing the problem in terms of maximum order statistics the problem is reduced to a bivariate problem. The differenced sub-utility functions ($\vartheta_{jc} - \vartheta_{j1} \ldots \vartheta_{jc} - \vartheta_{jC}$) are informative about the joint distribution of the maximum order statistic and the earnings equation

error term. Following the insight of single-index models, selection bias can be controlled for in a regression context through the inclusion of a flexible function in the differenced subutility functions:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \sum_j M_{ijc} \times \lambda_{jc}\left(\vartheta_{j1} - \vartheta_{jc}, ..., \vartheta_{jC} - \vartheta_{jc}\right) + v_{ic} \tag{1.9}$$

where M_{ijc} is a dummy variable which takes value 1 if individual i was born in j and resides in c. λ_{jc} implements a flexible function, such as a higher order polynomial, in the differenced subutilty functions. The regression above would yield unbiased estimates of β_{1c}.

Of course, these subutility functions are not known in practice and so estimation of (1.9) is not feasible. The approach of Lee (1983) is to assume that the differenced sub-utility function errors are not informative about selectivity and assume normality of the joint distribution. Dahl follows an alternative approach which forgoes these strong assumptions and allows for non-parametric estimation. Following the insight of Ahn and Powell (1993), it is possible to express selectivity as a function of the probability of selection. As summarised succinctly by Dahl, this is possible in latent index models because the selection mean or the error term is an invertible function of the the selection probability. As such, Dahl generalises the single-index formulation of Ahn and Powell (1993) to the current multiple-index setting.

This allows the joint distribution to be written as follows[8]:

$$\begin{aligned} &g_{jc}\left(u_{ic}, \max_m\left(\vartheta_{jm} - \vartheta_{jc} + \omega_{ijm} - \omega_{ijc}\right) | \vartheta_{j1} - \vartheta_{jc}, ..., \vartheta_{jC} - \vartheta_{jc}\right) \\ &= g_{jc}\left(u_{ic}, \max_m\left(\vartheta_{jm} - \vartheta_{jc} + \omega_{ijm} - \omega_{ijc}\right) | p_{ij1}, ..., p_{ijC}\right) \end{aligned} \tag{1.10}$$

where p_{ijc} is the probability that an individual of type i, born in location j, now resides in location c. These probability terms can be estimated in practice, allowing researchers to side-step estimation of the utility functions. The method by which these terms are estimated will be discussed at length in later sections. The estimating equation can now be expressed as:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \sum_j M_{ijc} \times \mu_{jc}\left(p_{ij1}, ..., p_{ijC}\right) + v_{ic} \tag{1.11}$$

where selection bias is controlled for by a flexible function in the transition probabilities across choice locations. As in equation (1.9), this formulation would yield unbiased estimates of β_{1c}.

[8] Provided the implicit function theorem holds, see Dahl (2002) for details.

The Single Index Sufficiency Assumption

Thus far, re-framing the problem in terms of the maximum order statistic sidesteps the need for estimation of a high-dimensional integral, and instead selection bias can be feasibly controlled for using a control function approach. However, estimation here is still problematic as controlling for selection bias requires the inclusion of a flexible set of interactions between migration probabilities. The number of required terms is potentially large and will grow exponentially as the number of locations grow.

Feasible estimation then requires some restriction on which terms to include in the model. In general however, it is difficult to determine which terms are most relevant to estimation and most accurately capture the underlying pattern of selection bias, the magnitude and direction of which is difficult to pin down in a multi-choice setting. Though few assumptions are used to derive at the formulation in (1.11), feasible estimation requires some restriction on the joint distribution in (1.10). The challenge for researchers is to impose a sufficiently weak restriction such that the control function closely approximates the true unobserved value of $E[u_{ic}|\vartheta_{j1}-\vartheta_{jc},...,\vartheta_{jC}-\vartheta_{jc}]$.

The restriction imposed by Dahl (2002) is what he refers to as the 'Single Index Sufficiency Assumption' (SISA), so-named as it relies on a single migration probability, the probability of the observed choice, to fully characterise the selection problem. The assumption underlying Dahl's empirical implementation is that the full set of information contained in the differenced subutility functions is captured by the probability of the first best migration choice, which of course is the choice observed. A-2 summarises this restriction:

$$\begin{aligned} &g_{jc}\left(u_{ic},\max_m\left(\vartheta_{jm}-\vartheta_{jc}+\omega_{ijm}-\omega_{ijc}\right)|\vartheta_{j1}-\vartheta_{jc},...,\vartheta_{jC}-\vartheta_{jc}\right) \\ &=g_{jc}\left(u_{ic},\max_m\left(\vartheta_{jm}-\vartheta_{jc}+\omega_{ijm}-\omega_{ijc}\right)|p_{ijc}\right) \end{aligned} \tag{A-2}$$

In practice, this assumption reduces the estimating equation to:

$$y_{ic}=\alpha_c+\beta_{1c}s_i+\beta_{2c}x_i+\sum_j M_{ijc}\times\mu_{jc}\left(p_{ijc}\right)+v_{ic} \tag{1.12}$$

which is easily implemented.

The SISA imposes restrictions on the covariance on the selection equation error terms. Specifically it requires that

$$cov(u_{ic},\omega_{ijm}-\omega_{ijc})=K,\quad \forall m\neq k \tag{1.13}$$

which imposes a constant covariance between the selection equation differenced errors and the

outcome equation error draw. Intuitively, this means that there cannot be unobserved factors which are differently valued across choice locations. For instance, non-cognitive skills cannot be differently valued across locations. It is highly likely that this restriction is violated in practice. For instance, if one considers the return to education across locations, then such an assumption implies different education levels are rewarded differently across locations, but not other dimensions of skill unobserved by the researcher. It seems highly plausible that the return to a college education might correlate with other dimensions of skill that are correlated with academic achievement. In this case then, equation (1.12) will fail to account for the full pattern of selection.

To make this restriction more concrete, consider the case that the error term u_{ic} is a function of an $i-c$ specific component, $\tilde{u}_{ic}$ and an individual specific ability draw $\tilde{u}_i$, such that $u_{ic} = \delta_c \tilde{u}_i + \tilde{u}_{ic}$. The variance of each of these components is given, respectively by $\tilde{\sigma}_{1c}$ and $\tilde{\sigma}_{2c}$. If I assume that individual taste shocks are uncorrelated with unobserved ability (which may not be true in practice), and are uncorrelated across birth and residence locations, then the covariance can be expressed as:

$$\begin{aligned} & cov(u_{ic}, \omega_{ijm} - \omega_{ijc}) \\ & = cov(u_{ic}, u_{im} - u_{ic} + \epsilon_{ijm} - \epsilon_{ijc}) \\ & = cov(u_{ic}, u_{im}) + var(u_{ic}) \\ & = \delta_c(\delta_m - \delta_c)\tilde{\sigma}_{1c} - \tilde{\sigma}_{2c} \end{aligned} \tag{1.14}$$

where the error terms $\tilde{u}_i$ and $\tilde{u}_{ic}$ are uncorrelated. Note that if $\delta_m \neq \delta_c$ then the return to unobserved ability is not equal across locations and hence the SISA is violated. Alternatively, the SISA is violated if taste shocks are correlated with $\tilde{u}_{ic}$ in a manner which is not constant across locations.

It is not immediately obvious how to relax this assumption. That is, the first best migration probability is a natural choice for inclusion, but the question facing researchers is which of the remaining terms are relevant, and which are most informative? There is no clear theoretical guidance over which subset of terms to include. Furthermore, it is impossible to determine whether the SISA holds in practice and so researchers cannot determine the degree to which imposing the SISA is biasing their results. In practice, Dahl imposes a slightly weaker restriction than (A-2). He includes the probability that individuals remain in their birth location as an

additional term.[9] There is no clear theoretical guidance motivating the inclusion of this term, nor is it clear whether this term will allow the control function to better capture the covariance structure of the error draws.

1.3.2 An Improvement: Post-Double Lasso

By using ML tools it is possible to relax the SISA and impose a less stringent restriction on the data. Specifically, it is possible to control for selectivity without imposing a covariance restriction, or any direct restriction on the pattern of selectivity, and instead rely on an assumption of approximate sparsity.

A fully flexible specification in the probability terms would capture selectivity, but dimensionality prevents researchers from implementing a fully flexible specification, and in general it is not possible for researchers to determine which probability terms to select for inclusion, or how to weight the importance of each term. Though Dahl recommends including the next best probabilities (the second and third best for instance), in general it is not possible to identify which probabilities correspond to each individuals ordinal ranking. It is also not possible for the researcher to determine which probability terms are more informative regarding the covariance structure of the selectivity distribution. If the return to unobserved ability in locations 1 and 2 is the same, the differences in the value of the migration probabilities associated with each location is due to differences in tastes or returns to education. Including the migration probability of being in location 2 provides no additional information regarding selectivity. If however, the returns to unobserved ability were different in location 3, then the probability term associated with this location is informative regarding selectivity based on unobserved earnings draws. Even if the second, third and fourth best locations could be identified, it is not clear that including these would yield better estimates than including the fifth best probability.

ML tools designed for variable selection can be used to select the set of probability terms to include in the model. Generally, these tools are designed to select a limited number of terms, weighing the noise-signal trade-off inherent in the inclusion of additional terms. The approach here then specifies $\mu_{jc}(.)$ as a set of flexible interactions in the migration probabilities and selects the terms from this set which are most relevant to estimation. The final estimating equation

[9] His assumption then becomes:

$$\begin{aligned} & g_{jc}\left(u_{ic}, \max_m\left(\vartheta_{jm}-\vartheta_{jc}+\omega_{ijm}-\omega_{ijc}\right) | \vartheta_{j1}-\vartheta_{jc}, ..., \vartheta_{jC}-\vartheta_{jc}\right) \\ & = g_{jc}\left(u_{ic}, \max_m\left(\vartheta_{jm}-\vartheta_{jc}+\omega_{ijm}-\omega_{ijc}\right) | p_{ijc}, p_{ijj}\right). \end{aligned} \tag{A-3}$$

then includes the set of terms chosen from this set such that the equation becomes:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \sum_j M_{ijc} \times \hat{\mu}_{jc}\left(p_{ij1}, \dots p_{ijC}\right) + v_{ic} \tag{1.15}$$

where $\hat{\mu}_{jc}\left(p_{ij1}, \dots p_{ijC}\right)$ represents the set of terms from the fully specified model chosen to be included in the final specification.

ML tools then make data driven decisions on term inclusion, weighing the desirability of including additional terms in the model. As such they overcome the dimensionality problem inherent in approximating an unknown function of a large set of terms. It is not possible to determine the underlying covariance structure of the joint distribution in (1.10), and hence select only the migration probability terms necessary to capture selectivity. What ML tools provide is a data driven way to flexibly characterise selectivity without imposing rigid assumptions on the problem.

Importantly, the terms necessary to include in the model are those which correlate with both earnings, and the explanatory variable of interest. Intuitively, if utility preferences did not correlate with schooling then estimates of the effect of schooling on earnings would be unbiased. As established by Belloni et al. (2014), when considering confounders (omitted variables or selectivity control terms for instance) that drive both the treatment and correlate with the observed outcome, it is necessary to perform a double selection procedure which explicitly accounts for this two-sided correlation. The validity of this procedure is determined by its ability to select the relevant, limited set of controls that well approximates the true underlying control functions summarizing their relation to both the outcome and the treatment variable.

The main variable selection procedure considered in this contribution is that of post-double-Lasso (PDL) as formulated by Belloni et al. (2014). Lasso coefficients solve the $\mathscr{L}^1$ regularization problem:

$$\min_{\beta} ||y - X\beta||_2^2 \text{ subject to } ||\beta||_1 \leq t \tag{1.16}$$

where t is a tuning parameter that determines regularization and $||\beta||_1$ is the standard $\mathscr{L}^1$ norm.[10] Lasso was introduced by Tibshirani (1996), and performs variable selection by shrinking coefficients to zero, effectively excluding them from the model (Hastie et al., 2015). As such, it yields a sparse solution by providing feature selection. Like OLS, Lasso provides estimates of β by minimising the residual sum of squares, but unlike OLS, Lasso imposes a constraint on the absolute sum of the K coefficient estimates. This constraint serves to place limits on the coefficient vector such that relatively low weight is placed on variables that tend to perform poorly

[10] Specifically $||\beta||_1 = \sum_{k=1}^{K} |\beta_k|$.

as predictors. That is, variables which tend to introduce a lot of noise relative to explanatory power.[11] The $\mathscr{L}^1$ penalty shrinks coefficients to zero, therefore assigning a positive explanatory weight to the most powerful predictive terms which introduce the least noise.[12]

Belloni et al. (2014) establish a two-step procedure which uses the variable selection properties of Lasso to select control variables most relevant to estimation. Specifically, this method selects terms highly correlated with either the outcome variable, or with the core explanatory variables. The latter is important in this context as the control terms approximate the non-zero component of error mean term. If selection bias is driving estimates, then selectivity differs across education groups and these control variables will be correlated with educational attainment. The basic procedure is as follows:

1. Perform Lasso of the outcome variable on the full set of potential controls. Store variables with a non-zero coefficient.

2. Perform Lasso of the key explanatory variables on the full set of potential controls. Store variables with a non-zero coefficient.

3. Perform OLS of the outcome variable on the key explanatory variable plus the intersection of variables retained from steps (1) and (2).

In their equation (2.8) they show that the post-double-selection estimator is given by:

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha,\beta}{\arg\min}\{\mathbb{E}\left(y_i - s_i\alpha - x_i'\beta\right)^2 : \beta_j = 0, \forall j \notin \hat{I}\} \tag{1.17}$$

where $\hat{I}$ is the set of x controls selected using steps 1 and 2 above.

I use PDL as the main tool of variable selection in this contribution. By selecting terms both relevant to model fit and strongly related to the explanatory variables of interest, in this case education, I control for selection bias which operates through different selection across locations by educational attainment. By using PDL, the curse of dimensionality is overcome in a manner that still allows for flexible estimation of the control function. The final estimating equation then assumes approximate sparsity such that the control function is well approximated by a subset of the terms included in the most general regression formulation in (1.11).

Belloni et al. (2014) establish that the set of included control terms need only well *approximate* the true underlying control function, for estimates of the core parameter of interest to

[11] Importantly, prior to performing Lasso regularization all variables must be standardized to have the same standard deviation and mean, otherwise the degree of regularization performed on each variable will be (in part) due to its scale rather than its explanatory power.

[12] Other forms of regularization, such as $\mathscr{L}^2$ regularization exist, but do not shrink coefficients to zero, hence not providing variable selection.

be asymptotically unbiased. Consider the following relationship between probability controls, earnings, and education for a single location (assuming away dependence on birth state for ease of exposition):

$$y_i = \alpha + \beta_1 s_i + \beta_2 x_i + g(p_{i1}, ...p_{iC}) + v_i \tag{1.18}$$

and

$$s_i = f(p_{i1}, ...p_{iC}) + \epsilon_i \tag{1.19}$$

Earnings and education are correlated with preferences across C locations. $g(p_{i1}, ...p_{iC})$ and $f(p_{i1}, ...p_{iC})$ are the true underlying functions of the probability terms, which are unknown in practice. Provided that these control functions are well approximated using a limited set of terms, such that the problem permits a sparse representation, then estimates of β_1 will be consistent.

To put this more concretely, define $\tilde{g}(p_{i1}, ...p_{iC}) = \gamma_g \tilde{P}_i$ and $\tilde{f}(p_{i1}, ...p_{iC}) = \gamma_f \tilde{P}_i$ as the PDL approximated control functions, where $\tilde{P}_i$ is a higher order polynomial expansion of the full set of C probability terms, and γ_g and γ_f are weighting matrices which assign zero to excluded terms. If the choice set is such that the approximation errors $r_{gi} = g(p_{i1}, ...p_{iC}) - \tilde{g}(p_{i1}, ...p_{iC})$ and $r_{fi} = f(p_{i1}, ...p_{iC}) - \tilde{f}(p_{i1}, ...p_{iC})$ are small, then estimates of β_1 will be consistent (Belloni et al. (2014)).[13][14] This is the assumption of approximate sparsity, that the control functions need only be approximated to a reasonably small margin of error. It is this assumption that replaces that of single-index sufficiency when using PDL to select terms rather than using only the first-best migration probability to capture selection bias. Importantly, approximate sparsity does not impose any a priori restrictions on the distribution characterising selectivity and can therefore characterise much richer patterns of selectivity.

There are clear reasons to believe the final estimating equation is indeed a sparse formulation of (1.11). Specifically, the SISA may indeed hold, and in this case only a single migration probability term is required to identify selection bias. Secondly, the returns to unobserved factors may be similar across subsets of locations, yielding a sparse formulation where only a subset of migration probabilities need be to included. In the case that the returns are unique across all choices, their values may be close enough across some locations such that a full empirical implementation is not required. Furthermore, some migration paths may be largely unobserved for the sample under consideration such that these terms are not particularly informative. Finally, the use of higher order polynomials serve only as an approximation, and some higher

[13] Plug-in standard errors from the final step OLS will also be consistent.

[14] The conditions for sparsity given by Belloni et al. (2014) are that at most s«n elements of the selection matrices γ_g and γ_f are non-zero (where n is the size of the full set of potential controls), and additionally, that the approximation error obeys: $(E[r_i^2])^{1/2} \lesssim \sqrt{s/n}$. Additional regularity conditions necessary to establish the asymptotic properties of the estimator are provided by the authors.

order interaction terms may provide little information to the model relative to the noise they are introducing.

Square-Root Lasso

In the Monte Carlo experiment I will use square root Lasso as formulated by Belloni et al. (2011):

$$\min_{\beta} \left\{ (y - X\beta)^T (y - X\beta) \right\}^{1/2} \quad \text{subject to} \quad ||\beta||_1 \leq t \tag{1.20}$$

where t, the penalty level, is a tuning parameter which determines the amount of regularization. Proper calibration of this parameter is important to avoid over-fitting, and for appropriate variable selection. Typically Cross Validation (CV) is used to select the penalty level. CV involves generating K distinct bootstrap sub-samples, or folds, from the population. Each fold is then split into a training and testing sample, and the model is estimated using the training data only. The testing sample is then used to generate out-of-sample predictions using the collected parameter estimates. In this manner model error is calculated by comparing the predicted outcomes to the true outcomes in the testing sample. The final estimate of the model error is calculated by averaging across the errors associated with each fold. In replicating the procedure for alternative parameter values, in this case for alternative values of the penalty level, the value associated with the lowest model error is selected. An alternative and less computationally burdensome means of selecting the tuning parameter would be to use Adaptive Validation (AV) which requires that researchers estimate the model for each value of the tuning parameter just once rather than once per fold (Chichignoud et al. (2016)).[15]

However, there exist theoretical bounds on the Lasso penalty level which achieve near-oracle rates of convergence, defined as the fastest rate at which parameter estimates converge to the true value (Belloni et al. (2011)). Belloni et al. (2011) present this closed form solution for the Lasso penalty level but highlight how it relies on both, assuming the model errors are normally distributed, and knowing the standard deviation of the errors. Instead the authors propose square-root Lasso (as defined above) and derive a closed form solution for the optimal penalty level. The benefit to using square-root Lasso over traditional Lasso is that this penalty level is pivotal with respect to the standard deviation of the model errors, meaning that knowledge of this parameter is not required. Additionally, this penalty level is arrived at without assuming normality, meaning it will achieve oracle-performance if the model errors are not normally distributed. As shown by the authors, CV is dominated by square-root Lasso estimates using

[15] It is worth highlighting here that this optimal choice of the penalty level is available only for Lasso in the linear regression context and in similarly straightforward implementations. No such property has been established in the literature for generalized linear models such as the penalized multinomial logit. CV or AV methods are therefore the only feasible approaches to penalty level selection at present.

the established theoretical penalty level, and, of course, is much more computationally burdensome. The optimal penalty level for square-root Lasso is given by:

$$t = \frac{c}{\sqrt{n}}\Phi^{-1}(1-\alpha/2p) \tag{1.21}$$

where n is the sample size, p is the number of choice terms, c is some constant greater than 1, Φ is the standard normal CDF, and α determines the probability that the bias is below the amount determined by oracle performance.

It is worth pointing out that this square-root Lasso formulation assumes homoskedasticity and uses a penalty level that is constant across variables. Belloni et al. (2012) present an alternative Lasso formulation which allows for separate penalty loadings for each variable:

$$\min_{\beta} ||y - X\beta||_2^2 \text{ subject to } ||\Psi\beta||_1 \leq t \tag{1.22}$$

where Ψ is a K dimensional column vector, with each element corresponding to a penalty weighting associated with β_k. When the ideal penalty loadings are known, the authors establish a solution for the penalty level t, which achieves near-oracle convergence. This result is arrived at without assuming normality and holds if errors are heteroskedastic. The ideal penalty loadings (Ψ), however, rely on having an accurate estimate of the model errors. Belloni et al. (2012) show how an estimate can be obtained using an iterative procedure which begins using conservative penalty loadings and updates the penalty loadings on each iteration until convergence.[16] In the Monte Carlo exercise that follows, and in Chapter 2 I primarily use square-root Lasso. I consider this alternative formulation in Chapter 2 to examine the robustness of my estimates.

1.3.3 Identification

The technical details regarding implementation above do not make clear how the model is identified. As is well known in the literature on Roy model estimation, identification relies on an exclusion restriction where some variable, or grouping of variables influences the likelihood of moving, but does not enter the wage equation directly. Identification then requires that migration probabilities are estimated using some variable that drives the likelihood of moving to a particular location, but does not drive wages.

Individuals differ in the value of migration probabilities which captures variation in average preferences for individuals who are observably similar. If the migration probability estimated

[16] Specifically, the authors provide values for the initial guess of the standard deviation and the penalty t in their appendix A. The algorithm then proceeds by estimating the residuals from the post-Lasso estimator and repeating the procedure. Details provided in their appendix algorithm A.1.

using these variables is large, then average preferences for moving to (or staying in) the location are relatively high. Correspondingly, these individuals are unlikely, on average, to have migrated based on high unobserved ability draws. In comparison, individuals with a low migration probability are likely to have moved to an area due to a preference or earnings shock. In general, the full set of migration probabilities will be informative, such that even if the probability is high in the current location, the fact that it is low for another location may reveal some additional information about selectivity. Selection bias is identified by comparing earnings across individuals in the same location who are observable equivalent but differ in the likelihood that they live in each location.

Finally, it is worthwhile to highlight that the identification of each migration probability requires a separate exclusion restriction. That is, there must be as many exclusion restrictions as locations. Birth location is appealing as an exclusion restriction for this reason and as mentioned above, it can be interacted with some other set of variables such that birth effects can still be controlled for in the wage equation and selection effects are identified across some other set of characteristics within birth locations.

1.3.4 Restricting the Control Function

Thus far the multiple-index model has been formulated allowing for a separate control function for each birth location. In practice, with many choice locations, such an implementation may be intractable. Importantly, the choice of whether to allow for a separate control function for each birth location is case dependent and depends crucially on the size of the sample originating from each birth location. If SISA held in the data, then implementing a separate control function for each birth location implies only the coefficient on the first-best probability be estimated for each birth location. For all but the very smallest birth location populations this should be implementable.

When SISA does not hold, then ideally researchers will select a subset of terms such that approximate sparsity holds *for each birth state*. Approximate sparsity requires that the subset of migration probability terms selected in the model captures well approximates the true underlying control function which characterises selectivity. Importantly, the distribution characterising selectivity is defined separately on each birth location, implying a different set of terms is required to capture selectivity for individuals from each location.

Variable selection will then be performed separately for each birth location. When the sample of individuals from each birth location is large, then a relatively large number of control terms will be selected, and it is likely that the underlying control function will be well approximated for all birth locations. In this sense, coefficient estimates obtained using PDL correction are consistent, as the control function will be well approximated as the sample size gets large.

However, in applications where the sample size is small for some birth locations, then relatively few terms will be selected for these locations, and it is less likely that the control function is well approximated. Note that the problem may still permit a sparse representation, merely that the sample size is not large enough to permit selection of the terms defining the sparse representation. The distribution characterising selectivity will therefore be better approximated for subsets of individuals depending on their originating location. It is not necessarily the case that selectivity is poorly characterised when the sample is small, but without foreknowledge of the covariance structure of the joint distribution in (1.10), it is not possible to conclude whether a large set of control terms is needed, or a limited set will suffice. In general though, when the sample is relatively large, and many control terms are selected via PDL, then there is greater certainty the control function is well approximated.

Given the inability of researchers to draw firm conclusions on whether limited variable selection for poorly populated birth locations is indicative of poor characterization of selectivity for these locations, a simpler formulation of the problem which sidesteps separate estimation by birth location may be desirable. A reasonable assumption to make in this instance is that the joint distribution of the error terms $g_{jk}(\cdot,\cdot)$ is independent of the birth location j. That is:

$$g_{jk}(\cdot,\cdot) = g_k(\cdot,\cdot), \quad \forall j. \tag{1.23}$$

This assumes that migration probabilities capture the same pattern of selection regardless of birth location. This restriction is imposed by Dahl for simplicity in estimating the control functions,[17] and is generally made in empirical applications. In general, it is not necessary to impose such a restriction in the case that there are relatively few locations or the sample is large, such that each birth location is well populated. One alternative would be to group birth locations together. This relaxed assumption implies that the covariance between earnings and sub-utility is approximately the same across these location groupings. However, even here it is possible that Lasso will under-select terms given that there are relatively few individuals that move across locations relative to stayers.

The implications and desirability then of such a restriction will vary across empirical contexts. Allowing for separate control functions will restrict the number of terms included to control for selection bias for each location. This increases the likelihood that selectivity is poorly characterized for poorly populated birth locations, and hence that coefficient estimates deviate from the true value. Imposing the same control function across locations ensures that greater potential selection bias is captured by the selection terms, but imposes the restriction that the covariance driving selectivity is identical regardless of birth location. Vitally, the choice of

[17] Dahl's restriction is slightly different, and allows for different control functions for 'movers' and 'stayers'.

whether to impose this restriction is context dependent and the method outlined in this chapter does not rely on an imposition of (1.23). Instead researchers must decide whether the sample size of individuals originating in each birth location is sufficiently large as to permit a rich enough set of control terms to be included in the estimating equation such that this set of terms likely defines a sparse representation of the true underlying control function. I explore the desirability, and the sensitivity of results to the imposition of this restriction in the empirical exercise in Chapter 2.

1.4 Monte Carlo Experiment

In this section, I provide numerical evidence on the efficacy of the method outlined above. To do so, I present results from a Monte Carlo experiment which is similar to that of Dahl (2002), allowing for a direct comparison. I compare the results obtained using PDL to methods obtained using Dahl's, method which imposes the SISA, and a fully specified model based on (1.11).

1.4.1 Model

Following Dahl the model takes the form:

$$\begin{aligned} y_{ic} &= \beta_c x_i + u_{ic} \\ u_{ic} &= \tau_c a_i + b_{ic} \\ t_{ijc} &= \gamma_{jc} z_i + \epsilon_{ijc} \\ V_{ijc} &= y_{ic} + t_{ijc} \end{aligned}$$

where N individuals (i) are drawn and assigned a birth location (j). Their utility for living in each location (c) is given by V_{ijc}, which is composed of their earnings in that state (y_{ic}) plus their tastes t_{ijc}. u_{ic} is the component of earnings unobserved by the econometrician whose goal is to estimate β_c. Individuals move based on their utility preferences such that:

$$y_{ic} \text{ is observed if and only if } V_{ijc} \geq V_{ijm} \quad \forall m.$$

The exogenous variables and error draws are drawn from the same distributions as in Dahl. x_i is drawn from a uniform distribution and takes integer values between 1 and 5, while z_i takes integer values between 1 to 10 also with equal probability. Remaining draws are as follows: $a_i \sim \mathcal{N}(0, 1)$, $b_{ic} \sim \mathcal{N}(0, 1)$ and $\epsilon_{ijc} \sim \mathcal{N}(0, 1)$.

Without loss of generality, I consider only estimation of β_1 which I set equal to 1. The

remaining β values are set arbitrarily and take values between 0.5 and 1.75. Similarly, arbitrary values for γ take value between -0.25 and 0.25. So far, the experiment matches identically with that of Dahl[18] The model set up is such that only a single control function is required for each birth location.

1.4.2 Implementation

I first consider a specification where $\tau_c = 1$ for all c and hence the SISA holds given the model above. As a second case, I consider relatively weak violations of the SISA, where for a handful of locations, the covariance of the earnings equation error terms and utility function errors differs. Finally, I consider a strong violation of the SISA where $\tau_c = \beta_c$ and hence the covariance structure is different for all choices. This exercise mirrors the SISA violation considered by Dahl and sets the return to unobserved ability equal to the returns to education. The results of this section are robust to other strong violations of SISA where τ_c and β_c are weakly positively correlated, negatively correlated, or uncorrelated. In all cases, I consider a small sample of 1000 and a larger sample of 10000 individuals born in each location. Dahl's method is implemented using a cubic in the first-best probability. I also run a flexible specification which includes a cubic interaction in all choice probabilities, which I refer to as the 'full' model.[19] Starting from this specification PDL is used to select which of these terms to include and which to exclude from estimation. In all cases the key variable x_i is included without modification.

Migration probabilities are estimated non-parametrically by grouping individuals into cells. Individuals are grouped according to their values of z_i and x_i.[20] A single correction function is used for each birth location, imposing the assumption on the joint distribution of the error terms discussed above. This assumption holds here given the specification of the error draws.

1.4.3 Results

In this section, I present results across a variety of alternative specifications which differ according to the number of locations considered, the method used for estimation, and the distribution of the error terms.

[18]Dahl (2002) sets parameter values within this range but I do not have access to the exact parameter set used by Dahl and so the exact values of $\beta_2 \dots \beta_N$ and γ are not the same.

[19]In results not shown I also present estimates based on Lee (1983). These results match those of Dahl (2002) very closely which is in line with results found by Dahl (2002) and Bourguignon et al. (2007). Given normality of the error draws this is unsurprising, but as shown by Bourguignon et al. (2007), Lee's method will perform very badly when the data generating process deviates from normality.

[20]I will discuss the estimation of migration probabilities in the following section, but using cell probabilities here is appropriate given that the researcher knows the true data generating process. Given the extra noise coming from using estimates rather than true values, this should bias against the PDL method.

Table 1.1: Monte Carlo Output: 5 Locations

	(a) $N = 1000$					(b) $N = 10,000$				
	Bias	RMSE	Std. Dev.	Cov.	Av. Samp.	Bias	RMSE	Std. Dev.	Cov.	Av. Samp.
	SISA HOLDS: $\tau_c = 1 \ \ \forall \ c$									
OLS	0.081	0.092	0.044	0.526		0.081	0.082	0.014	0.000	
DAHL	0.039	0.063	0.050	0.874	608	-0.006	0.018	0.017	0.935	6079
FULL	0.048	0.088	0.074	0.901		0.014	0.036	0.033	0.923	
LASSO	0.047	0.082	0.067	0.891		0.005	0.027	0.027	0.946	
	SISA STRONG VIOLATION: $\tau_c = \beta_c \ \ \forall \ c$									
OLS	-0.045	0.062	0.043	0.809		-0.046	0.048	0.014	0.074	
DAHL	-0.083	0.096	0.048	0.581	604	-0.116	0.117	0.017	0.000	6031
FULL	-0.010	0.073	0.072	0.945		-0.003	0.034	0.034	0.947	
LASSO	-0.010	0.066	0.065	0.947		-0.007	0.028	0.027	0.946	
	SISA WEAK VIOLATION: $\tau_j \neq 1, \tau_c = 1 \ \ \forall \ c \neq j$									
OLS	0.137	0.143	0.041	0.083		0.137	0.137	0.013	0.000	
DAHL	0.092	0.103	0.047	0.483		0.040	0.043	0.017	0.313	
FULL	0.078	0.105	0.071	0.786	611	0.024	0.040	0.032	0.880	6102
LASSO	0.073	0.097	0.064	0.777		0.011	0.029	0.027	0.923	
PARTIAL	0.077	0.098	0.061	0.737		0.010	0.025	0.023	0.921	

Notes N is the number of individuals 'born' in each location. 2000 replications are performed for each specification. 'DAHL' refers to the inclusion of a cubic in the first best probability. 'FULL' includes cubic interactions in all probabilities. 'LASSO' selects terms using post-double-Lasso. 'PARTIAL' implements a cubic in the first best probability and in the probability corresponding to the location with a unique value of τ_c. For each specification, mean bias, root mean-square-error, the standard deviation, coverage probabilities and the average sample size is reported. To estimate coverage probabilities (Cov.), I calculate 95% confidence intervals for each replication and calculate the proportion of replications where the true value ($\beta_c = 1$) falls within the confidence band. Bootstrapped standard errors are used to construct confidence bands for 'FULL', 'DAHL', and 'PARTIAL' specifications.

5 locations

The results of the Monte Carlo experiment provide evidence on the efficacy of the PDL selection method. In Table 1.1, the results are presented for the case of five locations. The baseline case in which the SISA holds is presented in the upper panel for small and large samples. Appendix Figures A.2-A.4 presents histograms and smoothed kernel density estimates of the distribution of estimates across replications for all specifications presented in this table. In the small sample case neither method removes selection bias completely. Dahl's method performs best however, removing 52% of bias compared to 42% for the Lasso method. Dahl's method also performs best in terms of Root Mean Square Error (RMSE) and the variation of estimates across replications is notably lower. The standard deviation of estimates is about 40% higher using Lasso compared to Dahl's method. This is unsurprising given the inclusion of additional terms using Lasso. The variability in Lasso estimates is notably lower than the full model and this result is replicated across all Monte Carlo specifications. In estimating confidence bands for each replication I find evidence however that even though Dahl's method performs best in terms of bias reduction, RMSE and variability, all three methods perform reasonably well in terms of coverage. The true coefficient value lies within the 95% confidence interval around 90% of the time for each method.

In moving to the large sample results in (b), what becomes clear is that Dahl's method no longer dominates Lasso in terms of bias reduction, although it does still outperform the full specification. Given the additional variability in estimates when using PDL, Dahl's method still performs better in terms of RMSE.[21] However, again all three methods lead to approximately appropriate coverage rates. When the SISA holds then it is clear that Dahl's method performs better than PDL and the full specification. Lasso tends to perform as well, if not better than, the full specification in terms of bias reduction, and notably reduces the variance of results. Since the full specification cannot be feasibly implemented in many empirical instances this implies that estimates based on Lasso will do at least as well as the full model in bias reduction.

In the second panel I consider the performance of each alternative method when the SISA is strongly violated, meaning that the return to unobserved factors are differently valued across each location. Specifically, I specify that the return to unobserved ability is equal to the return to education. As the SISA is violated in this manner, the ideal empirical implementation would include a flexible function of all migration probabilities as in the full specification. As is clear

[21] I leave it as an open question whether the sampling distribution between Dahl's methodology and the PDL method will converge in large samples. However, it is worth highlighting that Dahl's method chooses the most limited subset of terms necessary to approximate selectivity (when SISA holds), whereas PDL will continue to select confounders beyond this limited subset. Importantly these terms will still correlate with the explanatory variable, and with the choice decision, but they will include no additional information regarding selectivity once the first best probability has been included. This suggests the possibility that the sampling distributions need not overlap as the sample size grows.

in the small sample case, the full specification does indeed remove much of the bias in estimation. Lasso performs as well in terms of bias reduction but again performs better in terms of the variability of estimates and in RMSE. Both methods also have appropriate coverage probabilities which are almost exactly 95%. Dahl's method performs very poorly in this case.[22] In fact, results obtained imposing the SISA completely mischaracterize the selection problem and increase the distance of estimates from the true value. Dahl's method here picks up an upward bias in estimates, when in fact estimates are downward biased. As a result, Dahl's method does poorly in terms of RMSE and has poor rates of coverage implying inference based on Dahl's method would not be valid in this instance.

In the large sample estimates, both the full specification and Lasso perform slightly better in terms of bias reduction than in the small sample case, with the full specification removing over 90% of mean bias. Lasso performs slightly worse in terms of bias reduction but does better in terms of RMSE due to reduced variability of estimates. Both methods obtain appropriate coverage probabilities here. Dahl's method performs even worse in the large sample case, increasing bias by more than 100%. Given the narrow band of estimates around the incorrect value, coverage probabilities are nearly 0 for both OLS estimates and estimates obtained using Dahl.

Finally, I consider the case of a weak violation of the SISA. In this case, the returns to unobserved ability are differently valued in just one location. This specification would suggest a partial implementation where a subset of migration terms are included. I consider such a subset by including the first best migration probability and the probability associated with the location differing in τ_c. I include a cubic function in these two terms. I consider this partial specification to be the ideal scenario when the underlying utility function is known to the researcher.

In the small sample case in panel (a), both the full, partial, and Lasso specifications perform as well in terms of mean bias, although they remove less than 50% of the bias on average. Dahl's method removes less bias but performs better than the full specification in terms of RMSE as estimates are less variable. However, Dahl's method performs the worst in terms of coverage due to its relatively poor performance in bias reduction. As neither method removes bias completely, neither has coverage probabilities close to 95%. Considering the large sample, both Lasso and the partial specification remove over 90% of bias. Notably, Lasso performs better than the full specification, removing an additional 10% of the selection bias on average. Dahl's specification in contrast removes just 70% of the total bias and has a coverage rate of just 31%. In contrast, Lasso performs as well as the partial specification in terms of coverage, with both

[22] It is important to highlight that this need not be the case. Dahl's method could indeed perform well in this case. Whether it performs well or poorly is purely coincidental and depends on the underlying pattern of selection. There are no clear cases in which Dahl's method will perform well when the SISA does not hold, and as shown in this case, Dahl's method could completely mischaracterize the pattern of selection.

Table 1.2: Monte Carlo Output: 10 Locations

	(a) $N = 1000$					(b) $N = 10,000$				
	Bias	RMSE	Std. Dev.	Cov.	Av. Samp.	Bias	RMSE	Std. Dev.	Cov.	Av. Samp.
	SISA HOLDS: $\tau_c = 1 \ \ \forall \ c$									
OLS	0.092	0.103	0.047	0.505		0.091	0.092	0.015	0.000	
DAHL	0.058	0.078	0.052	0.800	532	0.008	0.021	0.019	0.922	5312
FULL	0.073	0.128	0.105	0.888		0.023	0.052	0.046	0.924	
LASSO	0.067	0.097	0.070	0.814		0.023	0.039	0.032	0.878	
	SISA STRONG VIOLATION: $\tau_c = \beta_c \ \ \forall \ c$									
OLS	-0.066	0.081	0.046	0.696		-0.066	0.068	0.015	0.005	
DAHL	-0.104	0.115	0.050	0.446	527	-0.146	0.147	0.018	0.000	5271
FULL	-0.011	0.103	0.102	0.953		-0.005	0.046	0.046	0.943	
LASSO	-0.022	0.072	0.069	0.930		0.004	0.032	0.032	0.944	
	SISA WEAK VIOLATION: $\tau_j \neq 1, \tau_c = 1 \ \ \forall \ c \neq j$									
OLS	-0.041	0.061	0.045	0.866		-0.042	0.044	0.015	0.182	
DAHL	-0.074	0.088	0.049	0.697		-0.112	0.114	0.018	0.000	
FULL	0.004	0.101	0.101	0.958	530	-0.002	0.046	0.046	0.949	5300
LASSO	-0.002	0.068	0.068	0.950		0.003	0.032	0.032	0.951	
PARTIAL	0.008	0.068	0.068	0.949		0.005	0.029	0.028	0.942	

Notes N is the number of individuals 'born' in each location. 2000 replications are performed for each specification. 'DAHL' refers to the inclusion of a cubic in the first best probability. 'FULL' includes cubic interactions in all probabilities. 'LASSO' selects terms using post-double-Lasso. 'PARTIAL' implements a cubic in the first best probability and in the probability corresponding to the location with a unique value of τ_c. For each specification, mean bias, root mean-square-error, the standard deviation, coverage probabilities and the average sample size is reported. To estimate coverage probabilities (Cov.), I calculate 95% confidence intervals for each replication and calculate the proportion of replications where the true value ($\beta_c = 1$) falls within the confidence band. Bootstrapped standard errors are used to construct confidence bands for 'FULL', 'DAHL', and 'PARTIAL' specifications.

methods having rates of 92%. It is also interesting to note here that the Lasso results fall somewhere between the partial and full specification. In particular Lasso still includes extra terms which explain the increased variability of estimates relative to the partial model, but the standard deviation of estimates is still lower relative to the full specification.

There are clear implications to be drawn from Table 1.1. Specifically, when the SISA holds, Dahl's method is clearly the preferred correction method. It achieves the greatest bias reduction and the lowest RMSE. However, the Lasso method also performs well in terms of bias reduction and both methods yield appropriate coverage probabilities. The cost of using Lasso in this case is in the increased variability of estimates which is between 1.3 and 1.5 times higher. However, when the SISA is violated, Lasso is clearly preferred to Dahl's method which at best removes only some of the selection bias, and can actually mischaracterize the pattern of selection bias entirely. In both of these cases, Dahl's method performs poorly in terms of coverage rates which means inference based on these estimates is not valid. Lasso performs similarly to the full model in all specifications, achieving lower estimation variability in all specifications. It performs better in terms of bias reduction when the SISA holds, or is weakly violated, and slightly worse when the SISA is strongly violated. In general then, the PDL method can be used when the full specification is not viable. In fact, given the decreased variability of estimates it is preferred even when the full model can indeed be estimated. Furthermore, the Lasso specification ensures valid confidence bands and inference procedures regardless of whether the SISA holds. As researchers cannot observe whether it holds in practice, this suggests Lasso is the preferred specification across all alternatives. The price paid to ensure inference is valid is increased variability of estimates in the case that the SISA does in fact hold.

10 locations

In Table 1.2, I consider the case of 10 locations. In increasing the number of locations, I examine the performance of PDL in settings of high dimensionality. Largely the results of table 1 are confirmed here. When the SISA holds Dahl's method once again achieves the best performance in terms of bias reduction and RMSE. It removes 91% of bias in the large sample, compared to just 75% for the Lasso specification. However, as before, both specifications have reasonable coverage rates. Dahl's method does do better here with a coverage of 92%, but Lasso has a reasonable coverage rate of 88%. When the SISA is strongly violated I find that both Lasso and the full specification remove around 94% of bias in the large sample case, while Dahl increases the size of the negative bias. Through variable selection, Lasso reduces variability in estimates by around 30% while achieving the same degree of bias reduction as the fully flexible specification.

In the case of a weak violation, I set the return to unobserved ability equal to the return to schooling for three locations. Once again, I include the partial specification as a benchmark.

Table 1.3: Monte Carlo Output: 3 Locations

	Bias	RMSE	Std. Dev.	Cov.	Av. Samp.
	SISA HOLDS: $\tau_c = 1 \ \ \forall \ c$				
	(a) $N = 1000$				
OLS	0.087	0.093	0.032	0.226	
DAHL	0.030	0.048	0.038	0.869	1020
FULL	0.031	0.050	0.040	0.868	
LASSO	0.029	0.048	0.038	0.880	
	(b) $N = 10000$				
OLS	0.088	0.088	0.010	0.000	
DAHL	0.006	0.014	0.013	0.914	10205
FULL	0.008	0.017	0.016	0.905	
LASSO	0.000	0.014	0.014	0.930	
	SISA STRONG VIOLATION: $\tau_c = \beta_c \ \ \forall \ c$				
	(a) $N = 1000$				
OLS	0.127	0.130	0.030	0.011	
DAHL	0.046	0.058	0.036	0.733	1030
FULL	0.047	0.060	0.037	0.740	
LASSO	0.045	0.057	0.036	0.750	
	(b) $N = 10000$				
OLS	0.125	0.126	0.009	0.000	
DAHL	0.010	0.015	0.012	0.880	10304
FULL	0.012	0.018	0.014	0.857	
LASSO	-0.001	0.013	0.013	0.937	

Notes N is the number of individuals 'born' in each location. 2000 replications are performed for each specification. 'DAHL' refers to the inclusion of a cubic in the first best probability. 'FULL' includes cubic interactions in all probabilities. 'LASSO' selects terms using post-double-Lasso. 'PARTIAL' implements a cubic in the first best probability and in the probability corresponding to the location with a unique value of τ_c. For each specification, mean bias, root mean-square-error, the standard deviation, coverage probabilities and the average sample size is reported. To estimate coverage probabilities (Cov.), I calculate 95% confidence intervals for each replication and calculate the proportion of replications where the true value ($\beta_c = 1$) falls within the confidence band. Bootstrapped standard errors are used to construct confidence bands for 'FULL', 'DAHL', and 'PARTIAL' specifications.

Both the full and Lasso specifications here perform as well in terms of bias reduction as the partial specification, and confidence bands are well defined as can be seen from the coverage rates. Lasso has slightly higher estimation variability than the partial model, but notably less than the full specification. Dahl's method in contrast performs poorly here, increasing the downward bias in estimates. Furthermore, the true value never lies within the 95% confidence band in any of the 2000 specifications as estimates are precisely estimated around the incorrect value. The results presented here are well summarised visually in Figures A.5 - A.7

3 locations

Finally, for three locations, I consider the case where the SISA holds or is strongly violated. The results presented in Table 1.3 show that all three methods tend to perform well here. There is little increased variability of estimates when including additional terms, and all three methods tend to perform well in terms of bias reduction. Lasso performs better in terms of bias reduction when the SISA is strongly violated, which means it has better rates of coverage. This example shows that the gains to using PDL are greater when considering higher dimensional problems. In higher dimensional settings the gains to using PDL to select terms are clear: Dahl's method can perform very poorly when the SISA holds, whereas PDL does well in terms of bias reduction regardless of whether the SISA holds. These results are presented visually in Figures A.8 - A.9.

Cross Validation

Thus far I have implemented the post-double-Lasso method applying a bound on the Lasso threshold established by Belloni et al. (2011). An alternative is to use 10 fold cross-validation to choose the penalty level which minimizes model error. This method segments the data into training and testing data sets, specifically 10 alternative cuts of the data is made and the model is trained on each 'fold'. The testing data is then used to verify the model fit by predicting outcomes out of sample. The performance is averaged across the 10 folds, and the Mean Square Error (MSE) is compared across alternative threshold values. The threshold that either minimizes model error, or minimizes MSE to within one standard deviation of the minimum is selected.

The results for 5 locations are presented in Table 1.4. In comparing these results to those in Table 1.1, the first thing to note is that the standard deviation of estimates is markedly higher for the cross validated estimates. Specifically, the standard deviation of these estimates tends to match the full specification very closely. This suggests that using cross-validation tends to result in over-selection of terms such that no real improvement is made relative to the full model. This is particularly noticeable in the case of a weak violation where the bias reduction obtained using CV estimates is closer to that of the full specification than Lasso estimates from Table

Table 1.4: Monte Carlo Output: 5 Locations, 10-fold Cross Validated Penalty Level

	Bias	RMSE	Std. Dev.	Cov.	Av. Samp.
	SISA HOLDS: $\tau_c = 1 \ \forall \ c$				
	(a) $N = 1000$				
CV(MSE + 1sd)	0.049	0.085	0.070	0.885	608
CV(MSE)	0.049	0.087	0.072	0.902	608
	(a) $N = 10000$				
CV(MSE + 1sd)	0.015	0.034	0.031	0.940	6079
CV(MSE)	0.014	0.035	0.032	0.948	6079
	SISA STRONG VIOLATION: $\tau_c \neq \tau_j \ \ \forall \ c, j$				
	(a) $N = 1000$				
CV(MSE + 1sd)	0.007	0.069	0.069	0.946	604
CV(MSE)	0.008	0.072	0.072	0.951	604
	(a) $N = 10000$				
CV(MSE + 1sd)	0.002	0.032	0.032	0.947	6031
CV(MSE)	0.003	0.032	0.032	0.964	6031
	SISA WEAK VIOLATION: $\tau_j \neq 1, \tau_c = 1 \ \ \forall \ c \neq j$				
	(a) $N = 1000$				
CV(MSE + 1sd)	0.075	0.100	0.066	0.799	611
CV(MSE)	0.076	0.102	0.068	0.801	611
	(a) $N = 10000$				
CV(MSE + 1sd)	0.025	0.044	0.036	0.882	6102
CV(MSE)	0.025	0.044	0.036	0.806	6102

Notes: N is the number of individuals 'born' in each location. 2000 replications are performed for each specification. 10 fold cross validation is used to choose the penalty level. 'CV(MSE + 1sd)' corresponds to the penalty level that achieves a mean-squared error of within one standard deviation of the minimum. 'CV(MSE)'corresponds to the penalty level that achieves the minimum mean-squared error. To estimate coverage probabilities, I calculate 95% confidence intervals for each replication and calculate the proportion of replications where the true value ($\beta_c = 1$) falls within the confidence band.

1.1. In general then, cross-validation tends to over-select terms and imposes the additional cost of excess computational burden. The results in table Table 1.4 indicate that the ideal penalty factor suggested by Belloni et al. (2011) tends to perform better and with less computational expense than results obtained using cross-validation.

Alternative Methods

There are of course alternative methods which could be used instead of PDL, either as a means of variable selection, or as a means of dimensionality reduction. I consider three alternatives. The first method I examine is that of stepwise regression. As its name suggests, this algorithm proceeds in discrete steps, at each step considering which term would best improve model fit if included. Similarly, terms are removed from the model if this improves model fit. The drawback of stepwise regression in this context is in its inability to select terms based on their correlation with the explanatory variables.[23] A second alternative used is Principal Components Analysis (PCA). PCA creates an orthogonally transformed set of variables (principal components) from inputted variables. Dimensionality is therefore reduced by reducing the set of terms included in the final regression model. In comparison to stepwise regression however, no terms are dropped from consideration, the data is merely transformed into a reduced number of variable components. Rather than dropping specific terms, the variation contained in some principal components is excluded. Furthermore, PCA does not use the outcome variable to train the model and to derive the principal components. Therefore, PCA serves as a method of dimensionality reduction, but does not take into account model fit. Finally, Partial Least Squares Regression (PLSR) is similar to PCA but accounts for the outcome variable in deriving components. Specifically, PLSR creates components of the explanatory variables that are related to the outcome variable by simultaneously decomposing both Y and X variables and maximizing the covariance between X and Y explained by the components.

In Table 1.5, I present results from the Monte Carlo experiment using each of these methods. I use CV to select the number of components to include in PCA and PLSR, and I consider 4 alternative variables on which to measure model fit in the stepwise regression algorithm. I consider the performance of each model using just the bias estimate and RMSE. There is little literature informing the appropriate calculation of standard errors when using these methods. In focusing on the large sample results, it is clear than PLSR tends to perform worse in terms of bias reduction across all three cases. The remaining bias tends to be 2-3 times larger following a PLSR correction than PDL. PCA performs better in terms of bias reduction relative to PLSR, and only marginally worse than PDL. However, it performs worse in terms of RMSE across

[23] I also found stepwise regression to be extremely slow to run given that an exhaustive search over all terms included and excluded from the model is performed at each step.

Table 1.5: Monte Carlo Output, 5 Locations: Alternative Methods

	$\tau_c = 1$		$\tau_c = \beta_c$		$\tau_1 \neq 1$	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
	(a) $N = 1000$					
PCA	0.049	0.088	-0.006	0.070	0.075	0.101
PLS	0.018	0.075	-0.036	0.077	0.047	0.082
Step(SSE)	0.076	0.099	-0.001	0.058	0.107	0.127
Step(AIC)	0.062	0.094	-0.002	0.063	0.088	0.111
Step(BIC)	0.083	0.099	-0.009	0.057	0.124	0.137
Step($adjR^2$)	0.052	0.089	-0.005	0.068	0.078	0.103
LASSO	0.047	0.082	-0.010	0.066	0.073	0.097
	(b) $N = 10000$					
PCA	0.006	0.031	-0.010	0.031	0.013	0.032
PLS	0.017	0.044	-0.015	0.036	0.031	0.060
Step(SSE)	0.011	0.033	0.002	0.030	0.012	0.029
Step(AIC)	0.006	0.030	-0.005	0.030	0.011	0.029
Step(BIC)	0.027	0.044	0.012	0.031	0.016	0.034
Step($adjR^2$)	0.005	0.030	-0.009	0.031	0.011	0.030
LASSO	0.005	0.027	-0.007	0.028	0.011	0.029

Notes: N is the number of individuals 'born' in each location. 2000 replications are performed for each specification. For PLS and PCA the number of components is selected using 10-fold cross validation. SSE, AIC, BIC, $adjR^2$ refer to the statistic used to select terms. Respectively, mean-squared error, the Akaike information criterion, the Bayesian information criterion, the adjusted R-squared are used. Columns 2-3 present mean bias and root mean-square-error when the SISA holds. Columns 4-5 present the same statistics when the SISA is strongly violated, and columns 6-7 present results for the case when the SISA is weakly violated.

all specifications. Stepwise similarly performs reasonably well (at least when using the Akaike Information Criterion (AIC) or the adjusted R^2 as a means of model testing), but is still largely bested by PDL. There is some evidence then that stepwise regression or PCA could perform well in the large sample case. However, PDL tends to outperform all methods, is easiest to implement, and is fastest computationally. Moreover, stepwise regression tends to perform poorly in small samples when the SISA holds or is only weakly violated.

Separate Control Functions by Birth Location

As discussed above, one restriction I impose on the model is on the inclusion of separate control functions by birth location. This restriction is necessary when migration flows are relatively small and where there are few migrations from certain locations. Here, I examine the performance of PDL and alternative methods when this assumption is violated. In this case, I allow for correlation across birth locations for taste shocks (ϵ_{ijc}). I draw these error terms from a multivariate normal distribution, with correlation across the j birth locations. The results obtained using this specification are presented in Table 1.6. I consider the case of 5 locations for 10,000 individuals born in each location and consider only the instances when the SISA holds or when it is strongly violated. Dahl 1, Full 1, and Lasso 1 refer to results obtained using a single control function, and Dahl 2, Full 2 and Lasso 2, use separate control functions for each of the five birth locations.

The upper panel considers results when the SISA holds. The results here clearly indicate that the use of separate control functions for each location are necessary to reduce bias close to zero. All three methods perform poorly when using just one control function. Dahl's method clearly performs best when using multiple control functions. The increased variability of estimates obtained using Lasso 2 and Full 2 is notable. In fact, this increase is so large that even though bias is reduced, RMSE increases when including multiple control functions. Coverage rates are valid here for the more flexible specifications.

In turning to the case when the SISA is strongly violated, Dahl's method performs poorly in both instances, but both the full specification and PDL perform reasonably well when considering just one control function. Using multiple control functions does little to improve bias reduction and more than doubles the standard deviation of estimates and hence RMSE. Coverage is improved due to the increased variability of estimates, but I obtain reasonable confidence intervals using the restricted model.

In general then, when the SISA holds and separate control functions are required for each birth location there is a cost to assuming a single control function. However, the increased variability of estimates is so great that it may be impossible to infer anything of interest from the empirical results. When the SISA is strongly violated, it is possible that the benefits to using

Table 1.6: Monte Carlo Output: 5 Locations, Multiple Control Functions

	Bias	RMSE	Std. Dev.	Cov.	Av. Samp.
	SISA HOLDS: $\tau_c = 1 \ \ \forall \ c$				
OLS	0.142	0.143	0.016	0.000	
Dahl 1	-0.042	0.046	0.020	0.414	
Full 1	0.062	0.077	0.045	0.723	
LASSO 1	0.060	0.069	0.034	0.552	5048
Dahl 2	0.021	0.034	0.026	0.872	
Full 2	0.040	0.106	0.098	0.938	
LASSO 2	0.035	0.084	0.077	0.930	
	SISA STRONG VIOLATION: $\tau_c \neq \tau_j \ \ \forall \ c, j$				
OLS	-0.052	0.055	0.016	0.090	
Dahl 1	-0.202	0.203	0.019	0.000	
Full 1	-0.022	0.051	0.047	0.905	
LASSO 1	-0.019	0.038	0.033	0.903	5058
Dahl 2	-0.120	0.122	0.025	0.003	
Full 2	-0.018	0.100	0.098	0.950	
LASSO 2	-0.024	0.082	0.078	0.935	

Notes: 10,000 individuals 'born' in each location. 2000 replications are performed for each specification. 'Dahl 1', 'Full 1' and 'LASSO 1' refer to the use of a single control function for each birth state. 'Dahl 2', 'Full 2' and 'LASSO 2' refer to the use of a separate control function for each birth state. To estimate coverage probabilities, I calculate 95% confidence intervals for each replication and calculate the proportion of replications where the true value ($\beta_c = 1$) falls within the confidence band. Taste shocks are drawn from a multivariate normal with correction across j birth states necessitating the need for a separate control function for each birth state.

separate control functions are outweighed by the costs. Results obtained using a single control function do reasonably well here in terms of bias reduction and coverage.

1.5 Estimating Migration Probabilities

Empirical implementation of any of the selection correction methods discussed above relies on the estimation of migration probabilities. As these migration probabilities summarize the information in the unobserved sub-utility functions, their estimation is of great importance for identification. Typical approaches have relied on imposing strict assumptions on the error draws of the utility equations, usually by estimating a conditional logit model which imposes the undesirable independence of irrelevant alternatives.[24] Dahl (2002) employs a non-parametric method which relies on fully saturating the state-space and separating individuals into mutually exclusive cells.

Non-parametric estimation has clear benefits, but there are accompanying downsides to this approach. As mentioned, one clear benefit is that researchers do not have to model utility by assuming some functional relationship, or have to decide which variables to include to proxy for location specific amenities. The main accompanying drawback to non-parametric estimation is in the choices faced by the researcher in determining which variables to use to group individuals, and how fine to make cells. If cells are too fine, researchers risk measurement error. Researchers are also restricted to consider variation coming from only a few characteristics given data limitations. Finally, it is often necessary to discretize continuous variables, or aggregate groupings of categorical variables, therefore limiting the useful variation inherent in these variables.

An alternative non-parametric approach is to use machine learning algorithms to estimate migration probabilities. In a similar Roy model context Ransom (2016) makes some progress along this dimension. Specifically, he uses a classification tree algorithm to estimate cell probabilities following Hothorn et al. (2006). Furthermore, he presents a numerical experiment similar in nature to that from Section 3 to show the efficacy of this approach. The benefits outlined above apply to this classification tree algorithm, however, there are alternative machine learning algorithms which tend to perform better, both in terms of classification, and in probability estimation. In particular, it is well known that estimation using a single classification tree tends to over-fit the data and hence produce biased probability estimates. This over-fitting comes from the algorithm being 'greedy'. This results from the recursive nature of the algorithm where the sample is split in successive steps. Despite this, use of the classification tree in Ransom (2016) represents a significant improvement over a cell-based approach.

[24] See Kennan and Walker (2011), or Davies et al. (2001).

Here, I consider an alternative machine learning algorithm to that used by Ransom, specifically, the Random Forest (RF) algorithm. As shown by Niculescu-Mizil and Caruana (2005) RF tends to do a good job of predicting probabilities across a range of generated data sets. Of particular note, the RF algorithm performs much better than probability estimates from a single decision tree. In this section, I discuss the RF algorithm and compare probability estimates obtained using a RF to a naive non-parametric method which segments individuals based on observable characteristics. Through a simple Monte Carlo illustration, I show that the RF method represents a significant improvement over more ad hoc methods of probability estimation. It is important to note however, that the core intuition underlying identification is the same, regardless of the method used to estimate migration probabilities. What changes is the variation across probabilities and the degree to which RF estimation allows for more accurate estimation of migration patterns.

1.5.1 Discussion of Random Forest

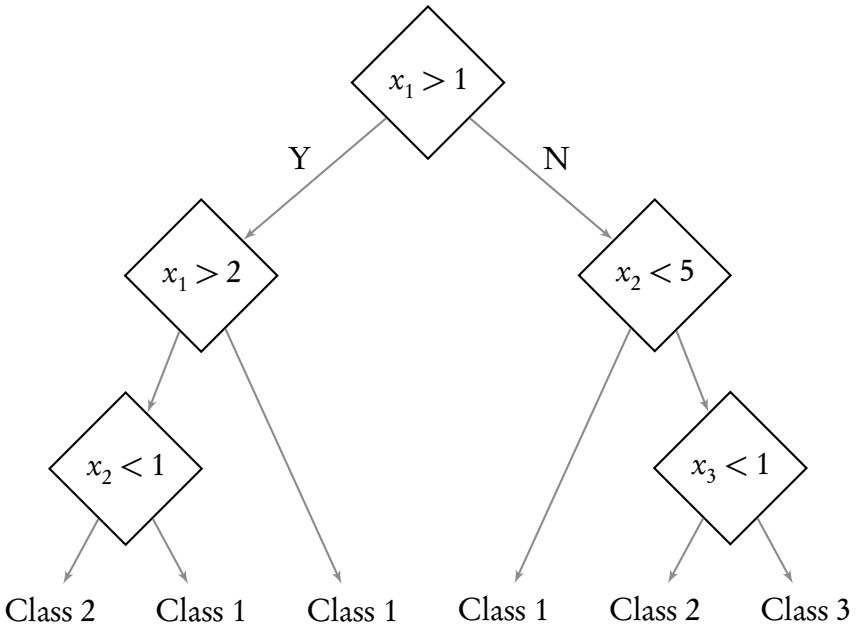
Before outlining the Random Forest algorithm, it will be instructive to discuss how to estimate a single classification tree.

Classification Decision Trees

Classification trees are estimated by recursively partitioning the data so as to predict a class label (an outcome) from covariates. Put simply, successive splits are made on X variables to group individuals into bins associated with a value of Y. If Y is a binary variable then each bin is associated with a prediction over 0 or 1. Regression trees estimate the outcome association with the set of covariates. An example of a classification tree is shown in Figure 1.1. In this example the researcher would like to determine how a set of x variables affects individuals choice of (or assignment into) three possible classifications. Beginning at the top of the diagram, covariates are split such that they are assigned to outputs given by the class labels at the bottom of the tree. Inputs can be discrete, in which case they are split in two (if split at all), categorical, or continuous. In estimating the classification tree, splits are made recursively, meaning successive splits are made in order. Each 'leaf' of the tree then represents a posterior probability distribution where the probability of the outcome or class label, conditional on belonging to that leaf, is estimated by calculating the fraction of individuals in the leaf who are observed in each outcome. Note that in this tree cuts are made on variables x_1, x_2 and x_3, but this does not mean that other variables were not considered by the researcher. Additional variables could have been included as predictors by the researcher but assigned no weight in classification by the algorithm.

The tree is estimated by iterating on two steps. The first step involves selecting the covariate

Figure 1.1: Classification Decision Tree with 5 nodes and 6 leaves



on which to form a node split. The second step then splits this node. In the selection step the variable with the strongest association between x and the outcome variable y is chosen from among the set of covariates X.[25] In the second step, the algorithm chooses from among different subsets of this selected variable, and makes the split that creates the most distinct bivariate distribution. Whether a split takes place is based on 'impurity' and 'node error'. Node purity measures the degree to which members in the node following the split belong to the same class. A 'pure' node then, is one where all individuals have the same class label. A typical measure of node purity is the gini index. Node error calculates the proportion of misclassified classes at a node. The split of the variable is therefore made to maximize node purity and minimize node error. These are the basic criteria on which selection and splitting are performed, although many algorithms introduce additional tests prior to splits being made.

It is necessary to choose hyper-parameters when estimating a tree just as it is when choosing the number of nodes in a Neural Network for instance, or in setting the value of the penalty in Lasso. In particular, researchers must select the number of splits or the minimum size of a leaf. In general it is optimal to cross-validate to select these parameters using hold-out samples to predict classification error. Classification trees make successive cuts until they meet one or more of the stopping criteria embedded into the algorithm. In general, these criteria are such that no successive split sufficiently aids the predictive accuracy of the model, or any successive split would require leaves below the set minimum size.

Random Forest

One limitation of classification decision trees is that they tend to overfit the data. That is, trees tend to grow large and leaves tend to be small, leading to large misclassification errors out of sample. It is possible to solve this overfitting problem by 'pruning' the tree. Pruning, tends to improve the fit of classification trees, whereby limiting the number of splits, or setting a minimum size for leaves, improves prediction. However, pruning the tree only deals with measurement error coming from having a small sample size in the leaves. It does not deal with over-fitting coming from the capturing of noise in the data. That is, individuals may be split in a manner consistent with the data but which is based purely on sampling variability. Another issue with decision trees is that they are 'greedy' in the sense that in selecting variables on which to make cuts, they choose the variable which leads to the greatest node purity. However, a more accurate representation of the data might make the cut on this variable only after another cut has been made. As an example, consider the case where both education and location affect income. A cut on education may be prioritised by the algorithm, whereas the effect of education may

[25] The choice of measure on which to base the association between x and y is important here as some measures have biased preferences towards continuous variables.

be more accurately represented differently by location. The cut on location should be made before the cut on education and this ordering may affect cuts later in the tree.

The RF algorithm introduced by Breiman (2001) grows many classification trees, with the final prediction over classification being driven by the average prediction across trees. Growing similar trees is not fruitful, and hence trees are randomized by 'bagging' (Denil et al. (2014)). Bagging involves training each tree on a random subset of training samples using bootstrap re-sampling. Going one step further, the RF algorithm also randomizes the subset of features over which the algorithm chooses to make splits at each node. This 'bagging over the feature space' is beneficial once again to prevent overfitting to the training sample. By randomizing selection over the feature space, individual trees are no longer highly correlated. This is a limitation of the bagged tree approach given that decision trees are estimated using a greedy algorithm. The hyper-parameters necessary to parametrize the model are the number of trees to be grown, and 'mtry', the size of the set of features drawn over which the algorithm selects at each node. With these extensions each tree is estimated using the same method as above for a single decision tree.

In general, though the method of probability estimation is relatively simple, the RF algorithm tends to predict probabilities well. Niculescu-Mizil and Caruana (2005) show for instance that probabilities estimated using a RF largely outperform those estimated using a single decision tree. Bagged decision trees also perform well, as do probabilities estimated using a Neural Network. One limitation of probabilities estimated using a RF method, is that they tend to bias estimation away from the tails of the distribution. Several methods to transform the probability distribution to take account of this have been suggested in the literature.[26] However, it is worth noting that it is possible to show this feature post estimation in the probability estimates, and such a transformation is not always required. That is, this bias is data dependent. In Niculescu-Mizil and Caruana (2005), Platt scaling or an isotonic transformation have little effect on the accuracy of RF estimates, compared to other methods. Furthermore, this issue is less of a problem with many classes rather than in a simple two class problem. With many classes, the probabilities may still be biased away from 0 and 1, but this bias will likely be very small given the range of choices.

[26] Platt scaling and isotonic regression are the two most popular methods here (Boström, 2008), although both are parametric in nature. Furthermore, these transformations are more difficult in multi-class problems. One compelling method to transform the data is to feed the RF predicted probabilities through a Neural Network using a sigmoidal transfer function.

1.5.2 Illustrative Example

To outline the benefits of using the RF algorithm to estimate migration probabilities, I consider a simple Monte Carlo experiment where the data is generated as follows:

$$V_{ij} = \beta_{1j} x_{1i} + \beta_{2j} x_{2i} + \beta_{3j} x_{3i} + u_{ij}$$

and

$$D_{ij} = 1 \text{ if and only if } V_{ij} \geq V_{im} \quad \forall m$$

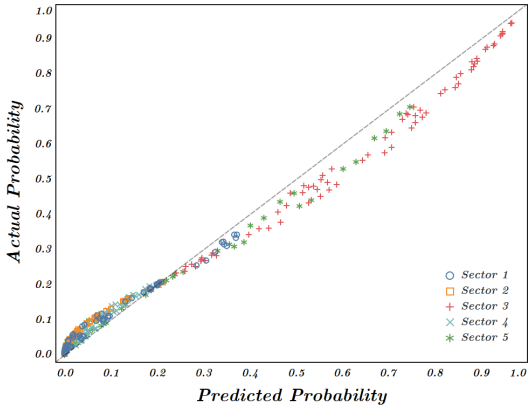
where D_{ij} is a dummy variable which takes value 1 if the individual chooses j. Utility for individual i of a given choice j depends on u_{ij} which is a draw from a Gumbel distribution, and on three variables. x_{1i} and x_{2i} take integer values between 0 and 5, and x_{3i} is a dummy variable. There are 72 possible combinations of 'type' given by these x variables and in the experiment I draw equal numbers of each 'type' of individual.[27]

Using this data, I create cells of each 'type' and calculate cell probabilities. I run the RF algorithm using the default number of candidate splits at each node ($\sqrt{p}$, where p is the number of inputs). As is standard, I make no restrictions on tree depth or leaf size. To present an additional alternative I consider probably estimates obtained using an Artificial Neural Network (ANN). Given the focus on RF estimates in my empirical application, and for brevity, I leave discussion of the ANN to appendix A. For ANN estimation, I consider both an ANN with a single hidden layer, and another with two hidden layers. In both cases, I use 10-fold CV to determine the number of hidden nodes.

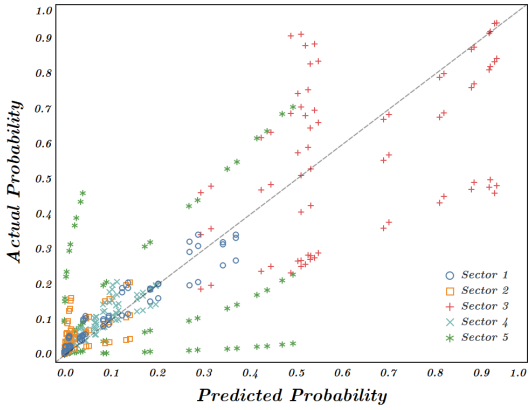
As the model errors are drawn from a Gumbel distribution, the true probability of making each choice is known. For the non-parametric cell approach and each ML method, I compare the predicted probabilities to the true values, averaging across observations for each individual type, and calculating the RMSE. The hypothesised benefits to ML methods in estimating cell probabilities is in preventing arbitrary choices on the part of the researcher, and in balancing the noise to signal in the data such that it prevents the formation of small cells. In models 1 and 2 below, I compare methods along the first dimension, and in the third model I consider the importance of sample size.

[27] Given the focus on discrete variables here it is worth noting that the random forest algorithm is likely to perform better in this setting than when continuous variables are used in modelling utility preferences. However, given the selection model presented here, and the need for a separate exclusion restriction for each choice probability, a reliance on a large set of discrete variables is more likely in practice. In chapter 2 of this thesis choice probabilities are estimated using mostly discrete variables.

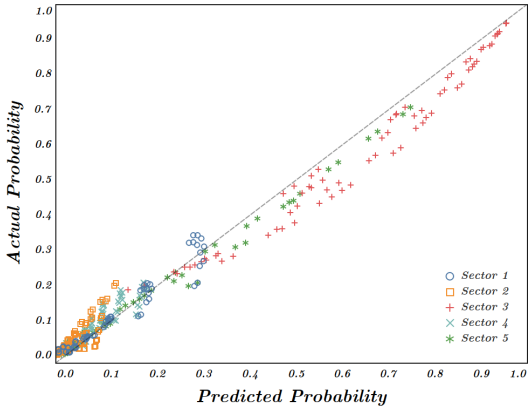
Figure 1.2: Model 1: Actual versus Estimated Probabilities



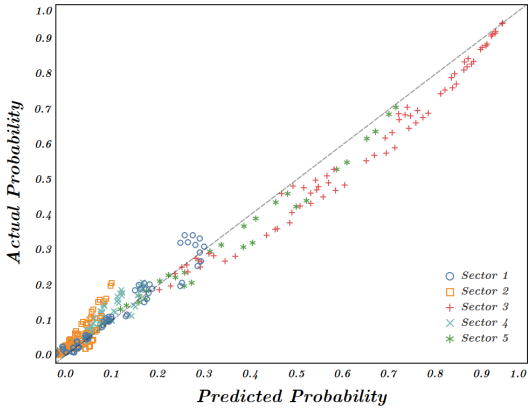
Cell Correct



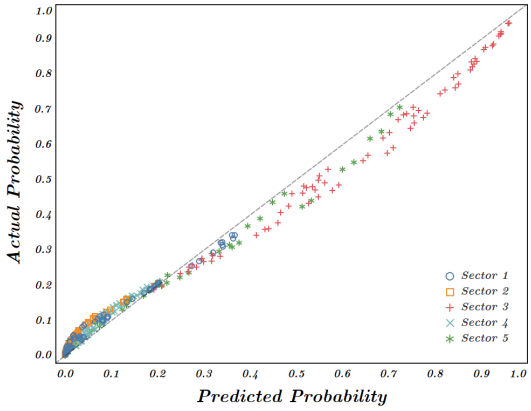
Cell Incorrect



ANN 1



ANN 2



Random Forest

Notes: True probabilities are plotted against probability estimates obtained using either machine learning tools, or grouping individuals into cells. Data is generated and estimates obtained as described for Model 1. I average probabilities for each 'type' of individual (based on x variable draws) in each location across replications. Each point represents the average estimate for a particular 'type' of individual in a given location.

Model 1

The first model is based on the data generating process as outlined above. I consider two alternative means of calculating cell probabilities. In the first case, the 'correct' cell breakdown is used whereby individuals are separated into 72 cells representing the full set of potential combinations of x values. As an alternative, I consider an 'incorrect' cell breakdown where the researcher aggregates x_1 into just two categories, therefore estimating probabilities for just 24 cells. The efficacy of the RF and ANN approaches here can then be judged against both of these cell breakdowns. To the degree that they are comparable with the 'correct' cell approach, I can see how these methods perform on a par with the best potential choices made by the researcher. Alternatively, when the researcher makes a choice not in line with the data generating process, I am interested in whether machine learning techniques lead to better performance.

The results are presented in the first panel of table Table 1.7. Additionally, the results are presented graphically in Figure 1.2. In comparing the correct cell approach to the incorrect method, I see a clear increase in estimation error for locations 3 and 5 in particular. In both cases, RMSE more than triples. Both ANN and RF methods perform well here, and indeed in many instances outperform even the best cell approach. Probabilities estimated using the RF method are closer to the true probabilities in every location than are those estimated using the best non-parametric approach which groups individuals into cells. ANN estimates also perform better than the cell approach in all locations except location 2. In general they perform slightly worse than the RF estimates, except in location 4, where ANN 1 does slightly better.

In this simple experiment it is clear that machine learning methods perform at least as well, if not better than the standard non-parametric approach when the cells are split in a manner consistent with the underlying data generating model and there are no concerns in terms of sample size.

Model 2

In model 1, I see that the ML methods for probability estimation presented here perform as well as the best non-parametric cell-splitting process. However, in this alternative specification, I examine the ability of ML methods to estimate probabilities when utility is based on transformations of the raw variables. If ML tools perform poorly along this dimension, then there is the possibility that researchers making decisions on how to split the sample is the best method for probability estimation.

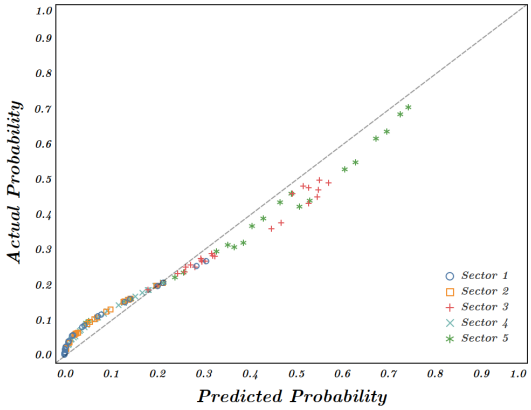
In this second experiment, I consider the case where the utility function uses a modified version of x_{1i}, $\hat{x}_{1i}$. $\hat{x}_{1i}$ is a dummy variable which takes value 0 for values of x_{1i} less than 2, and 1 otherwise. As before, I consider two possible cell breakdowns. In the 'correct' case, I separate individuals into 24 cells which present all possible combinations of $\hat{x}_1$, x_2 and x_3. I also consider

Table 1.7: Estimated versus Actual Probability Estimates: A Comparison of Approaches for each Data Generating Process

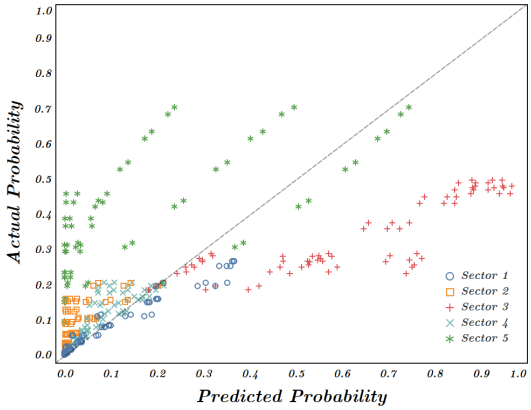
	P_1	P_2	P_3	P_4	P_5
			Model 1		
Correct Cell	0.035	0.032	0.074	0.036	0.043
Incorrect Cell	0.038	0.050	0.227	0.039	0.203
ANN 1 HL	0.035	0.038	0.073	0.033	0.034
ANN 2 HL	0.045	0.039	0.079	0.036	0.041
Random Forest	0.030	0.031	0.070	0.032	0.038
			Model 2		
Correct Cell	0.031	0.033	0.057	0.033	0.054
Incorrect Cell	0.081	0.079	0.345	0.077	0.266
ANN 1 HL	0.025	0.034	0.064	0.033	0.067
ANN 2 HL	0.027	0.036	0.066	0.033	0.061
Random Forest	0.032	0.036	0.062	0.036	0.059
			Model 3		
Cell	0.072	0.036	0.110	0.065	0.049
ANN 1 HL	0.038	0.038	0.069	0.033	0.041
ANN 2 HL	0.053	0.038	0.086	0.036	0.044
Random Forest	0.055	0.032	0.091	0.051	0.042
Random Forest ($N > 25$)	0.037	0.030	0.074	0.036	0.039
Random Forest ($N > 50$)	0.032	0.029	0.069	0.032	0.039

Notes: Data is generated for three separate models as described in text. Simulated agents optimize over 5 locations. For each 'type' of individual 100 observations are created and probabilities estimated for each 'type.' 500 replications are performed for each model and the Root Mean Square Error is calculated using actual and estimated probabilities for each replication. Presented in this table is the average RMSE across all 'types' of simulated individual. P_1-P_5 correspond to the estimates associated with each of these 5 locations. ANN 1 HL represents results from an ANN with 1 hidden layer. ANN 2 HL represents results from an ANN with 2 hidden layers. For the RF algorithm, I set the number of predictors from which to choose ('mtry') equal to the square root of the number of predictors and I grow 250 trees.

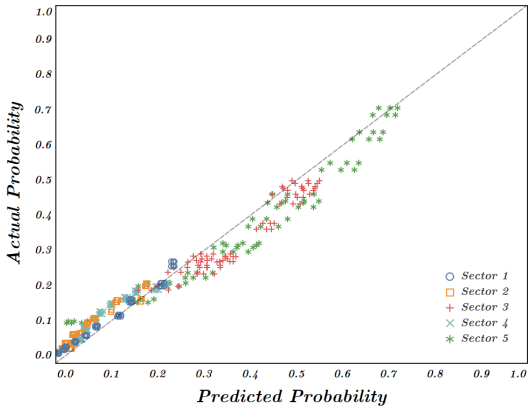
Figure 1.3: Model 2: Actual versus Estimated Probabilities



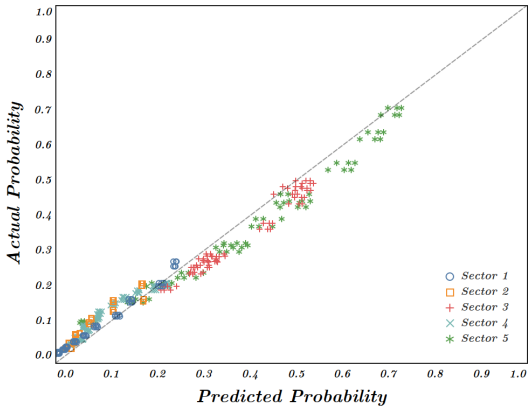
Cell Correct



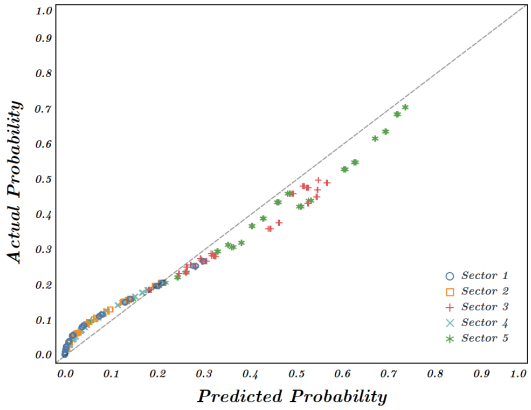
Cell Incorrect



ANN 1



ANN 2



Random Forest

Notes: True probabilities are plotted against probability estimates obtained using either machine learning tools, or grouping individuals into cells. Data is generated and estimates obtained as described for Model 2. I average probabilities for each 'type' of individual (based on x variable draws) in each location across replications. Each point represents the average estimate for a particular 'type' of individual in a given location.

estimates using an incorrect approach where the researcher assumes that utility is derived using x_1 rather than $\hat{x}_1$. In all ML methods here the raw variables x_1, x_2 and x_3 are used as inputs. The purpose of this model then is to test how ML methods perform when the true data generating process uses modified versions of the inputted variables.

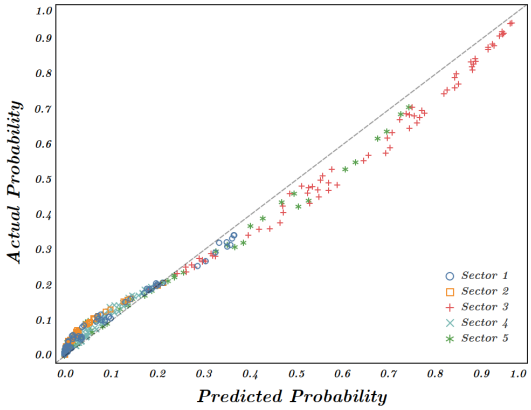
The results are presented in panel 2 of Table 1.7, and Figure 1.3. These results are again favourable to the ML methods considered here. In particular, all ML methods largely outperform the cell approach which creates cells based on the full set of potential outcomes. This suggests that ML algorithms can handle estimation when the true relationship uses a transformation of the inputted variables. Furthermore, all 3 approaches come remarkably close to matching the performance of the correct cell breakdown. ML methods therefore perform almost as well as the best potential method here. RF estimates perform marginally worse across the board, but represent a substantial correction over the incorrect cell approach. The same is true for ANN estimates. In comparing RF and ANN, the estimates are similar, with ANN performing best in 3 out of the 5 cases.

Model 3

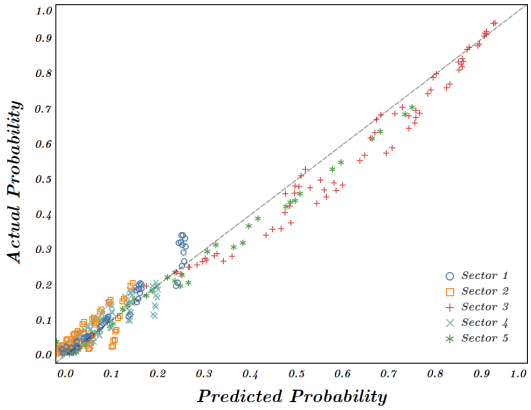
Another potential benefit of ML techniques here is in estimating more accurate probabilities than could be estimated using a non-parametric cell approach given the existence of small bins on which to base choice estimation. That is, I assume that there are indeed differences across small bins of individuals in their utility preferences, but I assess the ability of each method to estimate accurate probabilities given this small sample problem. Here, I compare the performance of ML methods to Dahl's non-parametric approach by arbitrarily drawing only a small number of observations for some cells. That is, for the full set of 72 potential cells in equation 8, I draw 100 observations for 36 of these cells, and just 10 observations for the remaining 36 cells. In the cell approach, I use the full set of 72 cells to estimate sectoral probabilities. ML methods have the potential to outperform this method then if they can use underlying similarities in choices across cells in such a way as to prevent excess cuts to the data.

The estimates in panel 3 of Table 1.7 and Figure 1.4 are clearly indicative of the benefits of ML estimation. In particular, both ANN and RF outperform the cell approach in all locations. ANN tends to perform better than RF here except in location 2. The RMSE in location 1, for instance, is two thirds that of the RF value. Though the default RF algorithm places no restriction on leaf size or tree depth, in rows 5 and 6 I restrict the minimum leaf size of cells. RF estimates represent a notable improvement over those estimated without the restriction. In particular, RF estimates now outperform ANN estimates in terms of RMSE in 3 out of 5 cases, and perform as well in the other 2 cases. The results here then indicate that in practical implementation it may be preferable to restrict the minimum leaf size in Random Forest esti-

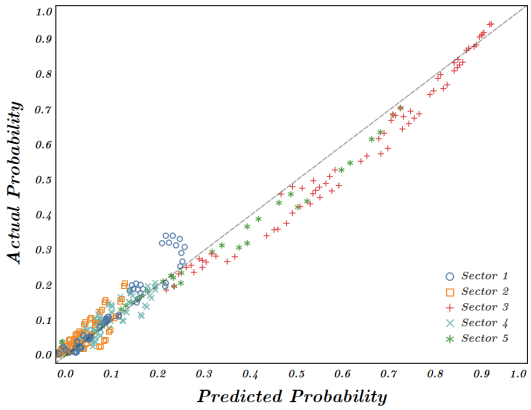
Figure 1.4: Model 3: Actual versus Estimated Probabilities



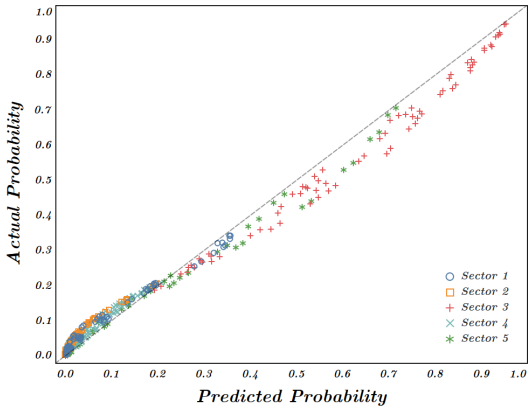
Cell Correct



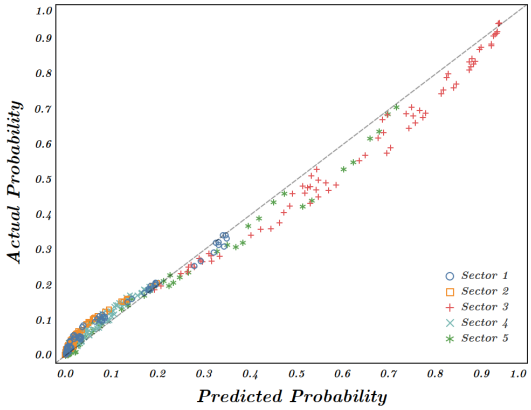
ANN 1



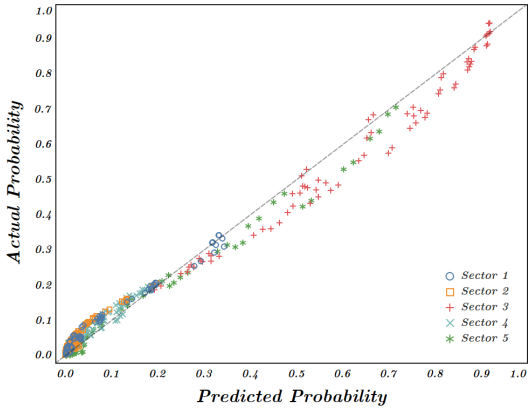
ANN 2



Random Forest



Random Forest ($N > 25$)



Random Forest ($N > 50$)

Notes: True probabilities are plotted against probability estimates obtained using either machine learning tools, or grouping individuals into cells. Data is generated and estimates obtained as described for Model 3. I average probabilities for each 'type' of individual (based on x variable draws) in each location across replications. Each point represents the average estimate for a particular 'type' of individual in a given location.

mation. Though the Random Forest algorithm is correct to make cuts at relatively low levels, the problem with doing so it that the class labels for these cells are poorly defined. When there is variation across small groupings of the data then, the RF algorithm will tend to detect these patterns of variation, but this source of variation is noisy and poorly estimated. Instead it is best to restrict the RF algorithm to prevent fitting to these narrow sources of variation.[28]

1.5.3 Discussion

This illustrative exercise shows that, in general, ML methods provide researchers with greater flexibility in the estimation of migration probabilities. Specifically, these methods prevent researchers from making choices that do not capture the true underlying data generating process, and instead perform as well as a non-parametric cell approach where the correct cuts along the input variables are made. Another important consideration here is that ML tools, allow for variables to affect utility in a continuous manner and hence will outperform the cell approach which requires the conversion of all continuous variables into discrete variables. The variation that drives identification using the cell approach is therefore much coarser than that utilized through ML tools.

The numerical exercise above then does not fully illustrate all potential benefits of ML methods. Indeed, ML methods allow for the inclusion of variables with potentially little explanatory power. These variables will be assigned a low weight in estimation and hence have little effect on probabilities. However, the inclusion of variables with little explanatory power in the cell approach could have potentially large negative effects on estimation as only a few choice variables can be included as the data must be fully saturated. The inclusion of redundant groupings introduces measurement error with no upside. In allowing for the inclusion of a greater number of terms, ML methods potentially allow for greater interactions between variables that drive the identification of selection and variables which enter the wage equation directly. In general then, RF and ANN estimation of cell probabilities is largely preferred to non-parametric cell based estimation.

1.6 Conclusion

This chapter introduces and discusses at length a means of correcting for selection bias that represents a distinct and clear improvement over traditional methods. Specifically, it establishes a more robust procedure to deal with selectivity in multi-choice settings, effectively overcoming the inherent dimensionality problem. It is arguably the first method to control for selection

[28] It is important to note here that this pruning does not equate the algorithm with the method of estimating a single decision tree. In particular, using a Random Forest but pruning the leaves still accounts for overfitting and correlation across trees which biases estimation and leads to poor out-of-sample prediction using decision trees.

bias where identification does not rely on strong distributional assumptions being imposed ex-ante. Instead selection bias is well captured if the control function governing selection is approximately sparse in the sense that it is well approximated by a small number of terms. I use machine learning tools to aid non-parametric estimation of migration probabilities which are central to identification, and to select terms relevant to selectivity. This framework is easy to implement, and be flexibly applied to a wide range of alternative applications.

Chapter 2

Re-estimating the Returns to Education

2.1 Introduction

This chapter implements an empirical application of the selection correction procedure outlined in Chapter 1. I estimate the returns to schooling across states using the 1990 U.S. census, mirroring the empirical exercise in Dahl (2002). This exercise serves to both provide new evidence of regional variation in earnings, specifically in the returns to educational attainment, and to compare estimates obtained using alternative means of selection correction. The results strongly indicate that the methodologies are not equivalent, in the sense that estimates obtained using each method are noticeably different.

The theoretical basis for estimation is provided using a generalized Roy model similar to that outlined in Chapter 1, where individuals choose in which state to live based on utility preferences, driven in part by earnings shocks, non-pecuniary benefits, and taste shocks. Individuals in this framework sort non-randomly across states based on variation in their relative preferences. Specifically, as in Dahl (2002) I take individuals' level of educational attainment as given, and model utility preferences over residential location, conditional on education. That is, selectivity is being driven by different utility preferences over states by level of education.

In comparing results obtained using Dahl's correction, and my improved method, I find that Dahl's method over- or under-estimates selection bias. I find in general that Dahl's method under-predicts the degree of bias in OLS estimates for the most and least educated individuals. For instance, in considering the returns to having an advanced degree, Dahl's estimate can understate the upward bias in OLS estimates by as much as 15%. Conversely, Dahl's method over-estimates the upward bias in OLS estimates of the returns to having a college degree. Despite these important differences, I confirm the upward bias in OLS estimates of the return to a college degree as found by Dahl (2002). The estimates obtained using this improved correc-

tion procedure, however, paint a more accurate picture of differences in returns to education across states. Additionally, in using these estimates to examine national level wage inequality I find evidence of noticeable upward bias in the national college premium due to sorting across states. Using my improved procedure I find that this bias varies between 9 and 12% over the period 1980-2000 and is overstated by Dahl's method which would suggest too low a level of wage inequality.

The remainder of this chapter is organised as follows: In section 2.2, I discuss the Census data used in estimation and the sample restrictions imposed. In section 2.3 I outline estimation of the migration probability terms using Dahl's method and RF estimation. I present estimates of the return to schooling using a variety of specifications in section 2.4 and in section 2.5 I explore the probability terms selected using the PDL methodology. Section 2.6 provides estimates of national level wage inequality, building upon the results in 2.4, while section 2.7 provides a test of the core implications of the Roy model. I conclude in section 2.8.

2.2 Data

The data I use for this analysis comes from the 5% public use sample of the 1990 US Census.[1] I limit the sample to white males aged 25-34 at the time of sampling, who report working in the previous year, earn at least 2000 dollars total income, work at least 10 weeks during the year, and work 20 hours on average. I drop individuals who live in group quarters, and those that attend school.[2] Individuals born outside the United States are dropped from the analysis as I consider mobility between birth state and residence state. Individuals' migration paths are summarised by movements from birth state j to residence state c.

Table 2.1 provides summary statistics for the sample and for 6 selected states.[3] The first takeaway from this table is that a large fraction, 35% of the US population of white males aged 25-34 in 1990, lived in a state that was not the state of their birth. There is clear variation across states in the proportion of residents born elsewhere. This figure is remarkably high for Florida at 71%, and lowest for New York at just 15%. There is greater similarity across the fraction of individuals born in the state who now live elsewhere. The range varies from 24% for Texas to 44% for Kansas. The remaining rows detail the average educational attainment of the population of the state and the fraction of individuals who are married or live in a metropolitan area. In

[1] Data is downloaded from https://usa.ipums.org/usa/.

[2] Further small cuts to the data are made given the splitting of individuals into cells. For instance, I drop individuals who are married but whose spouse I cannot identify in the data. I do this because I rely on splitting individuals by whether their spouse works or not as a means of identifying selection bias. Individuals who report living with both a room-mate and with relatives are dropped so that the cells align with those of Dahl (2002).

[3] The sample here differs slightly from that of Dahl (2002) as is clear from the empirical results and summary statistics. However, these differences are relatively minor and have no bearing on the general pattern of results of the implications drawn in this chapter.

Table 2.1: Descriptive Statistics

Variable	U.S.	California	Florida	Illinois	Kansas	New York	Texas
Migrants	35						
	(0.1)						
Inmigrant		42	71	22	36	15	38
		(0.2)	(0.3)	(0.3)	(0.6)	(0.2)	(0.3)
Outmigrant		30	37	35	44	36	24
		(0.2)	(0.4)	(0.3)	(0.6)	(0.2)	(0.3)
Less than High School	11	9	13	8	8	8	12
	(0)	(0.1)	(0.2)	(0.2)	(0.4)	(0.1)	(0.2)
High School	36	25	33	35	38	33	31
	(0.1)	(0.2)	(0.3)	(0.3)	(0.6)	(0.2)	(0.2)
Some College	29	36	31	30	31	29	30
	(0.1)	(0.2)	(0.3)	(0.3)	(0.6)	(0.2)	(0.2)
College	19	23	18	21	18	21	21
	(0.1)	(0.2)	(0.2)	(0.2)	(0.5)	(0.2)	(0.2)
Advanced Degree	6	8	5	7	4	9	6
	(0)	(0.1)	(0.1)	(0.2)	(0.3)	(0.1)	(0.1)
Married	67	58	62	66	71	61	71
	(0.1)	(0.2)	(0.3)	(0.3)	(0.6)	(0.3)	(0.2)
SMSA Residence	64	96	83	71	33	74	73
	(0.1)	(0.1)	(0.2)	(0.3)	(0.6)	(0.2)	(0.2)
Hourly Wage	12.20	14.89	11.42	13.08	10.45	14.06	11.78
	(0.01)	(0.04)	(0.05)	(0.05)	(0.07)	(0.05)	(0.04)
Observations	538127	46727	25579	26533	6234	36765	34890

Notes: Descriptive Statistics for the entire sample, and for selected states. Standard errors in parentheses.

terms of hourly wages, California pays the highest wage on average at an hourly rate of $14.89, while the average hourly wage in Kansas is just $10.45.

2.3 Estimating Migration Probabilities

One of the major contributions of Chapter 1 of this dissertation is in outlining the benefits to using the RF algorithm to estimate migration probabilities. To make clear the importance of this method in an empirical setting, I estimate migration probabilities using both the RF algorithm and the standard non-parametric splitting of individuals into cells and calculating the average transition path of the group. I derive empirical estimates using both cell and RF probabilities using the approach of Dahl, and results obtained selecting terms using the PDL method outlined previously. In this section, I discuss how I estimate the migration probabilities in each instance, and I compare the estimates obtained from each method. To make my results comparable to those of Dahl I maintain the restriction that birth state does not directly enter the wage equation.[4] The same exclusion restriction therefore applies as in Dahl's paper, that birth state and the family characteristics outlined below affect mobility decisions but not wages directly.

Non-parametric Cells

The first method used to estimate migration probabilities splits individuals into cells and then calculates the likelihood of following a particular migration path for the group as a whole. This approach relies on decisions by the researcher on where to cut the data so as to fully saturate the sample with dummy variables. To aid comparison, I group individuals in the same manner as Dahl (2002). Specifically, the cell groupings are defined differently for those who moved from their birth state ('movers') and those who remained ('stayers'), given that proportion of movers is smaller than that of stayers. Regardless, each cell is defined by birth state and individuals within each birth state are split into five education classes.[5] To estimate probabilities for stayers, individuals are further categorized by marriage, with married individuals being grouped into cells according to whether they have a working spouse, children younger than 5, and children between 5 and 18 years old. Unmarried individuals are categorized by whether they live alone, with a room-mate, or with a family member. To calculate cells for movers, individuals are also subdivided into cells by marriage, with unmarried individuals being classified according to whether they live with family or not. Married individuals are divided according to whether any children younger than 18 are present. Cells with fewer than 10 observations are dropped from

[4] As mentioned in Chapter 1 this is not a necessary restriction.

[5] A less than high school education, a high school education, some college education, a college degree, or an advanced degree.

Table 2.2: Summary of Cell Probabilities: Cell Variation

Education	# Cells	Mean	Std. Dev.	10th P'tile	90th P'tile
			STAYERS		
Less than High School	615	0.669	0.150	0.500	0.826
High School	706	0.648	0.167	0.455	0.803
Some College	695	0.558	0.163	0.352	0.729
College	633	0.477	0.168	0.250	0.688
Advanced Degree	449	0.414	0.176	0.200	0.641
			MOVERS		
Less than High School	3607	0.017	0.026	0.002	0.042
High School	6146	0.011	0.020	0.001	0.026
Some College	6301	0.013	0.022	0.001	0.032
College	5796	0.017	0.026	0.002	0.040
Advanced Degree	3919	0.025	0.034	0.004	0.058

Notes: This table examines variation across cell probabilities. Probabilities of staying/moving are calculated for each individual by grouping into discrete cells. The statistics presented here are calculated by comparing probability estimates across cells. Cells with fewer than 10 observations are excluded.

Table 2.3: Summary of Cell Probabilities: Individual Variation

Education	Mean	Std. Dev.	10th P'tile	90th P'tile
		STAYERS		
Less than High School	0.741	0.089	0.629	0.839
High School	0.747	0.080	0.642	0.841
Some College	0.660	0.094	0.550	0.791
College	0.579	0.135	0.424	0.782
Advanced Degree	0.495	0.151	0.328	0.733
		MOVERS		
Less than High School	0.033	0.041	0.004	0.075
High School	0.027	0.044	0.003	0.057
Some College	0.033	0.042	0.004	0.070
College	0.042	0.051	0.005	0.094
Advanced Degree	0.047	0.048	0.008	0.104

Notes: This table examines variation across cell probabilities. Probabilities of staying/moving are calculated for each individual by grouping into discrete cells. The statistics presented here are calculated by comparing probability estimates across individuals rather than across cells.

the analysis.

Migration probabilities capture the average preference for similar individuals over a particular migration path. Deviations from the average migration path then capture shocks in relative preferences for a particular migration and can be used to separate out earnings shocks and selectivity. Furthermore, to estimate the returns to schooling, the coefficient of interest β_{1c} is identified through differences in the returns to schooling given approximately equivalent patterns of migration. Individuals with different levels of education, but with the same estimates for migration probabilities have the same mean preferences and hence the same selectivity. Any remaining differences between their earnings is then attributable to the difference in education.

The variability of these probability estimates is summarised in Tables 2.2-2.3. Table 2.2 summarises the key moments across cells, while table Table 2.3 summarises the same moments across individuals in the sample. The first thing to note is the wide range in probability estimates within education classes. This variability is necessary for identification, ensuring that there is wide variation in preferences by birth state, or family characteristics. Both tables also make clear the increased propensity of individuals with a higher education to migrate from their birth state. The average likelihood of moving is over 50% for the college educated, and around 35% for those with a high school diploma. There is also a clear decline in the number of populated cells for the most and least educated groups as they compose a relatively small share of the sample.

RF Estimates

As discussed in section 1.5, there are clear benefits to using ML tools to estimate migration probabilities. Whether this involves the splitting of individuals into cells as in Ransom (2016), or in estimating continuous probability estimates using either the Random Forest algorithm or estimation via Artificial Neural Networks. In this section, I present summary statistics for migration probabilities estimated using the RF algorithm and compare these to the estimates from the previous section.[6]

In estimating the RF I proceed by using the same variables to predict migration as before. However, these variables can be used without restricting variation by transforming them to dummy variables. In practice this involves using birth state, five education categories, dummy variables for being married, being married with a working spouse, being divorced, living with roommates or with family, the number of children in the household, and the number of children less than 5 years old. In this manner, I allow for there to be greater variation in how the number

[6]Estimation using ANN was also considered, however, I found RF estimates to be much more stable to perturbations in the training sample, and in random initialization. ANN estimates tended to be highly correlated with both RF estimates, and cell probabilities. More generally, when plotted against cell estimates, they tended to have a similar relationship to that identified using RF estimates. They were simply too imprecisely estimated to be used for reliable inference.

Table 2.4: Summary of Random Forest Estimated Probabilities

Education	Mean	Std. Dev.	10th P'tile	90th P'tile
		STAYERS		
Less than High School	0.717	0.075	0.638	0.800
High School	0.700	0.074	0.622	0.785
Some College	0.669	0.075	0.585	0.763
College	0.608	0.092	0.504	0.742
Advanced Degree	0.558	0.090	0.470	0.688
		MOVERS		
Less than High School	0.031	0.036	0.005	0.066
High School	0.029	0.039	0.004	0.060
Some College	0.031	0.039	0.004	0.066
College	0.037	0.044	0.005	0.076
Advanced Degree	0.037	0.038	0.007	0.076

Notes: This table examines variation across Random Forest estimated probabilities. The statistics presented here are calculated by comparing probability estimates across individuals.

of children affects moving. Furthermore, I do not need to estimate separate functions for stayers and movers. To estimate the Random Forest I grow 250 trees and cross validate to select the number of random features over which to select at each node.[7] The Random Forest algorithm is implemented using 'treebagger' in matlab.[8]

Table 2.4 presents summary statistics for RF estimated migration probabilities. This summary is made across individuals and is comparable to Table 2.3 for probabilities estimated by grouping individuals into cells. The first important takeaway from this table is that the variability of estimates both within and across education classes is notably lower. When comparing RF estimated probabilities to those in Table 2.3 it is clear that there is less variation across education classes. The mean probability of staying ranges from 75% to 50% for cell probability estimates, but varies between 72% and 56% using RF estimates. The range in moving probabilities is also narrower. Within education categories I see a decline in variation using RF estimates when looking at the 90-10 percentile difference. Despite this, there is clearly sufficient within education class variability in estimates to allow for separate identification of the coefficients on the selection terms from the returns to education.

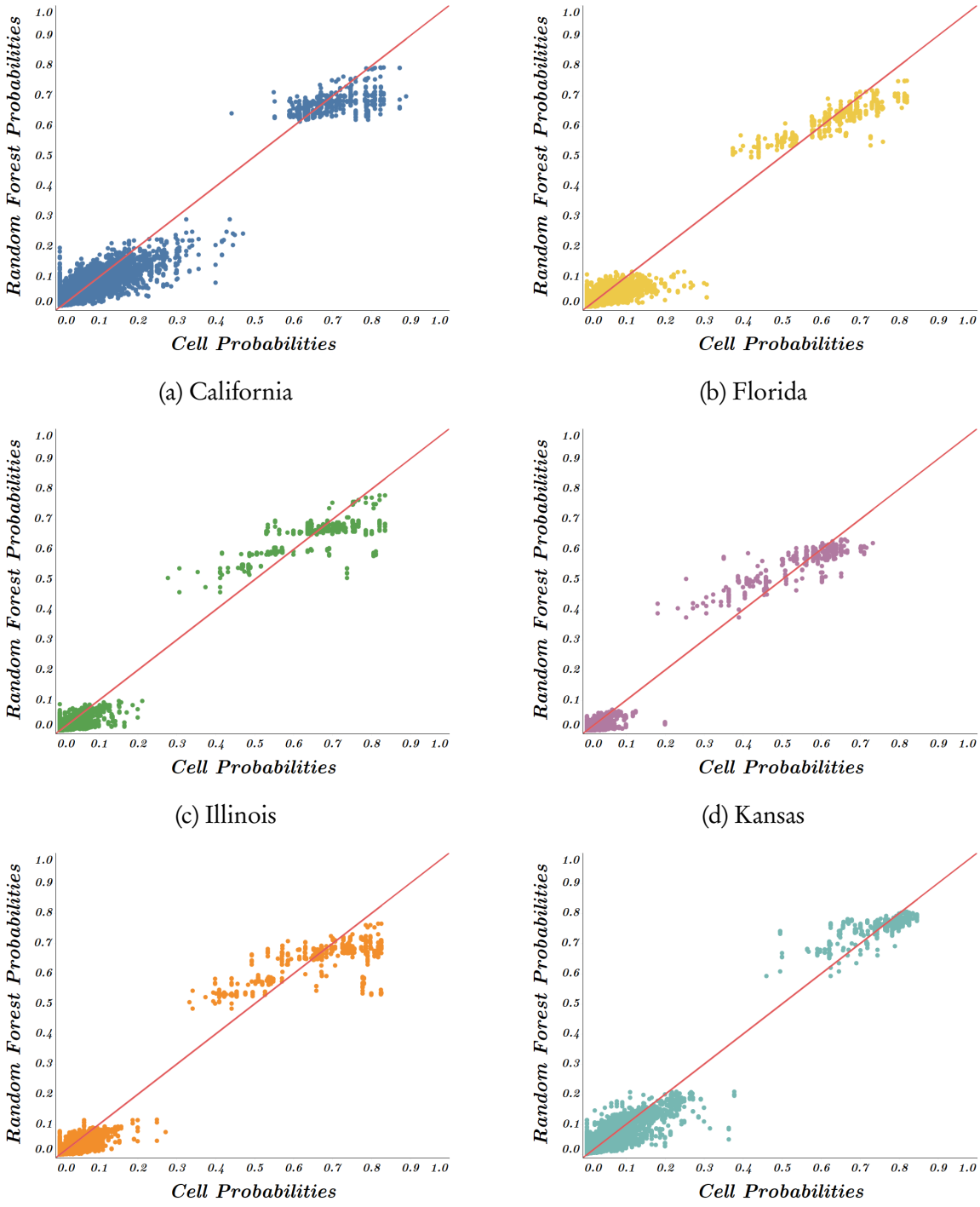
In the six key states for which I present detailed empirical results, the Random Forest estimated probabilities are plotted against cell probability estimates in Figure 2.1. This figure summarises what is clear from summary statistics: RF estimates tend to estimate migration probabilities within a narrower range. Given the benefits to RF estimation, and the degree to which ML tools tend to prevent measurement error due to poorly made cuts on the data, this suggests that the cell approach used by Dahl tends to accentuate variation in the data. This additional variation likely reflects measurement error as opposed to variation in the likelihood of moving. Splitting the sample in an ad-hoc manner then tends to increase variation across cells due to sampling variation. RF estimates (and ANN estimates not presented here) find that this variation is not caused by variation across the covariates used to predict migration patterns. For comparison the second panels of Figures 1.2-1.3 from the numerical experiment in Section 1.5 illustrate how poorly defined cells can drive excess variation around the true probability estimates.

It is important to highlight here that though RF estimates produce a narrower range of probability estimates, there is potentially more variation across individuals within this narrow band given that I do not need to split individuals into cells, and because the number of children can affect probability estimates in a continuous manner. The RF algorithm also allows for

[7] This value is 13, and with 66 inputted features is larger than the default value $\sqrt{66}$.

[8] As shown in the numerical example above, limiting the minimum leaf size will prevent noisy estimation for small cells. I present the results without limited the minimum leaf size as is standard in the literature. In results not presented, I consider estimates limiting the minimum leaf size to 50. The final empirical results are remarkably similar in both cases.

Figure 2.1: Probability Estimates: Cells versus Random Forest



(a) California

(b) Florida

(c) Illinois

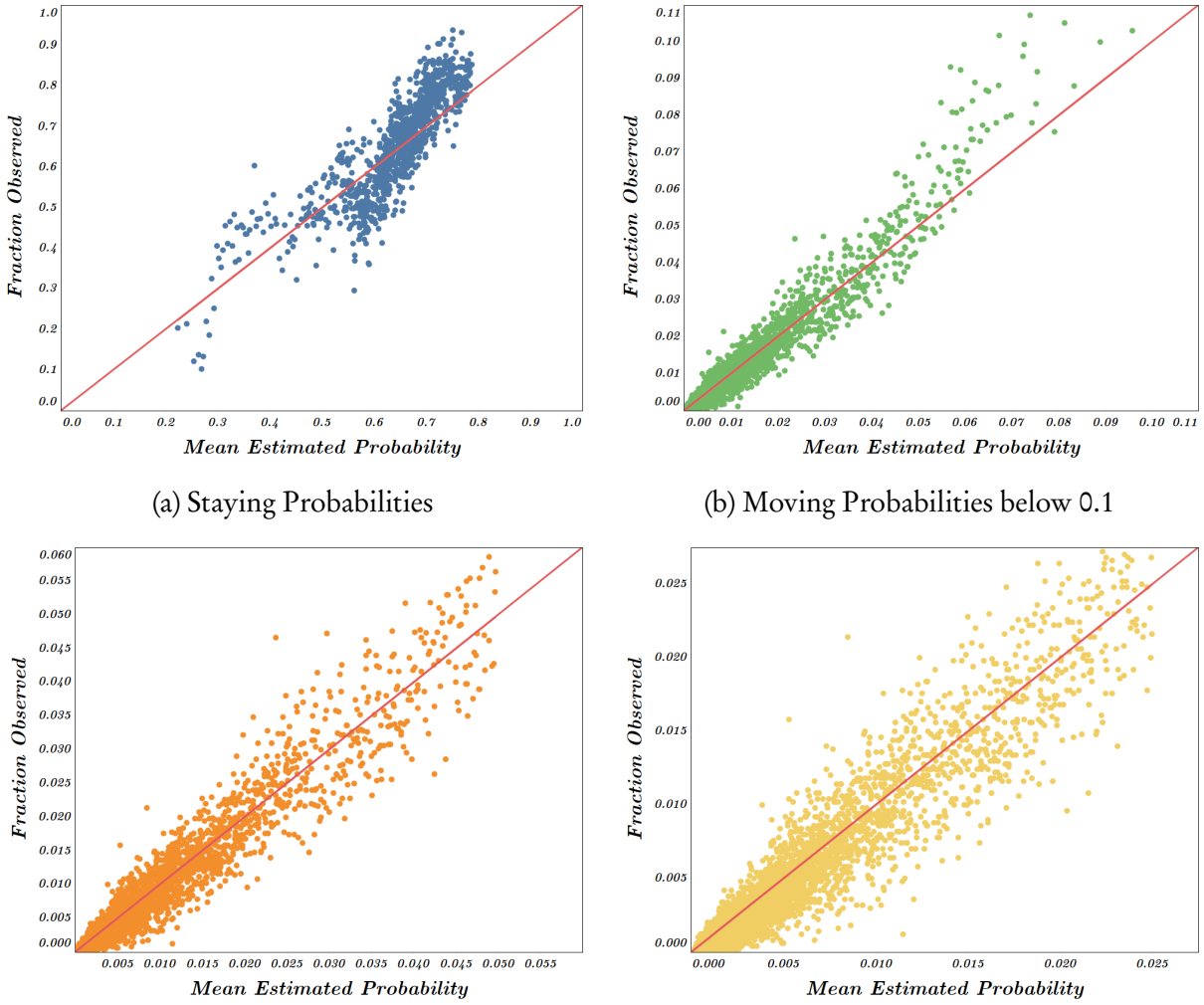
(d) Kansas

(e) New York

(f) Texas

Notes: Probabilities estimated using Dahl's method on the horizontal axis are plotted against Random Forest estimated probabilities on the y-axis. Results for selected states are presented and the 45 degree line is drawn for comparison.

Figure 2.2: Random Forest Reliability Curve



(a) Staying Probabilities

(b) Moving Probabilities below 0.1

(c) Moving Probabilities below 0.05

(d) Moving Probabilities below 0.025

Notes: Individuals are ranked according to their predicted moving probability and grouped into bins. Within each bin the fraction of individuals who undertake the specific transition path are calculated. This fraction is displayed on the y-axis, and the average of the estimated probability for the bin is plotted on the x-axis. Problems with the RF algorithm would be indicated by deviations from the 45 degree line which is drawn in red.

more flexible interactions between covariates and does not limit the variation used in estimating moving probabilities. Furthermore, the summary statistics make clear that there is sufficient variability in estimates within education categories as to permit identification of the coefficients on the probability terms.

The Reliability Curve

One potential downside to estimation of probabilities using a RF is that RF estimation can bias estimates away from 0 and 1. Given the relationship between RF estimates and cell estimates in Figure 2.1, it is important to ask if this issue is driving estimates here. Fortunately, it is possible to examine whether this bias exists in the empirical estimates by appealing to the reliability curve. In Figure 2.2 I plot segments of the reliability curve to answer this question. The reliability curve is based on basic intuition, that if I group individuals who have a 20% chance of being observed in a state, then 20% of them should be observed in that state. To assess the reliability of my estimates, I plot the fraction observed in a state against the mean estimated probability of being observed in that state. I group individuals in large bins of 5000 individuals by ranking their estimated probability. If the curve deviates from the 45 degree line in a systematic way, then there is evidence of systematic bias in RF estimates. It is important to note that variation around the 45 degree line will always exist as probabilities are being predicted using a limited subset of observable characteristics. However, this bias is specific to the RF algorithm and operates only at the very tails at the distribution.

In panel (A), I plot the upper end of the curve which captures the probability of staying in the state of birth. Here, it would appear that perhaps there is some evidence of downward bias for RF probability estimates in the range of .7 to .8 . Below .7, estimates hover around the 45 degree line with no clear trend in whether they fall below or above the line. This implies that the bias operates exclusively at the very upper end of estimates, suggesting the likely effect on estimates will be relatively minor. Additionally, even over this range there is still a strong positive correlation between the expected probability observed in their birth state, and the average probability estimate for those observed.

However, this pattern has the potential to explain why RF estimated probabilities exceed cell estimated probabilities in the upper tail estimates in Figure 2.1. For this pattern to fully explain that shown in Figure 2.1, all estimates below .6 would have to fall below the 45 degree line. Instead, there is a dip below the line in the range .6 to .7 and then estimates tend to however around the 45 degree line. It is unlikely then that this pattern fully explains that in Figure 2.1 for staying probabilities. Additionally, Florida, Illinois, and Kansas observe RF probability estimates for stayers that are on average lower than cell estimated probabilities below the range .7. For Texas we see a substantial proportion of RF estimates above .7 which exceed cell estimates.

At the lower end I see that in plotting the reliability curve for estimated probabilities below 0.1 and .05 in panels (B) and (C), there is little evidence of a bias away from 0. In panel (D) I plot the curve for values below 0.025. There is perhaps some evidence here of a bunching below the curve for values less than 0.0075. This would suggest that at most the bias in RF estimates is shifting estimates from 0 to 0.0075 and hence is not going to affect estimates considerably. In summary then there is no evidence that the RF algorithm biases results away from 0, but some evidence of a downward bias in estimates over the range .7 to .8. This may explain in part the compression of probability estimates in RF estimates in Figure 2.1, but it likely does not explain fully the pattern observed for staying probabilities, and has no bearing on the pattern observed for moving probabilities. Furthermore, it is arguable that variation away from the 45 degree line over the range .7 to .8 is simply due to variability in estimates coming from the limited subset of variables used to predict probabilities rather than inherent RF bias.[9]

2.3.1 Transition Matrices

In Figure 2.3, I present the transition matrix between the six states considered in detail in my empirical results. The transition matrix gives a clear visual representation of flows between states. Each figure in the matrix corresponds to the proportion of individuals from the birth state living in this state.[10]

Diagonal elements of the matrix display the staying probability for each birth state. For most states there is a clear decline in the probability of staying for college educated individuals. However, this decline is markedly lower for Texas and is not observed for California. This result is consistent with the Roy model of migration presented in section 1.2 where individuals move based on tastes and earnings potential in a manner that varies across education groups. For instance there are very likely non-pecuniary factors driving the decision to stay for highly educated individuals in California relative to other states, especially since, as noted by Dahl (2002), the returns to education relative to California are higher in other states such as Texas, and equivalent to those observed in Florida and New York for instance. If non-pecuniary factors are unimportant then similar migration flows should be observed in California as in these states.

In examining the other panels for California, there is some evidence of varying preferences across education groupings. College educated individuals are more likely to move from California to New York and to a lesser degree, Illinois. Evidence of variability across education categories is much stronger for other birth states. For Floridians, the college educated are more than twice as likely to move to California compared to the least educated workers. College

[9]The reliability curve for Cell estimated probabilities, also indicates some bunching above the curve for probability estimates around .8, and a bunching below for estimates in the range .6 to .7.

[10]I average across Random Forest estimated probabilities.

educated workers born in Florida, Illinois, and Kansas are also much more likely to move to Texas. These patterns provide evidence then of variations in preferences for migration across education categories within birth states, which strongly suggests workers are non-randomly allocated across states in a manner that varies by educational attainment.

The transition matrix also provides further evidence regarding migration flows more generally. For instance, California tends to attract a relatively high proportion of migrants relative to other states. The same is true for Texas and Florida. In contrast, relatively few individuals move to Kansas or Illinois. The relationship between New York and Florida is also particularly interesting. Specifically, less educated workers are more likely to move from New York to Florida, but those with higher levels of education are more likely to follow all other migration paths. The size of these migrations is also rather large. The reverse relationship does not hold, as flows from Florida to New York are not large and are more likely to be undertaken by the college educated. This reflects a more general relationship where flows in both directions are not necessarily similar. Additionally, it is interesting to note that there is only limited evidence of flows between states that are relatively close to one another. Though flows from Kansas to Texas are large, flows between Kansas and Illinois are much smaller than flows from Kansas to California.

Figures B.1-B.2 present migration paths for aggregated regions of the country, and for states to regions respectively. Regions are defined by aggregating states as in Table B.1. In both figures the population share of each region is presented for context and the probabilities presented here represent raw estimates across the sample rather than averages across RF estimated probabilities.[11] Stayers are dropped here and so bars represent the share of movers migrating to this region conditional on moving from the their state of birth.

Figure B.1 confirms that migration paths vary with education, and are distinct between pairs of birth and residence states. For instance, individuals born in the Northeast, the South, and the West, have a clear home bias as the share of moves to the home region is much greater than the population share of these regions. For the Midwest, the share of moves to other states in the same region matches closely the population share, suggesting flows are not geographically concentrated for individuals from this region. In general, flows to the Midwest are lower than the population share of this region for all sending states. The same is true for flows to the Northeast while flows to the South and the West tend to be larger. In general then, there is evidence of home bias, and a relative preference for the South and the West over the Midwest and the Northeast. Across educational groupings, the South tends to be preferred by less educated workers, while more educated workers have relatively high migration paths to the Northeast.

In looking at the flows from the six key states to each region I again observe relatively small

[11] Though using RF probabilities yields very similar patterns

Figure 2.3: Transition Matrix using RF Probabilities

Notes: Transition paths are calculated for each education group by averaging across individuals' predicted probabilities.

flows to the Northeast from states in other regions of the country. Individuals that do move to the Northeast are more educated on average from these sending states. There is strong evidence of home bias for Californians, with around 40% of individuals moving to states in the west, despite just 20% of the population living in these states. A preference for moving to the West is also observed for Texas, Kansas, and Illinois. For New York, flows to the South are relatively large, perhaps due to the massive flows between New York and Florida. As in the previous figure, flows to the South are larger on average for less educated workers compared to those with a college education.

In summary these transition matrices, in addition to providing evidence on migration flows between states and regions, provide clear evidence of variation in preferences across sending states and education groupings. This variation in mean earnings and preferences drives non-random sorting across locations in a manner consistent with the Roy model in section 1.2.

2.4 Results

In this section, I turn to estimation and present estimates for the returns to education using both the method outlined in Dahl (2002) and the selection procedure using post-double-Lasso as outlined in section 1.3. I also compare results obtained using the two different methods by which I estimate transition probabilities. The estimating equation for Dahl's method takes the form:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \sum_s M_{isc} \times \mu^*_{sc}\left(p_{ijc}, p_{ijj}\right) + v^*_{ic} \tag{2.1}$$

which imposes the weaker version of the SISA as outlined in A-3 and where $\mu^*_{sc}(.)$ represents a cubic polynomial in the first best and retention migration probabilities. s here captures whether an individual is a stayer or a mover as I follow Dahl in estimating a separate control function for each type. I also estimate a restricted form of the model which uses only a single control function for each type:

$$y_{ic} = \alpha_c + \beta_{1c} s_i + \beta_{2c} x_i + \mu^*_c\left(p_{ijc}, p_{ijj}\right) + v^*_{ic} \tag{2.2}$$

Given that the underlying theory implies a separate control function for each birth state, this model represents only a partial relaxation of the restriction. The assumption embedded in (2.1) is that the migration probabilities affect selectivity in a different manner for those who remain in the state than for movers to the state. Also, amongst movers, the relationship between migration probabilities and selection is identical regardless of birth state.

In using PDL to select terms, I formulate a general version of the model which includes a full

Table 2.5: Uncorrected Estimates for Selected States

	CA	FL	IL	KS	NY	TX
LHS	-0.154***	-0.160***	-0.186***	-0.194***	-0.194***	-0.186***
	(0.010)	(0.010)	(0.012)	(0.023)	(0.011)	(0.010)
Some College	0.118***	0.125***	0.098***	0.039***	0.146***	0.146***
	(0.006)	(0.008)	(0.008)	(0.014)	(0.007)	(0.007)
College	0.429***	0.451***	0.369***	0.347***	0.440***	0.517***
	(0.008)	(0.010)	(0.010)	(0.019)	(0.008)	(0.009)
Advanced	0.587***	0.662***	0.545***	0.497***	0.604***	0.684***
	(0.011)	(0.015)	(0.014)	(0.032)	(0.011)	(0.013)
Exper	0.081***	0.049***	0.075***	0.046	0.084***	0.048***
	(0.011)	(0.015)	(0.014)	(0.030)	(0.012)	(0.013)
Exper^2	-0.003***	-0.001	-0.003**	-0.001	-0.004***	0.001
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$\text{Exper}^3 \times 100$	0.005	-0.001	0.005	-0.001	0.008*	-0.008*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.004)
Married	0.153***	0.156***	0.184***	0.160***	0.167***	0.169***
	(0.005)	(0.006)	(0.006)	(0.013)	(0.006)	(0.006)
SMSA	0.170***	0.112***	0.254***	0.236***	0.235***	0.132***
	(0.012)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
R-Squared	0.138	0.167	0.190	0.180	0.192	0.192
Obs	46727	25579	26533	6234	36765	34890

Notes: Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

Table 2.6: Corrected Estimates for Selected States using Dahl's Approach: Separate Control Function for Stayers and Movers

	CA	FL	IL	KS	NY	TX
LHS	-0.151***	-0.157***	-0.215***	-0.199***	-0.202***	-0.184***
	(0.011)	(0.011)	(0.015)	(0.022)	(0.012)	(0.010)
Some College	0.125***	0.091***	0.060***	0.029*	0.113***	0.120***
	(0.007)	(0.009)	(0.009)	(0.016)	(0.008)	(0.008)
College	0.412***	0.400***	0.290***	0.319***	0.382***	0.469***
	(0.009)	(0.013)	(0.014)	(0.025)	(0.011)	(0.011)
Advanced	0.552***	0.608***	0.449***	0.453***	0.534***	0.614***
	(0.013)	(0.020)	(0.022)	(0.043)	(0.015)	(0.018)
Exper	0.082***	0.047***	0.069***	0.047*	0.081***	0.047***
	(0.012)	(0.016)	(0.015)	(0.029)	(0.012)	(0.014)
Exper^2	-0.003***	-0.001	-0.003	-0.001	-0.004***	0.001
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$\text{Exper}^3 \times 100$	0.006	-0.002	0.004	-0.001	0.008*	-0.009*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.005)	(0.005)
Married	0.109***	0.147***	0.154***	0.158***	0.144***	0.174***
	(0.007)	(0.007)	(0.007)	(0.014)	(0.007)	(0.006)
SMSA	0.179***	0.112***	0.255***	0.236***	0.235***	0.124***
	(0.012)	(0.008)	(0.007)	(0.012)	(0.006)	(0.007)
R-Squared	0.145	0.171	0.196	0.181	0.197	0.195
Obs	46727	25579	26533	6234	36765	34890

Notes: Bootstrapped standard errors (500 replications) in parentheses to control for variability in two-step estimation. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

set of second order polynomials and interactions between all 50 migration probabilities.[12] Using the PDL method outlined above, I perform Lasso twice, in the first instance regressing earnings on the control terms, and in the second instance, regressing education on the control terms.[13] I include additional controls for potential experience, being married and living in a metropolitan area. These controls are partialled out using an application of Frisch-Waugh-Lovell. I use square-root Lasso using the optimal penalty level in Belloni et al. (2011). I maintain the assumption of a single control function for all birth states in the main results.

Dahl's results

I begin by presenting the results using Dahl's method which imposes a restriction on the covariance of the error draws. Table 2.5 presents OLS results for comparison using dummies for a less than high school education (LHS), some college education, a college education, and for having an advanced degree (Advanced). The returns to education are expressed relative to having a high school education. I include a cubic in potential experience,[14] a dummy for being married and a dummy for living in a metropolitan area. The main results of this empirical section are presented for the same six states on which Dahl focuses. The uncorrected estimates suggest distinct differences in the returns to education across states with a much greater penalty for having a less than high school education in Kansas and New York relative to California and Florida. The returns to a college degree are much lower in Illinois and Kansas than in the other states.

Results obtained using Dahl's estimation strategy are presented in Tables 2.6-B.4. As estimation here requires the use of a two-step procedure where probabilities are estimated in the first step, estimates of standard errors must be adjusted to account for this additional sampling variability. Dahl (2002) proposes a closed form correction based on Murphy and Topel (1985). I correct for additional variability in estimates here by bootstrapping which yields estimates comparable to the closed form method.

Focusing on Tables 2.6-2.7 I see a clear decline in the coefficient estimates on having a college degree or an advanced degree relative to OLS estimates. Using a separate control function for movers and stayers tends to have a greater impact on estimates which may suggest separate control functions for stayers and movers do a better job of capturing the patterns of selection bias. In general, the results also indicate an upward bias in OLS estimates of the returns to some college education, except in California where this estimate is under-estimated. In Table 2.8, I present Hausman tests comparing the OLS and corrected estimates. This table shows that

[12] Dropping those which are collinear.

[13] As educational attainment is ordered, I perform Lasso with the dependent educational variable as a categorical predictor. Monte Carlo experiments confirm the efficacy of this approach. An alternative would be to select terms using a multinomial logit with a Lasso penalty.

[14] Age - Education - 6

Table 2.7: Corrected Estimates for Selected States using Dahl's Approach: Single Control Function

	CA	FL	IL	KS	NY	TX
LHS	-0.151***	-0.155***	-0.212***	-0.198***	-0.200***	-0.184***
	(0.010)	(0.011)	(0.015)	(0.022)	(0.012)	(0.009)
Some College	0.126***	0.115***	0.072***	0.033**	0.121***	0.131***
	(0.006)	(0.008)	(0.008)	(0.015)	(0.007)	(0.007)
College	0.420***	0.440***	0.314***	0.335***	0.397***	0.490***
	(0.008)	(0.011)	(0.013)	(0.022)	(0.011)	(0.010)
Advanced	0.562***	0.652***	0.480***	0.487***	0.558***	0.645***
	(0.013)	(0.018)	(0.019)	(0.036)	(0.014)	(0.016)
Exper	0.083***	0.049***	0.073***	0.047*	0.083***	0.048***
	(0.012)	(0.016)	(0.015)	(0.028)	(0.012)	(0.014)
Exper^2	-0.003***	-0.001	-0.003*	-0.001	-0.004***	0.001
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$\text{Exper}^3 \times 100$	0.006	-0.001	0.005	-0.001	0.008*	-0.008*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.005)	(0.005)
Married	0.124***	0.152***	0.159***	0.156***	0.148***	0.172***
	(0.006)	(0.007)	(0.007)	(0.014)	(0.006)	(0.006)
SMSA	0.179***	0.113***	0.255***	0.238***	0.235***	0.125***
	(0.012)	(0.008)	(0.007)	(0.012)	(0.006)	(0.007)
R-Squared	0.138	0.168	0.191	0.180	0.192	0.193
Obs	46727	25579	26533	6234	36765	34890

Notes: Corrected estimates of the returns to schooling. Bootstrapped standard errors (500 replications) in parentheses to control for variability in two-step estimation. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

Table 2.8: Corrected Estimates versus OLS

	CA	FL	IL	KS	NY	TX
	Double-Post Lasso					
LHS	3.06***	3.32***	3.29**	-0.19	0.01	2.57
	(51.88)	(28.73)	(4.78)	(0.01)	(0.00)	(7.17)
Some College	-1.02**	-2.02***	1.16	-0.95	-1.73	-3.32***
	(6.12)	(21.93)	(1.25)	(0.51)	(2.25)	(37.88)
College	-4.77***	-9.49***	-4.42*	-0.53	-6.02**	-4.73**
	(26.76)	(48.53)	(3.56)	(0.03)	(6.36)	(10.25)
Advanced	-9.73***	-14.73***	-9.46**	-4.66	-12.04***	-6.59***
	(32.29)	(42.74)	(5.77)	(0.67)	(11.41)	(5.55)
	Dahl 2 Control Function - Cell Probabilities					
LHS	0.24	0.39	-2.91***	-0.46	-0.82	0.26
	(0.28)	(0.85)	(13.19)	(0.30)	(2.16)	(5.07)
Some College	0.75**	-3.43***	-3.77***	-1.00	-3.36***	-2.62***
	(5.18)	(56.33)	(63.17)	(2.14)	(105.63)	(50.42)
College	-1.73***	-5.11***	-7.91***	-2.71*	-5.78***	-4.74***
	(15.16)	(37.27)	(68.92)	(2.77)	(64.7)	(55.74)
Advanced	-3.43***	-5.41***	-9.58***	-4.39	-6.95***	-6.96***
	(16.85)	(18.81)	(32.99)	(2.25)	(50.97)	(31.46)
	Dahl 1 Control Function - Cell Probabilities					
LHS	0.23	0.58*	-2.58***	-0.37	-0.60	0.27
	(0.37)	(3.57)	(10.81)	(0.18)	(1.45)	(1.93)
Some College	0.81***	-0.98***	-2.58***	-0.64*	-2.50***	-1.50***
	(21.57)	(26.54)	(46.8)	(2.78)	(95.93)	(47.2)
College	-0.87***	-1.09***	-5.45***	-1.13	-4.29***	-2.63***
	(9.15)	(9.33)	(44.82)	(1.08)	(46.63)	(35.7)
Advanced	-2.40***	-0.98	-6.50***	-1.01	-4.57***	-3.95***
	(11.58)	(1.04)	(27.04)	(0.36)	(35.28)	(16.56)

Notes: Corrected estimates compared to OLS estimates. Presented values are corrected estimates, less OLS estimate, multiplied by 100. Hausman test F-statistic in brackets. Hausman test significance: * 10%, ** 5%, *** 1%. Estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34.

for all states except Kansas the difference in coefficients between the uncorrected and corrected estimates of the returns to some college education, a college degree, or an advanced degree are significantly different. For Illinois, there is also a significant upward bias in OLS estimates of the LHS premium. In general then, results obtained using Dahl's method and cell probability estimates indicate an upward bias in the returns to higher levels of education in all states. The largest observed decline in the college premium is 21% for Illinois, while the lowest observed decline is 4% for California.

In Tables 2.6-B.4, the same results are obtained using RF estimated migration probabilities. Table B.6 presents the Hausman test results comparing these coefficients to OLS estimates. In general, the results obtained using RF estimated probabilities align closely with those in Tables 2.6-2.7. In particular the direction of bias correction is the same for all results which are significantly different from OLS estimates. The magnitude of results is notably different however. Focusing on the results using 2 control functions, the estimates using cell probabilities tend to find a greater upward bias in OLS estimates for all states in the return to a college or advanced degree. Largely, the estimates using cell probabilities predict an upward bias in the college premium 3-5% larger than estimates using RF probabilities would suggest. Regarding the returns to some college education, the results in table D2 find a greater upwards bias in OLS estimates for Florida and Illinois, New York and Texas, but a smaller downward bias for California. The results on LHS also find a reduced downward bias for Florida, and a large upward bias for Illinois.

Though these differences are small, they indicate a tendency for the estimates attained using cell probabilities to overstate the degree of upward bias in OLS estimates and underestimate the degree of downward bias.

Post-Double-Lasso Estimates

In Tables B.5 - 2.9, I present the main empirical results where control terms are selected via post-double-Lasso.[15] In Table B.5, PDL is used to select from among the control variables estimated

[15]In contrast to the results of Dahl's method here, I present unadjusted standard errors. The core theoretical result of Belloni et al. (2014) establishes that under sparsity conditions plug-in standard errors are consistent, asymptotically valid, and hence can be used in standard statistical tests and in forming confidence intervals. To make this idea more concrete, the assumption in their paper is that in approximating the control function using a subset of control terms, the approximations need not be exact, but approximation error must be low. Provided this approximation error is low, then standard errors are consistent. In the empirical application this approximation error will include deviations from the true value of the control function due to deviations of migration probabilities from their true value. The same assumptions necessary to establish the unbiasedness of estimates, namely a sparse approximation of the control function, is also necessary to establish the validity of standard error estimates. The identification assumption here then is that the model permits a sparse approximation and that approximation error is sufficiently low so as to not affect the consistency of coefficient estimates, or the validity of standard errors. This approximation error accounts for variability in migration probability estimates.

using cell probabilities. For each state at least 150 control terms are included, with more terms being included for larger states. The results in this table differ markedly from those in Tables 2.6-B.4, not just in terms of magnitude, but also in the direction of selection bias. For California, the PDL results find an upward bias in OLS estimates of a less than high school education, and a college education. Results in 2.6 - B.4 are indicative of the opposite result. Furthermore, PDL estimates suggest a much greater upward bias in OLS estimates of the college premium or the returns to an advanced degree. The same is true for Florida, Illinois, and New York.

My preferred estimates however are presented in Table 2.9 and use PDL selection combined with probabilities estimated using the RF algorithm. Comparing results in B.5 and 2.9, the first clear difference is that fewer terms are selected when using RF estimated probabilities. For the six states considered here around 1/3 fewer control terms are included in the model. Given the comparison of migration probability estimates in Figure 2.1, this is likely caused by the increased variability due to sampling variation in cell probabilities. This will naturally bias down coefficients leading to less shrinkage being applied in Lasso estimation. This will cause more terms to be included in the model due purely to increased noise in probability estimates and hence, control variables are over-selected in Table B.5 relative to Table 2.9.

The results of Table 2.9 are presented relative to OLS in Table 2.8 with accompanying Hausman tests for significant differences. It is important to note that coefficients in Table 2.9 are estimated with less precision than those in 2.6 - B.4. Given the results in the Monte Carlo experiment in section 1.4 this is unsurprising, as the inclusion of more probability terms will naturally lead to greater variation in estimates. From the Monte Carlo experiment, this cost in precision is accompanied by the assuredness of bias reduction and the validity of inference tests regardless of whether the SISA holds. Despite this increase in standard errors I still find evidence of an upward bias in OLS estimates for the two highest education groupings in all states excepting Kansas. Furthermore, the magnitude of this bias is much larger than that found using Dahl's method in California, Florida, and New York. The estimated college premium is 11% lower than the OLS estimate using this method, and just 4% lower using Dahl's method.

For Illinois, the college premium is notably different, but the returns to an advanced degree are similar. For Texas, PDL results overlap very closely with those in 2.6. In contrast to Dahl, I also find a positive downward bias in the returns to a less than high school education in California, Florida, and Illinois. For Illinois, this result contrasts sharply with the upward bias predicted by Dahl's method. Dahl's method also predicts the opposite pattern of selection bias for the return to some college education in California.

It is also important to highlight that due to the smaller standard errors of estimates in 2.6-B.4, Hausman tests are more likely to predict significant differences between estimates. However, given that Dahl's method performs poorly when the SISA is violated, these tests may lead to

Table 2.9: Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities

	CA	FL	IL	KS	NY	TX
LHS	-0.123***	-0.127***	-0.153***	-0.196***	-0.194***	-0.161***
	(0.011)	(0.012)	(0.019)	(0.029)	(0.021)	(0.014)
Some College	0.108***	0.105***	0.110***	0.030	0.129***	0.113***
	(0.007)	(0.009)	(0.013)	(0.020)	(0.013)	(0.009)
College	0.381***	0.356***	0.325***	0.341***	0.380***	0.469***
	(0.012)	(0.017)	(0.025)	(0.038)	(0.025)	(0.017)
Advanced	0.489***	0.515***	0.450***	0.450***	0.483***	0.618***
	(0.020)	(0.027)	(0.042)	(0.065)	(0.037)	(0.031)
Exper	0.087***	0.044***	0.069***	0.026	0.081***	0.052***
	(0.011)	(0.015)	(0.014)	(0.030)	(0.012)	(0.013)
$Exper^2$	-0.004***	-0.001	-0.003**	0.001	-0.004***	0.000
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$Exper^3 \times 100$	0.008	-0.002	0.006	-0.005	0.010**	-0.007*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.004)
Married	0.130***	0.138***	0.146***	0.154***	0.121***	0.152***
	(0.008)	(0.009)	(0.014)	(0.020)	(0.015)	(0.009)
SMSA	0.179***	0.110***	0.259***	0.237***	0.232***	0.126***
	(0.011)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
# Cells	199	186	170	87	211	212
R-Squared	0.151	0.173	0.205	0.190	0.202	0.200
Obs	46727	25579	26533	6234	36765	34890

Notes: Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

incorrect conclusions regarding the pattern of selection bias. For instance, the results in Table 2.8 indicate an upward bias in OLS estimates of the return to some college education in New York. However, if the SISA does not hold, then inference based on the standard errors estimated in Table 2.8 is not valid and conclusions cannot be drawn from the results of the Hausman test.

Given the notable difference between the PDL results and the results obtained using Dahl's method,[16] I conclude that the SISA likely does not hold in this empirical example (and for these six states) and that Dahl's estimates tend to mischaracterise the pattern of selection for the returns to a less than high school education, and underestimate the upward bias in OLS estimates for the returns to a college degree or an advanced degree.

Comparing PDL and Dahl's estimates

In Table B.7 I present differenced estimates obtained using my preferred PDL specification and alternative modifications to Dahl's method.[17] It is clear from these results that there are large differences between the results obtained using each method. Most of the difference between estimates is observed in California, Florida, and Illinois. Lasso estimates of the return to an advanced degree are notably lower in 4 out of 6 states, and the returns to a college education are lower in California and Florida. The PDL selection method also finds a smaller wage penalty for not having a high school diploma.

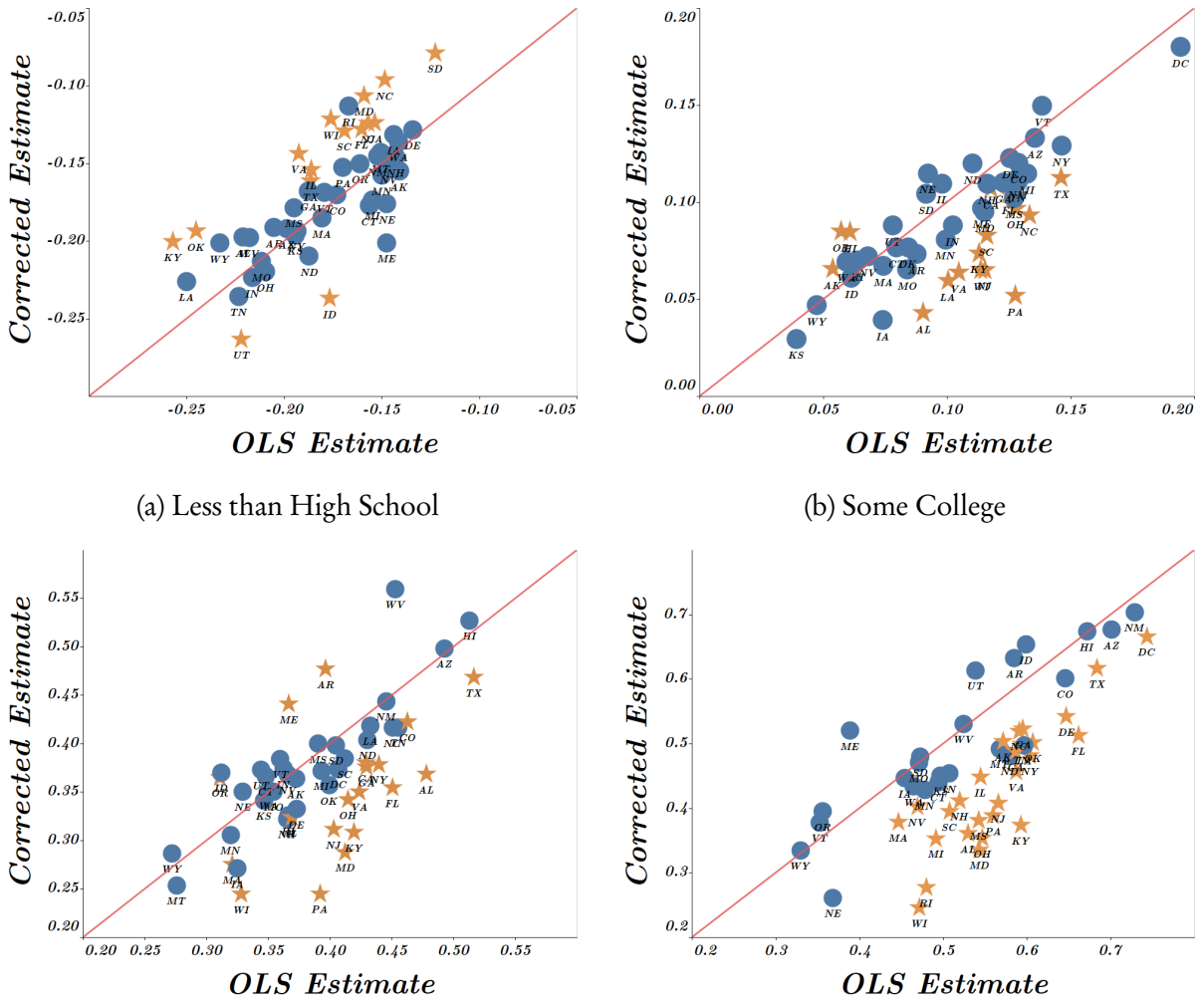
Results for all States

In Tables B.8 - B.15, I present the coefficient estimates for all states comparing PDL estimates to both OLS and corrected estimates using just the first best migration probability and the retention probability. These results are summarised visually in Figure 2.5 and Figure 2.4 which plots PDL estimates versus OLS estimates and estimates obtained using Dahl's method. The results are presented separately by education and estimates which are significantly different as indicated by the Hausman test are clearly indicated for the results relative to OLS. The Hausman test cannot be used to compare PDL estimates to Dahl's results. The first clear pattern in Figure 2.5 is that Dahl's estimates largely overestimate the negative impact of not having graduated high school. Combined with the first panel in Figure 2.4, it is clear that this results from underestimating the downward bias in OLS estimates due to selection bias.

[16] From the Monte Carlo exercise in chapter 1 the mean estimate would be the true value for both PDL and Dahl's method if SISA holds, but not if it is violated. The number of terms does not necessarily indicate that SISA does not hold. Migration patterns across states can differ by educational attainment (driving term selection) while the covariance structure of the error terms can be such that a single probability term is sufficient to characterise selectivity.

[17] Note that Hausman tests are not defined in this instance as it is not possible to discern the most efficient estimator between the two methods.

Figure 2.4: PDL Corrected Estimates Versus Uncorrected Estimates: All States



(a) Less than High School

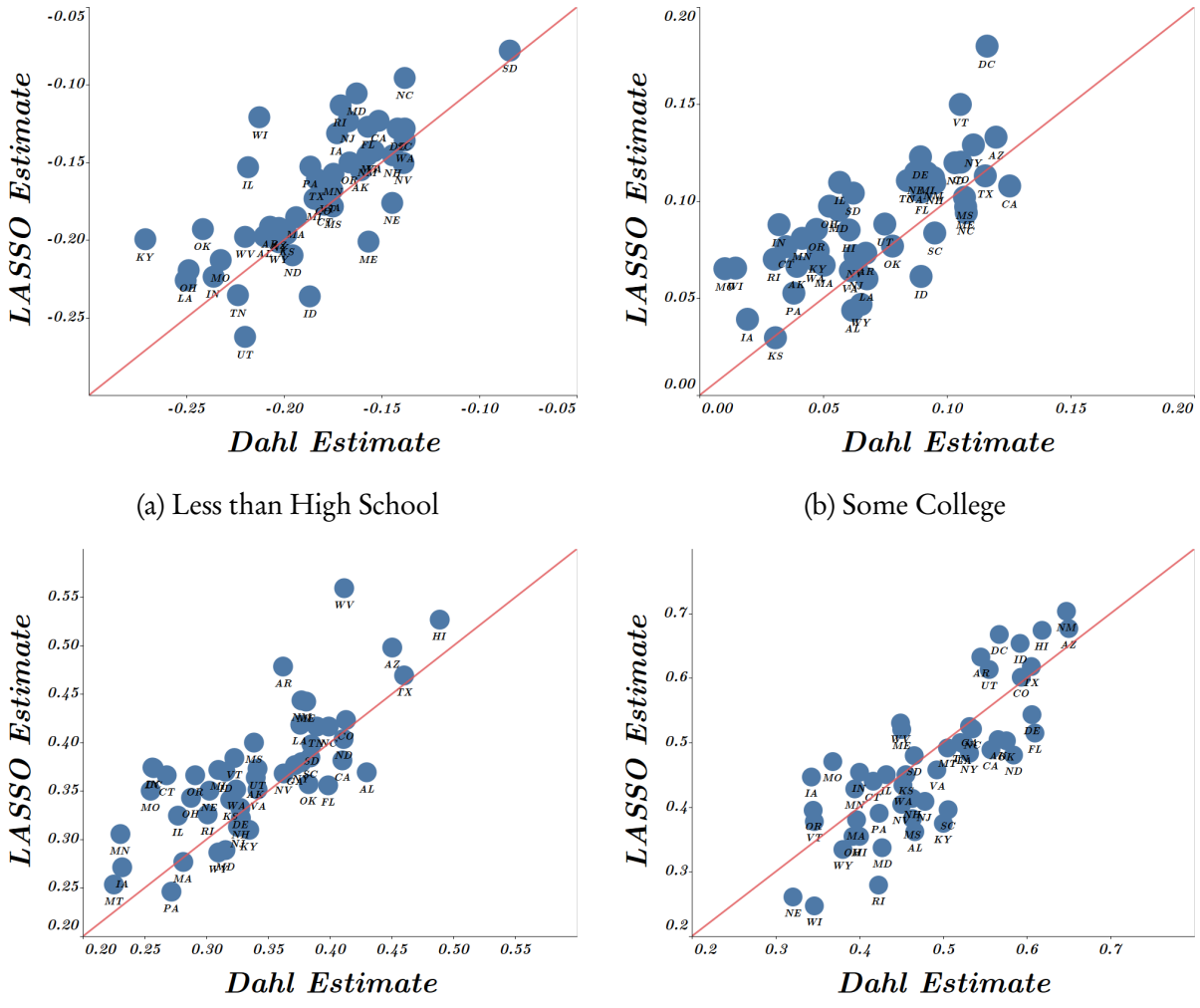
(b) Some College

(c) College

(d) Advanced Degree

Notes: PDL corrected estimate on the y-axis plotted against OLS estimate on the x-axis. Yellow stars correspond to a significant difference between estimates as calculated by the Hausman test at the 5% level. 45 Degree line drawn in red for reference.

Figure 2.5: PDL Corrected Versus Dahl Corrected Estimates: All States



(a) Less than High School

(b) Some College

(c) College

(d) Advanced Degree

Notes: PDL corrected estimate on the y-axis plotted against estimate obtained using Dahl's method on the x-axis. 45 Degree line drawn in red for reference.

In panel (b) of 2.5, I similarly find that the returns to some college education are understated using Dahl's specification. However, in Figure 2.4 panel (b), it is clear that Lasso estimates are on average lower than OLS estimates. OLS estimates are in general upward biased, while Dahl's estimates are on average downward biased (relative to Lasso corrected estimates) suggesting that Dahl's method overstates the degree of positive selection bias. The same result largely holds for the college educated, where PDL estimates are on average lower than OLS estimates, but higher than coefficient estimates obtained using Dahl's method.

Finally, relative to OLS estimates my preferred specification tends to find lower returns to an advanced degree. Dahl's method understates the upward bias in OLS estimates on average. However the reverse pattern is true for a handful of states with relatively high returns to an advanced degree. In conclusion, results obtained using a PDL selection procedure vary noticeably from those obtained using Dahl's method. Depending on the education grouping, Dahl's estimates can overestimate or underestimate the degree of selectivity. In general, the corrected estimates from my preferred specification suggest that OLS results are upward biased for all education levels (with some notable exceptions). I find that Dahl's estimates tend to over-correct for this bias when looking at some college and college estimates. However, the opposite is true for the returns to an advanced degree or a less than high school education.

2.4.1 Additional Results

Results obtained using machine learning tools to select covariates in Table B.5 and Table 2.9 use a single control function for all individuals regardless of birth state. In this section, I consider a relaxation of this assumption. Furthermore, I consider a Lasso specification which relaxes the assumption of a constant penalty loading.

Weighted Penalty Loadings and Heteroskedasticity

To relax the assumption of a constant penalty loading, I consider weighted penalty loadings as in (1.22). I use the algorithm outlined in section 1.3.2 to estimate the ideal penalty loadings. Otherwise, the PDL procedure is identical to the baseline case estimated using RF probabilities. The results of this exercise are presented in Table 2.10 and standard errors are robust to heteroskedasticity. What is immediately clear in comparing these results to the baseline case is the number of correction terms included in the model. Allowing for a more flexible penalty weighting leads to many more terms being selected through PDL. With a few notable exceptions, the results in Table 2.10 are comparable to the baseline estimates. For California and Florida, there is very little difference in coefficient estimates between the two frameworks. For Texas, the coefficients on all three education classes above high school are also very similar, al-

though the return to a less than high school education is noticeably higher and closer to the OLS estimate.

For New York, these results are indicative of greater upward bias in the college and advanced degree premium, but are in line with baseline estimates. However, there is a significant difference in the coefficient on a less than high school education, which is suggestive of much greater downward bias here than the baseline model captures. This coefficient estimate is more in line with the estimate in Table B.5 suggesting that a too restrictive penalty loading misses important selectivity patterns. In allowing penalty loadings to vary across coefficients, I also find an increased coefficient on a less than high school education in Illinois. In general though, with these exceptions in mind, the results in Table 2.9 and Table 2.10 are similar and capture selectivity bias in the same direction. A more flexible penalty loading then yields conclusions about selection bias similar to baseline estimates.

Multiple Control Functions

The baseline results also impose the restriction of a single control function for each birth state. There are clear reasons for imposing this restriction. Firstly, in many empirical instances it is not desirable to allow for a separate control function for each birth state. If the sample of individuals in some birth states is small, then relatively few control terms will be included to identify selection for these states. As discussed in chapter 1, this makes it less likely that selectivity is well characterized for individuals from these locations. The results of this section make clear the limits sample size places on variable selection and hence estimation. Additionally, relaxing this assumption and allowing separate control functions for subsets of states is somewhat arbitrary and imposes similarly strict assumptions on selectivity bias. In this section, I allow for a separate control function for movers and stayers, separate control functions by region of birth, and a modified control function which is the same across states except for terms relevant to the state of birth.

In Table 2.11, I implement a separate control function for stayers and movers. As the control function for each group is identified using a distinct sub-sample, I select terms separately for each group using PDL. In general, more cells are selected for movers than stayers, despite the relatively large population of stayers in most states (except Florida). Despite this, the number of terms selected for each group separately is lower than the total number in Table 2.9. If separate control functions are not required, then this specification will fail to properly account for selectivity bias, will include too few terms, and hence be too restrictive to capture selectivity bias for each group. This is a particular problem for stayers as only 70 terms on average are included to capture selectivity patterns for this group.

Across all six states (except Kansas), the coefficients on a college education and an advanced

Table 2.10: Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Controlling for Heteroskedasticity and Weighted Penalty Loadings

	CA	FL	IL	KS	NY	TX
LHS	-0.126***	-0.129***	-0.138***	-0.192***	-0.133***	-0.179***
	(0.012)	(0.014)	(0.021)	(0.035)	(0.021)	(0.016)
Some College	0.097***	0.110***	0.116***	0.001	0.150***	0.118***
	(0.008)	(0.011)	(0.014)	(0.025)	(0.014)	(0.010)
College	0.359***	0.366***	0.288***	0.374***	0.357***	0.462***
	(0.014)	(0.021)	(0.026)	(0.050)	(0.026)	(0.020)
Advanced	0.481***	0.512***	0.437***	0.419***	0.456***	0.615***
	(0.025)	(0.034)	(0.045)	(0.084)	(0.039)	(0.036)
Exper	0.084***	0.044***	0.066***	0.033	0.081***	0.050***
	(0.011)	(0.015)	(0.014)	(0.030)	(0.012)	(0.013)
Exper^2	-0.004***	-0.001	-0.003**	0.001	-0.005***	0.000
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$\text{Exper}^3 \times 100$	0.007*	-0.001	0.006	-0.004	0.010**	-0.007
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.004)
Married	0.122***	0.125***	0.147***	0.120***	0.132***	0.136***
	(0.009)	(0.011)	(0.015)	(0.025)	(0.016)	(0.010)
SMSA	0.181***	0.110***	0.260***	0.241***	0.234***	0.126***
	(0.011)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
# Cells	357	349	243	171	246	312
R-Squared	0.155	0.176	0.206	0.199	0.202	0.202
Obs	46727	25579	26533	6234	36765	34890

Notes: Ideal penalty loadings calculated using an algorithmic method described in text. Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

Table 2.11: Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Separate Control Function for Stayers and Movers

	CA	FL	IL	KS	NY	TX
LHS	-0.106***	-0.130***	-0.187***	-0.209***	-0.139***	-0.140***
	(0.016)	(0.014)	(0.024)	(0.038)	(0.030)	(0.017)
Some College	0.118***	0.101***	0.089***	0.003	0.140***	0.135***
	(0.010)	(0.010)	(0.017)	(0.024)	(0.019)	(0.012)
College	0.413***	0.378***	0.366***	0.351***	0.458***	0.512***
	(0.017)	(0.020)	(0.027)	(0.038)	(0.026)	(0.022)
Advanced	0.533***	0.554***	0.466***	0.398***	0.580***	0.626***
	(0.029)	(0.034)	(0.043)	(0.062)	(0.044)	(0.034)
Exper	0.081***	0.040**	0.060***	0.028	0.075***	0.054***
	(0.012)	(0.015)	(0.014)	(0.030)	(0.012)	(0.013)
$Exper^2$	-0.003***	-0.000	-0.002	0.001	-0.004***	-0.000
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$Exper^3 \times 100$	0.006	-0.003	0.003	-0.005	0.008*	-0.006
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.005)
Married	0.106***	0.150***	0.168***	0.129***	0.151***	0.145***
	(0.010)	(0.010)	(0.015)	(0.024)	(0.016)	(0.011)
SMSA	0.183***	0.111***	0.259	0.234***	0.231***	0.128***
	(0.011)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
# Cells Stayer	89	74	73	75	74	88
# Cells Mover	179	162	75	33	77	141
# Total Cells	278	246	158	118	161	239
R-Squared	0.156	0.175	0.208	0.194	0.203	0.205
Obs	46727	25579	26533	6234	36765	34890

Notes: Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

Table 2.12: Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Separate Control Function for Each Birth Region

	CA	FL	IL	KS	NY	TX
LHS	-0.189***	-0.213***	-0.217***	-0.246***	-0.201***	-0.206***
	(0.009)	(0.012)	(0.012)	(0.026)	(0.011)	(0.011)
Some College	0.115***	0.102***	0.101***	0.081***	0.103***	0.104***
	(0.007)	(0.009)	(0.009)	(0.020)	(0.008)	(0.008)
College	0.411***	0.415***	0.410***	0.389***	0.400***	0.394***
	(0.010)	(0.014)	(0.014)	(0.030)	(0.012)	(0.012)
Advanced	0.528***	0.546***	0.550***	0.443***	0.549***	0.527***
	(0.018)	(0.024)	(0.025)	(0.050)	(0.021)	(0.020)
Exper	0.104***	0.085***	0.087***	0.074**	0.089***	0.092***
	(0.012)	(0.016)	(0.016)	(0.032)	(0.013)	(0.014)
$Exper^2$	-0.005***	-0.004**	-0.004**	-0.002	-0.004***	-0.004***
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$Exper^3 \times 100$	0.011***	0.007	0.007	0.002	0.007	0.008*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.005)
Married	0.136***	0.147***	0.144***	0.122***	0.148***	0.140***
	(0.007)	(0.008)	(0.009)	(0.018)	(0.008)	(0.008)
SMSA	0.133***	0.084***	0.097***	0.052***	0.127***	0.126***
	(0.006)	(0.007)	(0.007)	(0.013)	(0.006)	(0.006)
# Cells NE	46	78	5	0	207	19
# Cells MW	79	68	185	95	13	57
# Cells South	48	138	9	8	6	186
# Cells West	208	11	1	9	5	29
Total # Cells	391	305	210	122	241	301
R-Squared	0.199	0.181	0.179	0.184	0.196	0.202
Obs	46727	25579	26533	6234	36765	34890

Notes: Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages. Separate control function included for states in the Northeast, the Midwest, the South, and the West.

degree are larger when allowing for two control functions when compared to the baseline estimates. Lower estimated selectivity bias here could be indicative of there being fewer included terms with which to capture selection bias for both groups, but especially stayers. This pattern also holds for the returns to some college education in California, Florida, New York, and Texas. The coefficient on a less than high school education is indicative of increased selection bias for some states, and decreased selection bias in others. This could be due to the control function for stayers being especially poorly estimated, and stayers having disproportionately low educated.

In moving from a single control function to two control functions, Dahl introduces greater flexibility in estimation. However, using my approach, allowing for separate control functions actually serves to restrict the number of terms used to pick up selection bias in the model. In allowing for two control functions, Dahl then picks up more selection bias, perhaps indicating that the inclusion of more terms is needed to capture selectivity for stayers and movers. Conversely, in using separate control functions for stayers and movers, less selectivity is predicted here as the variation used to capture selection patterns is muted. In comparing results between Tables 2.6 and 2.7, and across Tables 2.9 and 2.11 it is clear that implementing separate control functions serves to limit the amount of selectivity bias captured.

There is no clear theoretical reason why a separate control function for stayers and movers is preferable to say, a separate control for those born in Colorado and all others. In Table 2.12 I instead permit a separate control function for each region of the United States, including four control functions in total. This places a geographic restriction on the control function, such that those from the same region of the country are assumed to exhibit similar selectivity patterns. As is clear in this table, this restriction greatly limits the number of terms used to identify selection bias in those moving from regions far away. This is likely due to the sample size of these individuals being relatively small, however, it is possible that fewer terms are required to identify selectivity for these sub-populations. In general though, the inclusion of fewer terms with which to identify selectivity makes it harder to justify a conclusion the control function is well approximated, that is, that this limited set of terms defines a sparse representation. For Illinois, Kansas, and New York, including separate control functions by region amounts to assuming that selectivity can be well characterised using fewer than 10 control terms. This seems unlikely given that the choice over where to live is no less complex for these individuals than for individuals living in nearby states, and is no less linked to educational attainment.[18] It is perhaps no surprise then to see increased coefficient estimates on the returns to a college or advanced degree, although, in contrast to all previous estimates, the opposite is true for Texas. Estimates of the return to a less than high school education are also much lower than all previous

[18] See the transition matrices which makes this clear.

estimates and indicate upward bias in all OLS estimates. In general these results predict patterns of selection bias at odds with all prior specifications.

I present a final specification in Table 2.13 which is perhaps most similar in flavour to the restriction in Dahl, which instead of introducing a separate control function for each birth state, includes a separate migration probability term for the state of birth and allows this function to differ for the state of residence. In this manner, I include a separate migration probability for the 'staying' probability for those who do not leave their birth state, and the 'retention' probability for those who do. This function then allows for migration probabilities to capture the same selectivity across all individuals, conditional on not being born in that state. Instead, I allow for the probability of staying in the birth state to have a separate effect, and assume that this effect is identical across all states. The probability of remaining in Texas and New Hampshire has the same relationship to selectivity. This specification therefore relaxes the assumption of separate control functions by birth only partially and in a manner similar to Dahl (2002).

Though the coefficient estimates differ from the baseline estimates, these is no clear pattern in the direction of the difference, suggesting that this specification does not simply restrict the identifying variation used to capture selection bias. Instead coefficients are at times higher and lower, though in general the direction of effects aligns closely to estimates in Tables 2.9. This specification captures more upward bias in the returns to a college education in California, Illinois, and Texas, but suggests less upward bias is present in estimates for the other three states.

Data limitations often justify a necessary restriction on the distribution characterizing selectivity, such that estimation proceeds with a limited set of control functions defined across birth states. Without such a restriction, the number of terms chosen to characterize selectivity will be exceedingly small for some birth states, making it unlikely that selectivity is well captured. As stressed above, this is not necessarily the case, but it implies that the data permits a sparse representation defined by only a small handful of control terms. Deciding in what manner to restrict the set of control functions is not unimportant, as the results of this section make clear, but can be somewhat arbitrary. It may be hard to justify implementing a separate control function by birth region for instance. Additionally, the restriction should be chosen, keeping in the mind the likelihood that a sufficient number of terms are included such that each separate control function is well characterised, which as mentioned above is less likely to be true for the regionally defined control functions.

Implementing selection correction using a single control function imposes admittedly a strong restriction on the data. However, when the data does not permit the inclusion of a separate control function for each birth state it is difficult to outline a manner in which this restriction could be relaxed. Grouping birth locations by distance, or implementing separate control functions for stayers and movers need not better characterize selectivity, and are somewhat

Table 2.13: Corrected Estimates for Selected States using post-double-Lasso and Random Forest Estimates of Migration Probabilities: Modified Control Function

	CA	FL	IL	KS	NY	TX
LHS	-0.114***	-0.124***	-0.157***	-0.213***	-0.171***	-0.180***
	(0.012)	(0.013)	(0.019)	(0.030)	(0.019)	(0.014)
Some College	0.068***	0.098***	0.106***	0.035*	0.126***	0.117***
	(0.008)	(0.009)	(0.012)	(0.020)	(0.014)	(0.009)
College	0.338***	0.365***	0.311***	0.378***	0.394***	0.440***
	(0.014)	(0.019)	(0.023)	(0.041)	(0.026)	(0.018)
Advanced	0.443***	0.495***	0.467***	0.515***	0.545***	0.575***
	(0.025)	(0.032)	(0.039)	(0.066)	(0.038)	(0.030)
Exper	0.084***	0.045***	0.069***	0.030	0.079***	0.047***
	(0.011)	(0.015)	(0.014)	(0.030)	(0.012)	(0.013)
$Exper^2$	-0.004***	-0.001	-0.003**	0.001	-0.004***	0.001
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$Exper^3 \times 100$	0.007*	-0.001	0.006	-0.004	0.009**	-0.008*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.004)
Married	0.115***	0.127***	0.124***	0.149***	0.138***	0.142***
	(0.009)	(0.010)	(0.012)	(0.022)	(0.015)	(0.010)
SMSA	0.182***	0.110***	0.258	0.237***	0.234***	0.125***
	(0.011)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
# Cells	204	156	140	80	160	170
R-Squared	0.154	0.175	0.203	0.191	0.202	0.202
Obs	46727	25579	26533	6234	36765	34890

Notes: Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages. Stayers' control function includes flexible interactions with the first best probability (the retention probability) and movers' control function includes flexible interactions with the retention probability. All other probability terms enter similarly for movers and stayers.

arbitrary choices. Additionally, in further segmenting the data, the number of terms included to capture selectivity for these groupings can become small, and hence reduce the likelihood that each control function is well characterized.

In conclusion, data limitations may mean that it is not desirable to include a separate control function for each birth location. Including separate control functions across sub-populations should be preferred to a single control function only when researchers have a strong a priori justification for why variation in subutilities correlates with selectivity differently across states. Additionally, each control function included must permit a sparse representation such that a limited set of terms characterises selectivity. Selectivity is likely better characterized using a larger number of terms, even when the control function permits a sparse approximation, and hence the size of the sub-population must be sufficiently large as to permit inclusion of a sufficiently large set of terms.[19] Therefore, even when a priori reasons for grouping birth states does exist, the sample size of groupings may not permit flexible characterization of selectivity. The empirical exercise of this section makes clear that implementing separate control functions by birth region, and to a lesser extent, for stayers and movers, greatly limits flexible characterization of selectivity for these sub-groups. My preferred estimates remain those which include a single control function for all individuals regardless of birth state.

2.5 Exploring the Selected Control Terms

The empirical estimates presented above provide evidence of the importance of a flexible control function of migration probabilities for dealing with selection bias. Crucially, these results suggest the SISA may not hold and that estimates based on making this assumption may not be reliable. In this section, I examine the terms selected by the two-step procedure as it may shed light on factors affecting variable selection. In general, it is difficult to analyse patterns of selection based on these terms as they represent a complex relationship between interconnected choices.[20]

[19] It is of course possible that a sparse representation exists, consisting of just a few control terms, but this is unlikely when the patterns of selectivity in the data are rich.

[20] It is also worthwhile to point out that there are a number of caveats that may limit the importance of such an exercise. For instance, if the probability of moving to two locations is strongly correlated, then perhaps only one of those probabilities will be included. This does not mean that the probability of moving to the other state is any less relevant for selectivity. Additionally, the lasso methodology is not strictly selecting terms only relevant for selectivity bias. Instead, terms are being chosen that are potentially strong candidates for being the drivers of selectivity, that is, there are strong differences in preferences for where to live based on educational attainment.

Table 2.14: Summary of Terms Selected

	California	Florida	Illinois	Kansas	New York	Texas
First Best	0.077	0.038	0.075	0.096	0.080	0.065
New England	0.810	0.868	0.851	0.798	1.044	0.939
Mid Atlantic	1.748	1.735	1.802	1.697	2.329	1.799
East N-C	1.151	1.425	1.982	1.317	1.036	1.319
West N-C	0.767	0.881	0.815	1.454	0.792	0.822
West S-C	1.023	1.172	1.001	0.898	1.098	0.973
S. Atlantic	0.863	1.164	0.976	0.524	1.024	1.199
East S-C	0.991	0.993	0.901	1.647	1.054	1.589
Mountain	0.804	0.548	0.579	0.556	0.568	0.514
Pacific	1.304	0.658	0.661	0.539	0.771	0.624

Notes: Row 1 presents the share of control terms which include the first best probability. Evenness across selection would imply a value of 0.02. Rows 2-10 display the share of control terms related to each division scaled by the total share of states in that division. A number greater than one represents over-representation of terms from this division relative to an even selection. These results are calculated using the selected terms from Table 2.9.

Raw Patterns

The simplest way to summarise these results is to consider the share of terms selected which are constructed using the probability of moving to a particular state. To make matters simpler, I aggregate these shares by division. In Table 2.14, I express this share relative to the share of states in the division.[21] A value above one then signifies over-selection of terms in that division relative to randomness. In row one, I present the raw share of terms selected that contain the first-best migration probability. Terms containing the first-best probability make up just 2% of all terms and so the first clear takeaway from Table 2.14 is that the first-best probability does have particular importance in controlling for selection bias. Of the six states considered, Florida represents a lower bound, including twice as many terms using this probability as randomness would suggest, while Kansas defines the upper limit, selecting 5 times more of these terms. This result for Florida could be explained by the relatively high proportion of residents born outside the state, dampening the role played by the probability of staying in Florida. The importance of the first-best probability may provide an explanation for why Dahl's method tends to predict the correct direction of selection bias at least for the most educated groupings on average. However, given that at least 90% of selected terms do not contain this probability, there is clearly a degree of selectivity that this term alone cannot capture.

A second important takeaway from Table 2.14, is that the distribution of selected terms is not highly concentrated in a few divisions and for all states, terms are selected for inclusion across all divisions. Despite this, there is of course variation across states in which regions tend to be over-represented. Clearly terms are over-selected (relative to random selection) from the division of residence. This is explained in part by the first-best probability, but not entirely. In Figure 2.6 and appendix Figures B.3 - B.7, I present choropleth maps showing the relative selection of terms by state. In panels (a) and (b) of each figure, there is clear evidence of a selection preference for nearby states even when the home state is not considered.

Another pattern worthy of remark is the clear importance of terms constructed using the probability of moving to the the mid-Atlantic or the East North-Central. In contrast, terms related to the probability of living in the West of the U.S. are chosen much less often. In explaining these patterns in control variable selection it is important to keep in mind how terms are being selected using PDL. Terms are selected either due to their ability to improve model fit, or due to their correlation with educational attainment. Differences in preferences across states by educational attainment is what will bias results here. If there was no systematic difference in preferences by educational attainment, then unobserved earnings draws would not correlate with education, and our estimates of the returns to earnings would be unbiased. I explore

[21] This is the same as expressing the share relative to the share of all control variables containing probabilities of moving to states in this division.

the role of migration patterns, and more specifically, differences in migration by educational attainment in explaining variable selection.

Relation to Migration Patterns

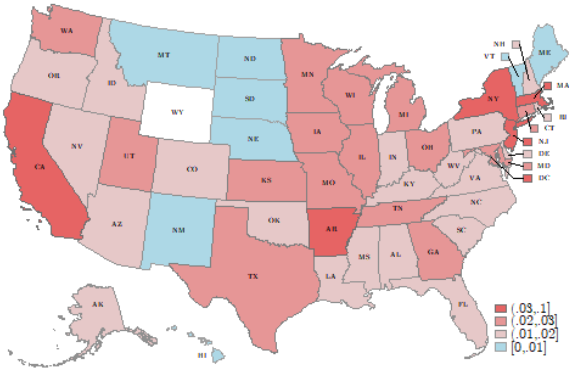
In panels (c) and (d) of Figure 2.6 and Figures B.3-B.7, I present migration patterns for each division. Specifically, I calculate the share of movers by place of origin and by destination respectively. It is immediately clear that there are differences in migration patterns by state which is consistent with conclusions drawn from the transition matrices. In general, there are clear preferences for locations nearby, although some areas, such as California, are a popular destination regardless of distance. For inmigrants, one constant across states is the relatively high share of migrants arriving from the Eastern-North Central and the mid-Atlantic.[22]

Given these flows and the patterns presented in Table 2.14, an obvious question is whether these migration patterns are in some way correlated with how terms are selected using PDL. For instance, simple intuition would suggest that as individuals are more likely to move into states nearby, those states are likely most informative regarding selection patterns. For instance, for New York over 50% of migration flows are to states in the South Atlantic and the mid-Atlantic. It makes intuitive sense that these terms are more informative regarding the selectivity of movers from New York, and furthermore, that there is likely to be substantial variation in these probability terms. This is only true if there is more variation in local migration probabilities across educational groupings. Given that there are less flows to and from states far away, it is likely that these flows are driven by large unobserved shocks, and unlikely to capture a distinct difference in preferences between education classes, and hence pick up differential selection on unobserved ability.

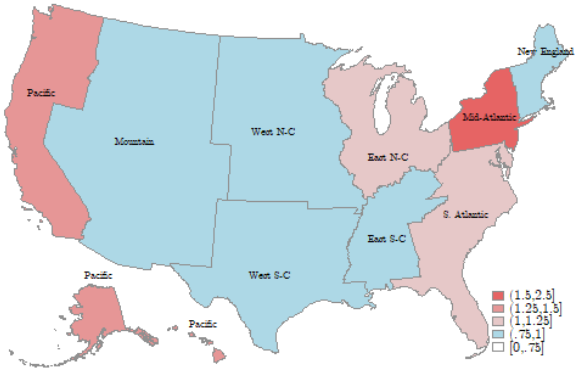
Given the importance of selectivity by education, it is worthwhile to consider also patterns in migration related to education. In Table 2.15, I present the average share of individuals with at least a college education by division of birth and division of residence for outmigrants and inmigrants. Immediately clear from this table is the existence of substantial within state variation in the educational attainment of arrivals across division of birth. For instance, only 31% of immigrants from East-South-Central to California have a college education or higher, versus 46% of inmigrants from the mid-Atlantic. These patterns also differ widely across states; in contrast to California, those who enter New York from the East-South-Central are generally more educated than arrivals from other regions. Of the six states considered, migrants to Florida have the lowest educational attainment on average, while migrants to New York are on average better educated. Patterns for Illinois are noteworthy as those who are arrive from New England and the mid-Atlantic are appreciably higher educated than arrivals from other states.

[22]Kansas being a notable exception here.

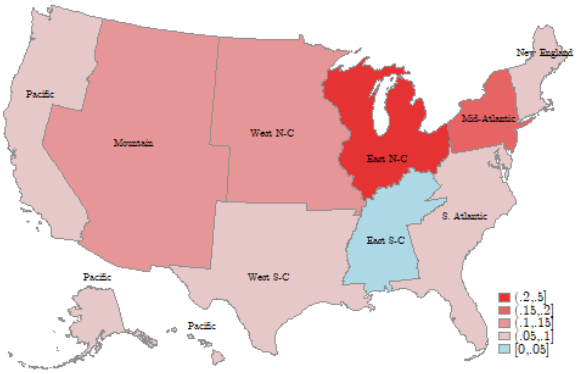
Figure 2.6: California: Selected Terms and Migration Patterns



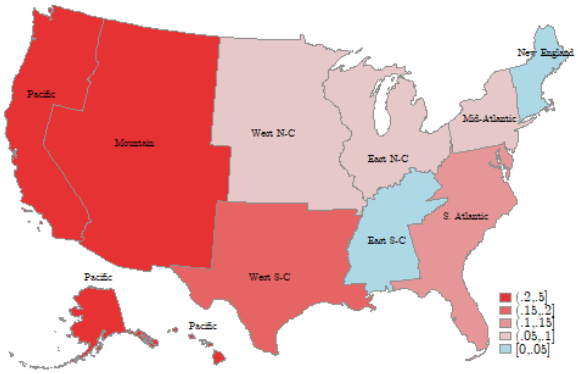
(a) Selected Terms by State



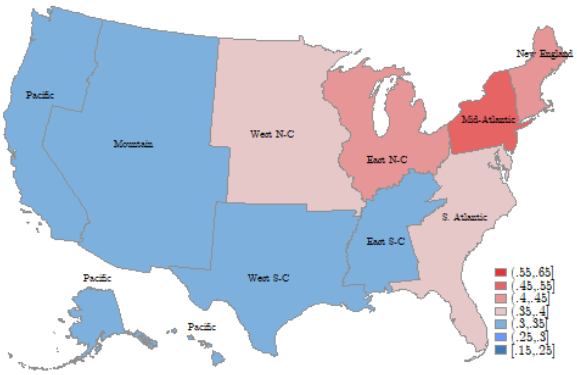
(b) Selected Terms by Division



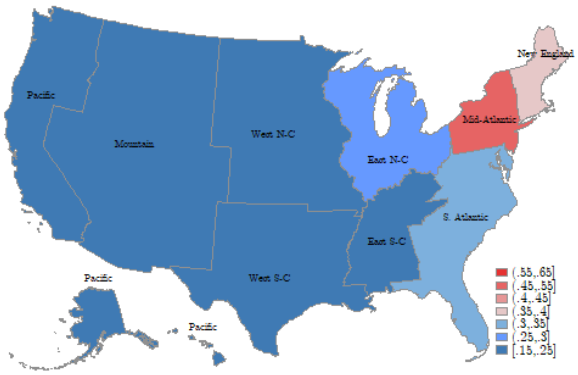
(c) Birthplace Share of Inmigrants



(d) Destination Share of Outmigrants



(e) BA share of Inmigrants by Birthplace



(f) BA share of Outmigrants by Destination

Notes: Panel (a) displays the share of control terms selected by PDL which are calculated using the migration probability from each state. Panel (b) calculates this share at the division level (excluding California) and scales by the total share of states in this division. Panel (c) displays the share of immigrants to California from each birthplace. Panel (d) shows the destination share of migrants out of California. Panel (e) plots the share of immigrants entering California from each division who have a College Education. Panel (f) displays the College Educated share of those entering each division who originated in California.

Table 2.15: Share of Migrants with a College Education

Division	CA	FL	IL	KS	NY	TX	Stayers	All
	Share of College Educated Inmigrants						Div Share	
New England	0.45	0.23	0.60	0.41	0.46	0.43	0.24	0.32
Mid Atlantic	0.46	0.27	0.63	0.41	0.46	0.44	0.24	0.28
East N-C	0.41	0.23	0.46	0.36	0.55	0.36	0.18	0.21
West N-C	0.36	0.29	0.38	0.35	0.52	0.40	0.17	0.21
West S-C	0.39	0.26	0.37	0.21	0.45	0.37	0.17	0.25
S. Atlantic	0.34	0.23	0.24	0.19	0.43	0.37	0.15	0.19
East S-C	0.31	0.29	0.31	0.25	0.52	0.35	0.19	0.24
Mountain	0.34	0.27	0.39	0.24	0.45	0.33	0.17	0.24
Pacific	0.35	0.28	0.36	0.22	0.51	0.27	0.23	0.28
	Share of College Educated Outmigrants						Div Share	
New England	0.40	0.34	0.61	0.38	0.59	0.44	0.24	0.32
Mid Atlantic	0.46	0.30	0.55	0.45	0.46	0.43	0.24	0.28
East N-C	0.30	0.24	0.32	0.38	0.54	0.23	0.18	0.21
West N-C	0.23	0.24	0.37	0.23	0.43	0.29	0.17	0.21
West S-C	0.34	0.29	0.37	0.37	0.38	0.32	0.17	0.25
S. Atlantic	0.23	0.24	0.23	0.33	0.41	0.24	0.15	0.19
East S-C	0.19	0.30	0.34	0.30	0.42	0.23	0.19	0.24
Mountain	0.21	0.30	0.36	0.30	0.35	0.27	0.17	0.24
Pacific	0.22	0.34	0.42	0.32	0.47	0.30	0.23	0.28
	Share of College Educated in State							
Stayers	0.24	0.18	0.23	0.18	0.27	0.20		
All	0.31	0.23	0.28	0.22	0.30	0.26		

Notes: The share of inmigrants measures the share of individuals that move from a division to a state who have at least a college education. The share of outmigrants captures the share of individuals that move from a state to a division who have at least a college education. Rows 8 and 9 give the total share of college educated individuals in each division. Row 8 calculates this share for those born in the division who reside there, and row 9 presents this value for all residents in the division. The final two rows in the table present the share of college educated individuals for each state. This is calculated either for the entire resident population, or for those born in the state who remain resident there.

Flows from these divisions to other states are also markedly lower in educational attainment.

Patterns in educational attainment of outmigrants are similar to those of inmigrants. For example, migrants from Illinois are in general much more highly educated when moving to New England and the mid-Atlantic. Furthermore, outmigrants from New York are on average better educated than outmigrants from other states. Panels (e) and (f) of Figure 2.6 and Figures B.3-B.7 provide a visual summary of these flows. These patterns represent distinct differences in preferences and utility for particular migration paths. For instance, the average educational attainment of individuals living in New York and California tends to be similar. Despite this, those who leave New York tend to be on average more highly educated regardless of destination. That is, the average New Yorker in most states is on average more educated than the average Californian, despite having relatively similar education levels in these states. Educational attainment in Illinois is on average lower than in California, but those born in Illinois are on average more educated than their California born counterparts in many regions.

Using PDL, control terms are selected based on their correlation with educational attainment across individuals in the state of residence. If there is a positive correlation between educational attainment in California and the probability of migration to New York, then a relatively high share of control terms created using the migration probability of New York should be selected. Table 2.15 makes clear that there is substantial variation in preferences over migration trajectories across education classes. Given differences in the educational attainment of migrants to and from a state and by state of origin or destination, the correlation between education and migration patterns will be very different for each state, and hence we should expect corresponding variation across states in which control terms are selected.

The reason why variation in migration probabilities across education classes is important for picking up selection bias is intuitive. If those with a BA who are living in California have a higher probability of moving to New York than those with a high school diploma, then the average unobserved shock necessary to induce living in New York is higher for the high school educated. To induce a move to California then, requires a relatively high California specific earnings shock for the BA educated in this instance.[23] If there was no variation in the probability of migration to New York across education classes, then this variable is not informative regarding differential selection into California. This simple intuition motivates selecting terms based on their correlation with educational attainment, though in practice the complex multidimensionality of the migration decision does not permit such a straightforward interpretation.

To test whether the selection of control terms is indeed related to migration flows and the relative educational attainment of migrants, I regress the measures presented in Table 2.14, and

[23]For any given New York specific earnings shock, preferences for New York are higher for the BA educated, and hence a higher California specific earnings shock is necessary to induce migration to California.

Table 2.16: Factors Affecting Control Term Inclusion

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
High Educated Share of Inmigrants	0.46**			0.43**	0.46***	0.47**	0.48***
	(0.23)			(0.17)	(0.17)	(0.19)	(0.18)
High Educated Share of Outmigrants	0.40			0.86***	0.90***	1.11***	0.94***
	(0.30)			(0.23)	(0.23)	(0.26)	(0.23)
Share of Inmigrants		3.02***		2.88***	2.41***	2.39***	2.80***
		(0.19)		(0.21)	(0.24)	(0.24)	(0.28)
Distance			-0.13***	-0.04***	-0.03*	-0.02	-0.02
			(0.01)	(0.01)	(0.01)	(0.02)	(0.01)
Own Division					0.33***	0.33***	
					(0.09)	(0.09)	
Low Educated Share of Outmigrants						0.95	
						(0.61)	
Low Educated Share of Inmigrants						0.062	
						(0.58)	
Constant	0.75***	0.71***	1.33***	0.36***	0.33***	0.16	0.26**
	(0.10)	(0.03)	(0.04)	(0.10)	(0.10)	(0.16)	(0.10)
Observations	450	450	450	450	450	450	400
R^2	0.02	0.36	0.15	0.42	0.44	0.44	0.28

Notes: Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Dependent variable is the fraction of control terms selected by PDL from each division, divided by the total share of terms from that division in the full specification. Independent variables are the distance between the state and the division, the share of high and low educated immigrants out of the state to each division, and the share of low and high educated immigrants into the state from each division. A dummy variable for own division is included and also included is the share of migrants to the state originating from each division. In column (7), flows between a state and its own division are excluded.

calculated for all states, on the educational attainment of outmigrants and inmigrants to and from each division respectively, the share of inmigrants from each division, and the distance in kilometers between the state and division centroids. In Table 2.16 I present the results of this exercise.

In column one, I consider just the role played by the educational attainment of inmigrants and outmigrants. I find a positive and significant relationship between the educational attainment of inmigrants and the probability of selecting terms associated with the sending division. In column 2, I find a positive relationship between selection and the share of inmigrants. Relative to random selection, a 5% increase in the share of inmigrants arriving from this division is associated with a 15% increase in the selection of terms from this same division. In column 3, I consider the role of distance, finding that fewer terms associated with the probability of moving to states far away are chosen.

In column 4, I estimate all of these effects jointly, finding a decreased role for distance, and a positive and significant relationship between variable selection and the educational level of outmigrants. In specification 5, I include a dummy variable which takes value one if the state under consideration belongs to the same division under consideration, which further reduces the effect of distance. In columns (7) and (6), I find no evidence that variable selection is correlated with the share of the least educated moving across states, and find that my preferred specification in column (4) is robust to dropping the division to which each state belongs.

These results provide evidence that flows by educational attainment and the relative size of flows are important factors influencing which terms are selected by Lasso. This is important as it is variation across migration probabilities by education class that captures differential selection on unobservables. Despite the correlation between term selection, and differential migration choices by educational attainment, it is important to stress that for each state, term selection is not strongly concentrated geographically. In general, control variables are selected via the PDL procedure which use variation in the probability of moving to all parts of the US. This high-dimensional choice problem then is best identified using variation in preferences across all 50 choices.

2.6 Wage Inequality

A large literature has documented the growth in the college wage premium in the US between 1980 and 2000 (Murphy and Welch (2001)). Diamond (2016) and Moretti (2013) amongst others have established the role played by location, whereby college educated workers tend to locate in large metropolitan areas. Moretti (2013), shows that in accounting for the local cost of living the increase in the college premium is less severe than naive OLS estimates would suggest. Given

both the distinct differences in sorting by college and high school educated workers, and the evidence of upward bias in OLS estimates of college premium at the state level, it is natural to ask how selection bias affects estimates at the national level. In this section, I determine how much of the increase in the college premium at the national level between 1980 and 2000, is the result of differences in sorting by ability. I compare estimates obtained using Dahl's correction method, and the improved method outlined in this contribution to highlight the importance of appropriate selection correction.

Traditional estimation of the college premium has proceeded by regressing individual log earnings on education dummies, controls for observable characteristics, and location (Moretti (2013)). Given the importance of distinct returns across states in this chapter I estimate the national college premium by aggregating state-level estimates. Specifically, I weight the college premium in each state by the national share of college workers residing in the state. This measure therefore captures the average premium over high school workers earned by college workers. As before I limit the sample to white males aged 24-35, born in the United States. To make comparisons over time I obtain estimates from the 5% public use samples of the 1980 and 2000 US census. Uncorrected and corrected estimates are obtained using the exact same procedures outlined above with regards the 1990 census.

In Table 2.17 I present estimates of the national college premium for each year. In 1980 there is clear evidence of upward bias in the OLS estimate due to selectivity across states. The degree of upward bias is overstated by Dahl's method, but using my preferred specification I still find evidence that the college premium is 9.5% lower than OLS estimates would suggest. In 1990 Dahl's estimates again overstate the upward bias in OLS estimates. My preferred specification estimates the national average premium to be 12% lower than a simple aggregation over OLS estimates. The same is true for 2000, with my preferred specification estimating a national college premium 9% lower than OLS. In general, selection bias at the state level translates into bias at the national level. Dahl's method overstates this bias in all years, while my preferred specification finds that in general using OLS estimates leads to an upward bias of the college premium of around 9-12%.

In the second panel of Table 2.17 I present the percentage growth in the national college premium. It is immediately clear that most of the increase in wage inequality over this period occurs between 1980 and 1990. Over the entire period, 1980-2000, my preferred specification yields a growth estimate of 75% which is very similar to the 73.8% obtained using OLS estimates. This is perhaps unsurprising given that the magnitude of selection bias is approximately equal in 2000 and in 1980, meaning the PDL estimate is roughly 9% lower than the OLS estimate in both years. However, given the increased selectivity in 1990 the growth rate between 1980-1990 and 1990-2000 is not the same as the OLS estimate. Specifically the growth rate is 5%

Table 2.17: National Level Estimates of the College Earnings Premium

	OLS	DAHL	PDL
	Weighted Average College Premia		
1980	0.252	0.221	0.228
1990	0.409	0.345	0.361
2000	0.438	0.387	0.400
	Changes in the College Premia		
1980 - 1990	0.624	0.559	0.579
1990 - 2000	0.071	0.122	0.108
1980 - 2000	0.738	0.750	0.751
	Decomposition: Fixed Population		
1990	0.407	0.341	0.360
2000	0.444	0.385	0.398
	Decomposition: Fixed Premia		
1990	0.258	0.226	0.232
2000	0.408	0.343	0.361

Notes: National level estimates of the college premium are constructed by taking a weighted sum over state level estimates, using the share of the college population in each state. In performing the decomposition I restrict either this population share, of the state college premia at the prior year value.

lower between 1980-1990 and 3% higher between 1990-2000 using my improved method. For 1990-2000 this increase is 53% larger than the increase estimated using OLS. In examining wage inequality over this period, self-selection plays an important role, with sorting on ability across states accounting for 9-12.5% of the college premium aggregated at the national level.

The aggregated measure of the national college premium used here is composed of two distinct components. In the final two panels of Table 2.17 I perform a simple decomposition exercise, fixing either the share of college workers residing in each state, or the state college premium, at the previous years value. This decomposition is informative about the relative importance of changes in wage premium at the local level, and changes in the distribution of workers across states. In first considering fixed population shares, it is immediately clear that between 1980-1990 and 1990-2000 the bulk of the increase in the national college premium is explained by increases in the relative wage premium at the state level. When holding the premium constant, there is essentially no change in the national level college premium. In general, the relocating of young, white, college educated males across states does not explain the average increase in their earnings relative to high school educated workers over 1980-2000. Instead, changes in the college premium at the state level account for almost all of the increase in the national average college premium.

However, there is still a role for location in explaining the national level increase in the college premium over this period. Firstly, the increase in the college premium at the state level could be explained by increased clustering of college educated workers in large metropolitan areas within a state (Moretti (2013)).[24] Secondly, composition changes in terms of the relative skill distribution of workers likely drives changes in premium over time. The location decisions of workers is related to the relative supply of workers and hence movements in the wage premium. Finally, these is also substantial variation in growth of the college premium at the state level. College educated workers are on average concentrated in states which have experienced relatively high growth in the college wage premium over 1980-2000. In summary, though Table 2.17 suggests the location of workers plays little role in driving national wage inequality, the relative supply and demand for educated workers, and hence the sorting of workers, is an important driver of changes in local wage premium and therefore national wage inequality.

In conclusion, I present evidence that OLS estimates overstate wage inequality at the national level by 9-12% over the period 1980-2000 for the restricted sample considered in this chapter. This upward bias is caused by the self-selection of workers across states. I further show that most of the increase in the national college premium over this period, is due to increases in the college premium at the local level, and that these increases are relatively high in the states

[24]The method presented in this chapter could easily be extended to obtain estimates of the college premium at the metropolitan areas or at any alternative level of geography.

Table 2.18: Migration Flows and Differences in Returns

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Intercept	0.62***	0.64***	0.69***	0.63***	0.63***	0.62***	0.71***
	(0.02)	(0.02)	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)
Δ Corrected		0.82***	0.52***	1.57***	0.92***	1.12***	1.24***
College Return		(0.20)	(0.21)	(0.26)	(0.27)	(0.22)	(0.32)
Δ Uncorrected	1.07***						
College Return	(0.23)						
N	1341	1341	1341	1341	1341	1341	1341
R^2	0.015	0.011	0.023	0.093	0.085	0.066	0.143
Quality of Life				YES			YES
Climate					YES		YES
State Budget						YES	YES
Distance			YES				YES
Δ Unemp. Rate			YES				YES
F-test for				18.55	17.17	13.84	10.44
Amenity Variables				(0.00)	(0.00)	(0.00)	(0.00)

Notes: Robust standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Dependent Variable is the log difference in migration flows between states for college educated and high school educated individuals. Δ operates on differences in premia across states for college and high school educated individuals. The corrected returns use here come from the post-double-Lasso estimates. See B.1.1 for the list of included amenity variables and other control variables.

where college workers concentrate.

2.7 Locational Choice and Returns to Education

In this section, I estimate the locational response of the college educated to differences in returns across states. This exercise is similar to that of Dahl (2002) and Ransom (2016) and uses cross birth state differences in migration flows between the college and high school educated to measure the importance of the returns to education in driving migration across states. Evidence finding a positive relationship between mean earnings and migration flows is supportive of the Roy model presented above, whereby individuals move based on relative earnings opportunities. A finding of no effect would suggest people do not strongly value differences in earnings across states and hence, is evidence against a finding of selection bias: if individuals do not respond to earnings differences then earnings shocks are more likely randomly assigned across states. This exercise mirrors that in Dahl (2002), but it is important to reconsider these results given the improved estimates of the returns to schooling presented in this contribution.

I assume that the following log-linear model appropriately captures the relationship between migration flows and state-specific education premium where the rate at which individuals transition from their birth state j to state k is a function of the difference in earnings w_{cjk} where c indexes college educated workers. Furthermore, differences in state specific amenities (A_k) and $j-k$ specific moving costs (D_{jk}) affect the likelihood of transition.

$$ln(P_{jc}^C) = \pi_0^C + \pi_1^C(w_c^C - w_j^C) + \pi_3^C(A_c - A_j) + \pi_3^C D_{cj} + \epsilon_{jc}^C \tag{2.3}$$

Note that the probability of moving between states here is aggregated for all college workers. Estimates of differences in earnings for college workers across states are expressed relative to the earnings of high school education workers. To identify the effect of earnings on migration flows then, I can compare the relative transitions of high school education, and college educated workers to differences in premiums for these two groups. Assuming that equation (2.3) holds true for high school workers, I can estimate the following equation which takes the difference of the log migration flows for college and high school educated workers:

$$ln(P_{jc}^C) - ln(P_{jc}^H) = \pi_0 + \pi_1\hat{w}_{jc} + \pi_2(A_c - A_j) + \pi_3 D_{cj} + \epsilon_{jc} \tag{2.4}$$

where $\pi_0 = \pi_0^C - \pi_0^H$, $\pi_1 = \pi_1^C - \pi_1^H$, $\pi_2 = \pi_2^C - \pi_2^H$, $\pi_3 = \pi_3^C - \pi_3^H$, $\epsilon_{jc} = \epsilon_{jc}^C - \epsilon_{jc}^H$, and $\hat{w}_{jc} = (w_c^C - w_j^C) - (w_c^H - w_j^H)$. Note that I can rewrite the differenced wage as $\hat{w}_{jc} = (w_c^C - w_c^H) - (w_c^H - w_j^C)$ which is the difference in the college premium between states relative to the high school wage. The true value of this term is unknown but I can substitute the PDL

estimate, $\hat{w}_{jc}$, into (2.4). Here, the difference in earnings between states k and j is just the estimate obtained from the estimation approach above. The coefficient estimates on amenity differences and path-specific migration costs represent the relative difference in response to these factors for college and high school educated workers. For instance, if distance between states is a greater driver of migration for high school workers than for college workers (where distance negatively affects migration), then this will be reflected in a positive coefficient estimate on distance.

I use similar proxies for amenity differences as Dahl (2002) and Ransom (2016), using measures for climate amenities, quality of life, and the local state budget. I use the great circle formula to calculate distance between state capitals. More information on these variables is provided in B.1.1. Estimation results are presented in Table 2.18. In general, the results are similar to those of Dahl (2002), although the magnitude of the relationship is lower in this sample. Across all specifications, regardless of the controls used, there is a clear positive relationship between higher mean earnings and migratory flows. If the difference in mean earnings between states is higher for the college educated than those with a high school diploma, then college workers are more likely to follow this migration path. F-tests performed in specifications 4-7 also provide limited evidence of differences in the importance of amenities in influencing the choice to move across education groups.

Another instructive exercise is to compare results obtained using the PDL correction method, and using Dahl's method. In Table 2.19, I obtain results using estimates from the specification in Table 2.6. In general, there is no clear trend in comparing estimates between these tables. Dahl's results are on average slightly higher, but this is not true of all specifications. In summary, this exercise confirms the results in Dahl (2002) and Ransom (2016), that migration flows are consistent with the logic underpinning the Roy model.

2.8 Conclusion

This chapter uses the novel selection correction method introduced in the first chapter of this dissertation to derive new and improved estimates of the returns to education across states. I compare these results to those obtained using the method pioneered by Dahl (2002) and find that the choice of correction method used is of great importance. The estimates obtained using my preferred correction procedure confirm the general direction of the bias in OLS estimates as found in prior studies. However, my results suggest that more traditional and restrictive correction methods tend to overstate the upward bias in OLS estimates of the returns to at least some college education, and understate the bias at either end of the education spectrum.

I have discussed at some length alternative specifications and practical issues related to im-

Table 2.19: Migration Flows and Differences in Returns: Alternative Return Measure

	(1)	(2)	(3)	(4)	(5)
Intercept	0.68***	0.61***	0.62***	0.60***	0.69***
	(0.03)	(0.02)	(0.02)	(0.02)	(0.03)
Δ College Return	.77***	1.07***	1.00***	1.63***	1.78***
	(0.25)	(0.28)	(0.35)	(0.27)	(0.47)
N	1341	1341	1341	1341	1341
R^2	0.0255	0.0737	0.0829	0.0729	0.142
Quality of Life		YES			YES
Climate			YES		YES
State Spending and Taxation				YES	YES
Distance	YES				YES
Δ Unemployment Rate	YES				YES
F-test for Amenity Variables		13.59	16.17	14.71	9.84
		(0.00)	(0.00)	(0.00)	(0.00)

Notes: Robust standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Dependent Variable is the log difference in migration flows between states for college educated and high school educated individuals. Δ operates on differences in premia across states for college and high school educated individuals. The corrected returns use here come from the estimates obtained using Dahl's correction procedure. See B.1.1 for the list of included amenity variables and other control variables.

plementation of this improved correction procedure in practice. Though this methodology does not require restrictions on the number of control functions used in estimation, as this empirical exercise makes clear, a restriction is necessary in some situations, depending on the dimensionality of the choice set, and the size of the data set.

This chapter also makes clear the importance of accurate estimation of the probability terms central to capturing selection bias. Furthermore, I present clear evidence that this estimation becomes even more important when coupled with model selection performed using PDL. If control terms are poorly estimated this can lead to poor selection of key control terms and introduce additional noise into estimation with little gain in terms of bias reduction. Following estimation, I explore the terms selected by PDL in my preferred specification, finding clear patterns in control variable selection which confirm the efficacy of this methodology. Additionally, I present new estimates of national level wage inequality between workers with a high school education and a college degree. I find that OLS estimates are upward biased, but by less than Dahl's methodology would suggest. I mirror the exercise in Dahl (2002) in section 2.7, confirming the core implications of the Roy model of mobility.

Chapter 3

Estimating Union Wage Spillovers: The Role of Bargaining and Emulation Effects

3.1 Introduction

Since Lewis (1963) the literature on union wage effects has hypothesised the potential role of spillover effects in driving fluctuations in nonunion earnings. Wage emulation in the nonunion sector, operating due to the 'threat' of unionisation (Rosen, 1969), should increase wages when union power is strong. The large and prolonged decline in the union sector observed since the late 1970s then could have potentially large effects on wages outside of this sector. Though a substantive number of papers have attempted to estimate the impact of union wage spillovers, these estimates have relied largely on variation in the percent of organized workers. Evidence based on variation in this proxy for union power is mixed and sensitive to varieties of control terms included, with the preponderance of studies finding a small positive spillover effect.

Early contributions in this literature use the percent of workers unionised either at the industrial or city/state level, exploiting cross-sectional variation, to capture the existence of general equilibrium spillovers. Freeman and Medoff (1981) find a non-significant positive correlation between the proportion union and nonunion wages, a result confirmed by Donsimoni (1981). Conversely, Holzer (1982) finds a large positive effect, with a ten percent increase in the proportion unionised associated with a .04 increase in log nonunion wages. Kahn (1980) provides additional evidence in support of a positive spillover effect, while Hirsch and Neufeld (1987) find a positive spillover effect at the industry level, but no significant effect operating in the local labor market. Dickens and Katz (1986) also find a positive correlation between the union proportion and wages, although Podgursky (1986) finds that this effect exists only for sufficiently large establishments.

A severe limitation of these early studies is the existence of omitted variables and the correlation of the proportion union with selection effects. In moving from estimation in the cross-section to cross-temporal estimation, Neumark and Wachter (1995) control for industry and year fixed effects, finding a significant negative relationship between the industrial proportion union and nonunion wages, in direct opposition to previous studies. In an important analysis Farber (2005) exploits more plausibly exogenous variation in union power using right-to-work laws, and in predicted estimates of union membership. Farber also carefully considers the role played by omitted variables, in turn introducing fixed effects at the industry and state level. His analysis finds great sensitivity in estimates to the source of variation used, providing some context for the variation in estimates across earlier studies. When controlling for a wider range of potential omitted variables his results indicate at most a small positive effect of union power on nonunion wages.

In a recent paper Fortin et al. (2019) estimate the role played by union threat effects, first by extending the analysis of Farber (2005) using variation in right-to-work laws, and secondly, using variation in the unionisation rate included as an additional covariate in their distribution regression approach. The results of their event study finds some evidence of a decline in nonunion wages following the introduction of right-to-work laws. Additionally, they find evidence of a positive spillover effect operating primarily at the part of the wage distribution just below the median. Building on the re-weighting procedure of DiNardo et al. (1996) they find that spillovers account for half the increased wage inequality explained by declining unionisation. Fortin et al. (2019) is the most comprehensive examination of union spillover effects that exists but it still faces difficulties in not fully addressing endogeneity issues.

In contrast to this literature that largely relies on reduced form estimation of the union spillover effect, our approach to estimating spillover effects formalises union spillovers in a search and bargaining framework, endogenising the process of union formation, and incorporating wage effects arising through differences in the bargaining process. In making clear what is being identified in the model and the variation used, we overcome the problem inherent in early studies of likely bias due to omitted characteristics and selection into the union sector, and we estimate an effect with a clear theoretical basis and interpretation. In using the percent of workers organised as a proxy for spillover effects it is impossible to determine the channel through which these effects operate.

Importantly, our model introduces a novel mechanism through which the union sector affects nonunion wages: the bargaining channel whereby the ability of nonunion workers to find high paying union jobs affects their bargained wages. Additionally, we incorporate wage emulation due to the union threat where the wages of nonunion workers are affected by variation in the alternative prospects of *union* workers. Our framework makes clear the difficul-

ties inherent in separately identifying these two effects, while controlling for selection into the union/nonunion sectors. The contribution of this paper then is provide new evidence on an novel channel through which union wage spillovers operate, and to separately identify this channel from traditionally hypothesised wage emulation effects. Furthermore, in basing estimation on a formal theoretical framework we are clear about what is being identified in our model, which distinguishes our estimates from much of the existing literature.

We are not the first to embed unionisation within a search and bargaining model. Our baseline model is based on that of Taschereau-Dumouchel (2017) (henceforth TD), whose work is informed by the early contributions of Pissarides (1986), Açıkgöz and Kaymak (2014), and Krusell and Rudanko (2016), amongst others. TD indeed formalises union threat effects in his model, though the channel of the effects deviates from the typical hypothesised wage emulation channel. Instead, threat effects are manifested in firm hiring decisions, where workers of different skill levels are hired to prevent a vote in favour of unionisation. Though this effect is certainly interesting, we believe it is likely of second order importance relative to emulation effects and so our model instead incorporates this channel of effects.

Additionally, our framework is informed by Beaudry et al. (2012) (henceforth BGS), which formalises the impact of changing alternative job prospects (outside options) on wages by means of changes in the industrial composition of work. Following BGS we model local labour markets composed of industries and firms with workers able to transition between jobs in proportion to their prevalence.[1] As in BGS we will exploit variation within industry, across city to identify the effect of declining unionisation, and changes in the composition of union work on nonunion wages over the period 1980-2010.

Though our focus is on the estimation of spillover effects affecting nonunion earnings, our paper is related both directly and indirectly to the large and substantial literature that traces both patterns in declining unionisation, and estimates both the union wage premium and the role of declining unionisation in driving increasing wage inequality. Card et al. (2004) and Card et al. (2018) provide a comprehensive summary of the research in this area following the early contribution of Freeman (1980). Freeman's analysis finds evidence of declining 'within' sector wage inequality which can potentially offset increased 'between' sector wage inequality arising due to the existence of a union wage premium. In an important contribution DiNardo et al. (1996) (henceforth DFL) introduce a semi-parametric re-weighting technique building on the work of Oaxaca (1973) which attributes around 14% of the increase in wage inequality over

[1] Though we specify the firm as our unit of analysis and the level at which workers become unionised, estimation exploits variation at the industry-city level. In that sense, firms can be thought of more accurately as establishments, or more generally, the level at which the unionisation vote takes place. Given the emphasis on the enterprise level in the Wagner Act, aggregation to the city level may ignore some complexity, particularly if there are notable transitions between different union/nonunion establishments in the same firm.

1979-1988 (for men) to declining unionisation.[2] Further studies by Card (2001), Card et al. (2004) and Gosling and Lemieux (2001) provide estimates of the union contribution to wage inequality, comparing patterns across male/female workers, private/public sector workers, and across countries. In a recent contribution, Card et al. (2018) compare the experience of Canada and the United States, focusing in particular in the increasing prominence of the public sector as a source of unionised employment.[3]

We begin our empirical analysis by estimating a simple version of our model which incorporates spillover effects operating through the bargaining channel, formalises selectivity, and makes clear barriers to identification. Specifically, outside options associated with the union sector may be correlated with unobserved local productivity shocks. As in BGS and Beaudry et al. (2014) we overcome this problem using Bartik-style instruments which exploit exogenous variation coming from national level changes in wage premia and industrial employment growth. To deal with selection into the union sector we use measures related to the number and success of union certification drives to proxy for local variation in the union threat.

The results from our specification indicate that average log wages would be 0.02 log points higher in 2010 had outside options associated with the union sector been fixed for nonunion workers over 1980-2010. In estimating results separately for sub-populations we find larger spillover effects operating over this period for the young and for the least educated workers. For men with at most a high school education nonunion wages declined .026 log points as a result of declining outside options. Conversely, changing outside options led to a mild increase in the wages of more highly educated males. Our decomposition of earnings also indicates interesting interplay between the various components of outside options, with an improvement in the composition of union jobs for highly educated workers largely offsetting impacts of declining union job availability. In an extension to the baseline model we allow for firms to respond to the union threat through emulating wages paid in the union sector. Our core wage equation is then extended to include an additional term which reflects the outside options of workers in the union sector. Additionally, the average industrial wage will be a weighted average of non-emulation and emulation wages across firms in the sector.

The remainder of the paper is organised as follows: In section 3.2 we present our baseline model, in section 3.3 we present a discussion of identification and our results based on the framework in section 3.2. In section 3.4 we extend our model to incorporate the firm response to the union threat and we conclude in section 3.5.

[2] A similar approach is used by DiNardo and Lemieux (1997). The study by Fortin et al. (2019) already mentioned updates the results of DFL to 2017, while extending the approach to account for spillover effects.

[3] See also recent studies by Farber et al. (2018) and Firpo et al. (2018)

3.2 The Model

3.2.1 Model Set-up

Our model is based on that of Taschereau-Dumouchel (2017) (TD) which places union formation and wage setting in a search and bargaining model, with unions being able to bargain a higher wage because they can threaten to take the whole workforce out of production while an individual, nonunion worker can only threaten to withdraw her own labour. In the TD model, firms employ workers of different skill levels who have different preferences about unions. In particular, since unions compress skill differentials, more skilled workers would vote against a union and less skilled workers would vote in favour. Firms facing a union threat can resist unionization by taking advantage of these preferences: hiring a larger proportion of skilled workers in order to construct a workforce that will vote against a union. Our interest is in the potential spill-over effects of unions on nonunion wages and, partly for that reason, we focus instead on the concept of nonunion firms resisting unionization through paying higher wages. We also view this type of wage emulation channel as well as direct legal and illegal campaigns by firms as likely a more immediate response to unionization threats than altering the skill composition of the workforce. In addition, we will allow for the possibility of both direct and wage emulation effects of unions on wages in a given industry affecting wage setting in other industries in a local economy. Following Beaudry et al. (2012), this can happen because having higher rent job options in a city increases the value of the outside option for workers in all industries as they bargain with their employers. Through the rest of the paper, we will refer to effects of unions on nonunion wage setting in order to resist unionisation as emulation effects and indirect effects through impacts on bargaining options as bargaining spillover effects. To focus attention on whether these channels are sizeable, we will alter Taschereau-Dumouchel's model by having only one skill level but multiple industries.

With that in mind, we start with an aggregate production function of the form:

$$Q = \left[\sum_i a_i Z_i^{\chi} \right]^{\frac{1}{\chi}}, \chi < 1 \tag{3.1}$$

where, Q is total output for the national economy and is equal to a CES aggregate of I intermediate goods indexed by i. The intermediate goods (Z_i's) are produced in local economies, or cities, of which there are C. Thus, $Z_i = \sum_c Y_{ic}$, where Y_{ic} is the amount of intermediate good i produced in city c. The price for Q is normalized to 1 and the price for intermediate good i is given by p_i. For simplicity, and to focus attention on wage setting, we will ignore firm entry and exit and assume that there is a fixed number of firms operating in each industry by city cell.

Thus, $Y_{ic} = \sum_c y_{icf}$, where y_{icf} is the output of firm f operating in industry i in city c.

Firms and workers operate in a labour market that includes frictions. Unemployed workers and vacancies posted by firms meet according to a matching function, $M(U_c, V_c)$, where U_c is the number of unemployed workers in city c and V_c is the number of unfilled vacancies. As is standard in the search literature, we will assume that the matching function is constant returns to scale. BGS show that in steady state this implies that the probability that a vacancy in city c meets an unemployed worker, q_{vc}, and the probability that an unemployed worker in city c meets a vacancy, q_{uc}, can be written as functions of the employment rate for the city, ER_c. We will assume that workers are not mobile across cities and that there is no on-the-job search.[4] Our model is partial equilibrium in that we take treat q_{vc}, q_{uc}, and ER_c as given rather than solving for them from the model. We make this simplifying assumption in order to maintain our focus on wage outcomes.

Once workers and firms meet, they bargain a wage to divide the match surplus. Following TD, in a nonunion firm, this bargaining is between the individual worker and the firm while in a union firm it is between the set of employees and the firm. Employees at nonunion firms have the opportunity each period to vote on whether to form a union. Once unionized, a firm stays unionized - there is no decertification. Both union and nonunion firms choose the optimal number of workers to hire taking account of the bargained wage among other factors.

3.2.2 Workers

We begin by characterizing the choices and environment faced by workers. We will work in discrete time and assume that all workers have the same skill level. At any moment, a worker can be unemployed and searching for work or employed at one of three types of firms in a particular industry. The first type of firm is a simple nonunion firm in which the firm and worker bargain a wage and the firm chooses an optimal number of employees without direct regard to the threat of unionization. As we will see, these firms will arise in situations where the costs and benefits of unionization are such that workers at the firm do not want to form a union. Workers in these firms have a value function given by,

$$W^n_{icf}(w^n_{icf}) = w^n_{icf} + \rho(\delta U^u_{ic} + (1-\delta) W^n_{icf}(w')) \tag{3.2}$$

where, ρ is the discount rate, w^n_{icf} is the nonunion wage that would be paid at firm f in the given $i-c$ cell, δ is an exogenous probability that the worker's job is terminated, and U^n_{ic} is the value of unemployment in city c for a worker whose previous job was a nonunion job in industry i. In this specification, w' corresponds to the wage that will be paid in the next period

[4] Beaudry et al. (2014) provide a model with mobility.

if the job is not terminated. Following TD, we assume that workers and firms believe that next period's wage will be set optimally and that they cannot affect it through actions this period. This assumption rules out, for example, the relevance or ability to affect reputations.

Alternatively, workers in the firm may vote to unionise, in which case they have a value function given by,

$$W^u_{icf}(w^u_{icf}) = w^u_{icf} + \psi_{icf} + \rho(\delta U^u_{ic} + (1-\delta)W^u_{icf}(w')) \tag{3.3}$$

where u superscripts correspond to unionised values, and ψ_{icf} is the non-wage benefit to workers from being in a union in this particular firm. The value ψ_{icf} could be related to the work environment and/or personal non-productive attributes of other workers in the firm. We assume that a value for ψ_{icf} is drawn separately for each firm from a common distribution, $f(\psi)$ and that the union amenities are created by the union itself rather than the firm and do not enter the cost function for the firm.

Firms can respond to a union threat by paying workers a wage that is high enough to make them indifferent between unionising and remaining non-union. Given this, there is a third possible type of firm which we will call a union emulating firm in which workers are paid a wage, $w^{n*}_{icf} > w^n_{icf}$ but do not get union benefits. As a result, the value function associated with working at this type of firm is the same as for the simple non-union firm but with a wage of w^{n*}_{icf}.

The other potential state for workers is being an unemployed searcher. For reasons that will become apparent momentarily, the value of unemployment differs by previous industry and by whether the person was a union or nonunion worker in their previous job. The value function for a person who was in a union job in industry j is given by:

$$\begin{aligned} U^u_{jc} = b + \rho\Bigg(q_{uc}\Bigg[T^u_{jc}\sum_i \eta^u_{ic} W^u_{ic} + (1-T^u_{jc})\sum_i \eta^n_{ic}\left(p^n_{ic} W^n_{ic} + (1-p^n_{ic})W^{n*}_{ic}\right)\Bigg] \\ + (1-q_{uc})U^u_{jc}\Bigg)\end{aligned} \tag{3.4}$$

where: b is the flow value of being unemployed; η^u_{ic} and η^n_{ic} are the proportion of employment in the union and nonunion sectors, respectively, in city c that is in industry i. For workers in city c and previously employed in a union job u in industry j, the probability of finding a job in the union sector is T^u_{jc}. With probability, $(1-T^u_{jc})$ workers match with a firm in the nonunion sector. We allow the probability of finding a union job to differ according to whether

the worker was previously in a union job and by their industry of previous employment.[5] This is the reason for indexing the value of unemployment by previous industry and union status.

We assume that search is random - that, for example, workers are unable to differentiate between nonunion firms paying regular nonunion wages and nonunion firms paying emulation wages.[6] As a result, they find employment in proportion to the relative abundance of these firms as measured by p_{icn} which is the fraction of non-emulation paying jobs in the nonunion sector in industry i in city c. $W^u_{ic} = E(W^u_{icf})$ is the expected value of employment in a union firm in the $i-c$ cell with the expectation taken across the distribution of union firms in that cell (the other expectations are defined analogously). Thus, an unemployed worker gets a flow value, b, and with probability q_{uc} meets a vacancy. Workers have different propensities for finding employment in the union and nonunion sectors, but otherwise find jobs in proportion to their share of local employment. This expression implies that the value of search is higher when the local employment structure has more high value industries and more unionized firms since that means searchers are more likely to find those high value jobs. To simplify other expressions in the model we will often summarize the expected value of a job as:

$$\tilde{W}^u_{jc} = T^u_{jc} \sum_i \eta^u_{ic} W^u_{ic} + (1 - T^u_{jc}) \sum_i \eta^n_{ic} (p^n_{ic} W^n_{ic} + (1 - p^n_{ic}) W^{n*}_{ic}) \tag{3.5}$$

In the empirical section of the paper we will consider alternative formulations of these outside option terms which are defined separately for nonunion and union workers, and by industry.

3.2.3 Firms

All firms in a given industry-city cell have a common production function given by,

$$y_{icf}(n) = \epsilon_{ic} n - \frac{1}{2} \sigma_i n^2 \tag{3.6}$$

where, ϵ_{ic} is a local productivity shock, n is the number of employees, and $\sigma_i > 0$ is a parameter reflecting potential span of control issues. This specification implies that technology is common across cities within an industry but that there is comparative advantage in producing an intermediate good by city. We assume that the technology is common to all three types of firms (unionized, non-union, and union emulators). The literature on union effects on productivity

[5] For nonunion job searchers the probability is T^n_{jc}.

[6] Our model therefore abstracts away from issues related to workers queueing for union jobs (see Abowd and Farber (1982) for a theoretical treatment of queuing and supportive empirical evidence). This queuing mechanism could imply an additional spillover channel whereby the existence of union firms drives down vacancy filling rates in the nonunion sector, pushing up wages. The prevalence of queueing is likely driven by union wage premium and the relative likelihood of finding union work such that queuing effects are likely to enter through the outside option channel.

seems to us to be inconclusive and so we adopt an agnostic take in which unions affect firm activity by affecting wages but not through technological adaptations.[7] We assume that the σ_i's are sufficiently smaller than 1 such that, combined with the assumption of a fixed number of firms in each ic cell, they imply that production of any good is spread across cities.[8]

Firms choose a number of vacancies to post in order to attain a profit maximizing number of employees given the bargained wage. The cost of hiring is linear in the number of vacancies posted. As a result, the value function for a nonunion firm is given by,

$$J^n_{icf}(n) = \max_v p_i y_{icf}(n') - w^n_{icf}(n')n' - \varkappa v + \rho J^n_{icf}(n') \tag{3.7}$$

subject to the equation of motion:

$$n' = n(1-\delta) + q_{vc}v \tag{3.8}$$

where $J^n_{icf}(n)$ is the value of non-union firm f when it ends the previous period with n employees. Of those, a fraction, δ, leave the firm for exogenous reasons. The firm posts v vacancies at a cost, $\varkappa$, per vacancy and fills them with probability, q_{vc}. Following TD, we will assume that the firm can only post positive vacancies and that the number of vacancies is sufficiently large that we can treat $q_{vc}v$ as a deterministic number according to a law of large numbers. This allows us to rewrite the value function as:

$$J^n_{icf}(n) = \max_{n'} p_i y_{icf}(n') - w^n_{icf}(n')n' - \varkappa \frac{n' - n(1-\delta)}{q_{vc}} + \rho J^n_{icf}(n') \tag{3.9}$$

Note that we have written the value function as if the firm exists in a stationary environment in which it assigns a probability of zero to its workers trying to unionise in the future. We will return to that assumption below. The value functions for firms when they are unionized or acting as union emulators are identical in structure to the nonunion firm value function, with the only change being the substitution of the relevant wage ($w^{n*}_{icf}(n')$ or $w^u_{icf}(n')$ in the expression. We next turn to using these value functions in combination with those for the workers to characterize wage bargaining solutions. Once we have those solutions, we will be in a position to discuss union formation.

[7] Hirsch and Link (1984) and Addison and Hirsch (1989) summarise the early research in this area which finds largely inconclusive and mixed evidence on the effect of unionism on productivity.

[8] We work with a quadratic production function to permit tractability in deriving our wage expressions. It is worth noting that TD showed that using a Cobb-Douglas type form for production has the unfortunate implication that general productivity shifts such as ϵ_{ic} do not determine wages. In that sense, our results are not perfectly generalizeable but a quadratic function captures the main points we want to emphasize.

3.2.4 Wage Determination

Collective Bargaining

Once the firm hires workers, the workforce will vote on unionization. If a union is formed then the union bargains collectively on behalf of workers. Wages are bargained according to Nash bargaining over the entire surplus to production from hiring n workers. By bargaining collectively the union is able to effectively threaten that the entire workforce will quit and enter the unemployment pool this period. Such an action would impose two costs on firms: first, the firm would produce zero units of output this period; second, as the firm hires n workers every period in steady state, the number of vacancies required to achieve this optimal workforce will be much larger, and hence vacancy filling costs borne by the firm will increase. These costs are taken into account when determining the firm's and worker's surplus.

Wages are set according to the Nash Bargaining condition

$$\beta S^u = (1-\beta) n (W^u_{icf}(w) - U^u_{ic}) \tag{3.10}$$

where S^u represents the firm's surplus. On the right hand side is the total sum of workers surplus, which is given by the gain to employment for all workers hired by the firm. Since the workers are identical, we use a specification that focuses on the total surplus and assume that the union members will all get an equal share of the part of the surplus captured by the union. This ignores issues related to seniority, for example.[9]

We will focus on a steady state in which the wage and optimal number of workers for a firm are constant across periods. Given this, the workers' surplus can be re-written as

$$\begin{aligned} W^u_{icf}(w) - U^u_{ic} = & \frac{1}{1-\rho(1-\delta)}(w^u_{icf}(n^u_{icf}) + \psi_{icf}) - \frac{(\rho-1)b}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} \\ & - \frac{(\rho-1)\rho q_{uc}}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} \tilde{W}^u_{ic} \end{aligned} \tag{3.11}$$

On the firm side, the surplus from a successful bargain with a union is given by the difference between producing this period with n_u workers (the optimal number of workers with a bargained union wage) and not producing this period along with the cost of rehiring the entire workforce the next period. The firm surplus can be expressed as follows (with a detailed derivation provided in the appendix):

[9] See Abraham and Medoff (1984, 1985) who present evidence of the importance of seniority for layoffs and promotions, and see Abraham and Farber (1988) for evidence that the seniority wage profile is steeper under collective bargaining.

$$S^u = \left(p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf}(n^u_{icf})\right) + \rho(1-\delta)\frac{\varkappa n^u_{icf}}{q_{vc}} \tag{3.12}$$

The surplus equals the current period profit plus the cost of replacing the portion of the work force that would not normally turn over, which is the relative cost of being in the bargaining break-down option.

Solving the bargaining expression for steady state wages yields:

$$\begin{aligned} w^u_{icf}(n^u_{icf}) = & \frac{\beta(1-\rho(1-\delta))}{(1-\beta\rho(1-\delta))}\frac{p_i y_{icf}(n^u_{icf})}{n^u_{icf}} + \frac{\beta\rho(1-\delta)(1-\rho(1-\delta))}{(1-\beta\rho(1-\delta))}\frac{\varkappa}{q_{vc}} \\ & + \frac{(1-\beta)(1-\rho)}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))}b - \frac{(1-\beta)}{1-\beta\rho(1-\delta)}\psi_{icf} \\ & + \frac{(1-\rho)\rho(1-\beta)q_{uc}}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc})}\tilde{W}^u_{ic} \end{aligned} \tag{3.13}$$

Individual Bargaining

Non-union wages are also set through Nash bargaining. However, the value for the firm of the option corresponding to a break down in negotiations relates only to the loss of an individual worker (and the indirect effects the removal of one worker has on the others) rather than to the complete stoppage of production that occurs under collective bargaining. The Nash bargaining condition is the same as before, and we derive the following expression for the firm surplus (see appendix for details):

$$S^n = p_i \frac{\partial y_{icf}(n^n_{icf})}{\partial n} - w^n_{icf}(n^n_{icf}) - n^n_{icf}\frac{\partial w^n_{icf}(n^n_{icf})}{\partial n} + \frac{\rho(1-\delta)\varkappa}{q_{vc}} \tag{3.14}$$

Substituting this expression into the Nash bargaining condition and solving the differential equation in wages yields the following expression for non-union wages:

$$\begin{aligned} w^n_{icf}(n^n_{icf}) = & \frac{1-\rho(1-\delta)}{1-\beta\rho(1-\delta)}\frac{\beta p_i}{1+\beta}\left(\frac{\partial y_{icf}(n^n_{icf})}{\partial n} + \beta\epsilon_{ic}\right) + \frac{\beta\rho(1-\delta)(1-\rho(1-\delta))}{(1-\beta\rho(1-\delta))}\frac{\varkappa}{q_{vc}} \\ & + \frac{(1-\beta)(1-\rho)}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))}b + \frac{(1-\rho)\rho(1-\beta)q_{uc}}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))}\tilde{W}^n_{ic} \end{aligned} \tag{3.15}$$

This wage differs from the union wage in two ways. First, under collective bargaining the union negotiates over the surplus generated from total output while for non-union workers

what matters is the marginal surplus. As a result, individual wages are determined by average output for union workers but marginal product for nonunion workers. Second, union wages include a compensating differential component with wages declining when union amenities are greater. On the other hand, outside options - which are determined by both the value of non-work time and the expected value of finding a new job after unemployment - have the same effects on the union and nonunion wages.

Firm Size

For several reasons, it will prove useful to derive expressions for the optimal number of workers hired by firms that are either unionized or that are non-union and pay the nonunion wage derived in the previous section.[10] These are straightforward exercises in which we set the derivatives of the firm value functions with respect to number of workers equal to zero.[11] The expressions for the optimal firm size in union and nonunion firms are given by:

$$n^u_{icf} = \frac{1}{\sigma_i p_i}\Bigg[p_i\epsilon_{ic} + \psi_{icf} - \frac{(1-\beta\rho^2(1-\delta)^2)}{1-\beta}\frac{\varkappa}{q_{vc}} - \frac{1-\rho}{1-\rho(1-q_{uc})}b - \frac{(1-\rho)\rho q_{uc}}{1-\rho(1-q_{uc})}\tilde{W}^u_{ic}\Bigg] \tag{3.16}$$

and,

$$n^n_{icf} = \frac{1+\beta}{(1+\beta\rho(1-\delta))}\cdot\frac{1}{\sigma_i p_i}\Bigg[p_i\epsilon_{ic} - \frac{(1-\beta\rho^2(1-\delta)^2)}{1-\beta}\frac{\varkappa}{q_{vc}} - \frac{1-\rho}{1-\rho(1-q_{uc})}b - \frac{(1-\rho)\rho q_{uc}}{1-\rho(1-q_{uc})}\tilde{W}^n_{ic}\Bigg] \tag{3.17}$$

The part of this expression in parentheses is the same as the union firm size expression with one exception: the union expression includes the value of amenities the union provides, ψ_{icf}. Because these amenities imply lower wages in steady state, higher amenities mean a larger firm

[10] In particular, though existing studies have established the importance played by workplace or establishment size (Podgursky (1986), Pearce (1990)), we are limited in using the CPS to measures of the number of employees at the firm level, which is a poor proxy for workplace size. Note that though we have used 'firms' throughout this paper as our terminology of choice, one can think of union certification at the level of the workplace or establishment without making any modification to the model. In solving for the size of the firm (or establishment) in this framework we are not ignoring the effect of firm size on union formation and wages, instead the manner in which size affects wages in the model will be embedded in the coefficient estimates on our other key variables. Firm size in this framework is determined through firm optimisation which is driven by productivity, model parameters, and variables affecting wage setting.

[11] See the appendix for derivation details.

size. Otherwise, the same terms in each are being multiplied by 1 in the union equation and by a term greater than 1 in value for the nonunion equation. That force on its own means that nonunion firms are larger. The amenities effect works in the opposite direction.

Wage Equations

In our empirical work, we will focus on log linearizations of wages around the point where all cities have the same proportions in each industry.[12] One key point of this exercise is to bring out the employment rate in the city, ER_c. Noting that the expected value of a job is a function of wages ($E^n_{ic}(w)$), we express outside options as a weighted average over wages.

In steady state, the worker job finding rate, q_{uc} and the firm worker finding rate, q_{vc} can be written as functions of ER_c. Thus, we get:

$$w^n_{icf} = \gamma^n_{0i} + \gamma^n_1 E^n_{ic} + \gamma^n_2 ER_c + \gamma^n_4 \epsilon_{ic} \tag{3.18}$$

and,

$$w^u_{icf} = \gamma^u_{0i} + \gamma^u_1 E^u_{ic} + \gamma^u_2 ER_c - \gamma^u_3 \psi_{icf} + \gamma^u_4 \epsilon_{ic} \tag{3.19}$$

where

$$E^u_{jc} = T^u_{jc} \sum_i \eta^u_{ic} w^u_{ic} + (1 - T^u_{jc}) \sum_i \eta^n_{ic} (p^n_{ic} w^n_{ic} + (1 - p^n_{ic}) w^{n*}_{ic}) \tag{3.20}$$

and,

$$E^n_{jc} = T^n_{jc} \sum_i \eta^u_{ic} w^u_{ic} + (1 - T^n_{jc}) \sum_i \eta^n_{ic} (p^n_{ic} w^n_{ic} + (1 - p^n_{ic}) w^{n*}_{ic}) \tag{3.21}$$

where, $p^n_{ic} w^n_{ic} + (1 - p^n_{ic}) w^{n*}_{ic}$ is the observed mean nonunion wage. Note that the union and nonunion wage equations have different error terms, with the nonunion error term consisting of the productivity shock, ϵ_{ic} while the union error term includes the productivity shock but also the unobserved (to the econometrician) value of union amenities, ψ_{icf}. Given that we have assumed, so far, that workers are identical, higher wages in an industry correspond to rents - differences in pay over and above what is required for the marginal worker to want to join that industry. Those rents are maintained because of the frictions in the labour market. It is important that we are considering rents since wage differences across industries that correspond to compensating differentials (say, because of having to work with asbestos) cannot be the basis of bargaining a higher wage with your current employer. If a higher wage corresponds to a compensating differential then there is no net gain to moving to the dangerous job and, so, no

[12] See appendix for details.

basis on which to threaten your current employer during bargaining.

3.2.5 Union Determination

Our theoretical analysis occurs at the level of the firm. At that level, we are interested in the question of whether firms become unionised. The other element in determining the overall unionisation rate in a local economy is decertification. However, decertification is much less common than certification.[13] To reflect this, and keep our model as simple as possible, we will assume that there is no decertification. Instead, all firms - union and nonunion - die with a probability, δ_d, in a period. The dying firms are replaced with new firms started by new entrepreneurs. In steady state, the number of dying firms equals the number of newly born firms. All firms are born nonunion and then workers at the firm decide whether to unionise. Recall that firms are born with a draw of a level of amenities that would be provided at that firm if it were unionised, ψ_{icf}.

We begin with a simple model in which firms are passive players in unionisation. That is, unionisation is determined entirely by the workers and firms do not try to respond by, for example, emulating union wages. In terms of the model derived so far, this implies that the probability of meeting an emulating firm $(1-p^n_{ic})$ is zero and the outside option wage terms are adjusted accordingly.

In the simple model, the workers at a firm compare the value of the job continuing as a nonunion job to the value of the job being union minus the cost of unionising. The wages they use in this exercise are the ones we arrived at in the previous section that reflect the optimal hiring decisions by firms, making this a monopoly union model. We will assume that the decision is made according to a median voter model with the median voter not at risk of losing her job when employment is reduced after unionisation. We also assume that workers do not care about the employment outcomes of those who do lose their jobs, implying that we can focus exclusively on wages. Given this, a firm becomes unionised if:

$$W^u_{icf}(w^u_{icf}) > W^n_{icf}(w^n_{icf}) - \lambda^*_{ct} \tag{3.22}$$

where, λ^*_{ct} is the fixed cost to workers of unionising a firm in a city c. Substituting in steady state expressions for the value functions based on (3.3) and (3.2), we can define an index function:

$$I_{icf} = (w^u_{icf} - w^n_{icf}) + \psi_{icf} - (1-\rho(1-\delta))\lambda^*_{ct} \tag{3.23}$$

[13] Using election data from the NLRB (discussed below) we find that certification elections outnumber decertification elections over 1980-2010 by at least 5 to 1. The same is true if we consider the number of workers involved in elections. See Fortin et al. (2019) who present a figure showing the ratio of eligible workers certified over the time horizon 1978-2017.

such that a firm is unionised if $I_{icf} > 0$ and remains nonunion otherwise. Here, we have assumed that all workers in the firm share the costs of unionisation equally. The term multiplying the fixed cost per worker puts the one-time fixed cost of unionising in the same present value terms as the flow of wages and union amenities.

We can substitute in the union and nonunion log-linearized wage expressions, (3.19) and (3.18) into (3.23) to give:

$$I_{icf} = \alpha_{0i} + \gamma_1^u E_{ic}^u - \gamma_1^n E_{ic}^n + \alpha_2 ER_c + (1-\gamma_3^u)\psi_{icf} - \lambda_{ct} + \alpha_4 \epsilon_{ic} \tag{3.24}$$

where, $\lambda_{ct} = (1-\rho(1-\delta))\lambda_{ct}^*$ and $\alpha_4 \geq 0$.[14]

This is a very standard selection set-up and implies,

$$E(w_{icf}^n | I_{icf} \leq 0) = \gamma_{0i}^n + \gamma_1^n E_{ic}^n + \gamma_2^n ER_c + \gamma_4^n E(\epsilon_{ic} | I_{icf} \leq 0) \tag{3.25}$$

Notice that the error mean term is a function of λ_{ct} - the cost of unionising - but the unconditional nonunion wage equation is not. Thus, measures of the cost of unionising are available as exclusion restrictions that identify selection effects from direct determinants of the nonunion wage. If we consider two cities, c and c', that are identical except that c' has higher costs of unionisation then unionisation will be lower in c'. Moreover, because the coefficient on the productivity term, ϵ_{ic} is positive in the index function and recalling that ψ_{icf} and ϵ_{ic} are assumed to be independent, union firms tend to have higher productivity. Thus, the marginal firms that would be unionised in c but non-union in c' will be at the low end of the productivity range for union firms but the high end for nonunion firms. This has implications for a simple specification using the proportion union to capture spillover effects, as the estimated coefficient on the union proportion would be biased downward. Industry-city cells with higher unionisation rates would be ones with lower productivity among nonunion firms.

In what follows, we first estimate the regression specification given by (3.25). This allows us to examine the effects of changes in union power on nonunion wages, that is, to investigate the presence and size of spillovers. We do not attempt to estimate a similar specification for local union wages because low unionisation rates, especially in the later years, imply sample sizes in industry x city cells that are too small to work with.

The Impact of Declining Unionisation

We can use equation 3.25 to discuss the channels through which declines in union power affect the mean observed non-union wages. The first channel is through a decline in the probability

[14] This condition on α_4 is shown in the appendix.

that a worker formerly on a nonunion job in industry j finds a union job. This is obviously related to the decline in the proportion of workers who are unionised, though one could imagine it declining either faster than that proportion (if older union workers keep their jobs but new job searchers have difficulty getting into a union job) or slower than that proportion (if the proportion declines quickly because union workers suddenly start taking early retirement). In our specification, this first channel shows up in two places. The first is T_{jc}^{n} (the probability a nonunion person finds a union job of any kind) and the second is changes in the distribution of union jobs across industries (the η_{ic}^{u}'s) relative to changes in the industrial distribution of nonunion jobs. To understand the industrial distribution effect, consider a simple example with a high wage industry (manufacturing) and a low wage industry (services). A decline in the probability, T_{jc}^{n}, refers to a common decline in the probability of getting a union job in either sector, but the impact on worker outside options will obviously be greater if it is mostly union jobs in the manufacturing industry that are lost. At the same time, we don't want to assign all industrial changes as union effects. Instead, we could argue that shifts in the industrial distribution for nonunion workers capture changes in the overall economy while a change in the industrial distribution for union workers relative to what happens for nonunion workers is a union decline effect - a way of capturing a more nuanced version of how unions have declined and the effect that has had on nonunion wages.

The second channel through which declining union power operates is through declines in union wages - if unions become less effective at unifying worker resistance during bargaining, or afraid to threaten the withdrawal of the whole workforce in a new policy environment then the union wage premium will decline. In that case, the value of the outside option of finding a union job for a current nonunion worker also declines. The third channel is through changes in union emulation - either through declining probabilities that nonunion firms feel that they have to emulate union firms (i.e., declines in $(1 - p_{ic}^{n})$) or declines in the wages they have to pay to emulate union firms (i.e., relative declines in w_{ic}^{n*}). The fourth channel is through selection, as shifting firms from being union to nonunion changes the productivity composition of nonunion firms. In our model, this effect implies an increase in the observed nonunion wage, offsetting the negative effects of declining union power operating through the first two channels.

It is useful, at this point, to compare the regression specification that has emerged from the simple version of our model with the standard approach that has been used previously to try to capture spillover effects of unionisation on nonunion wages. In papers including Freeman and Medoff (1981), Holzer (1982) and Hirsch and Neufeld (1987), the main regression takes the

form:[15]

$$w^n_{icf} = a_o + a_1 P_{ic} + a_2 x_{icf} + u_{icf} \tag{3.26}$$

where, P_{ic} is the proportion of workers in the i-c cell who are unionised[16], x_{icf} is a vector of other controls, a_2 is a parameter vector of the same length as x_{icf}, and u_{icf} is an error term. None of the references of which we are aware provides a derivation from theory for this specification and so it is likely viewed as a reduced form specification. In earlier papers the x_{icf} vector typically includes local demand and supply shifters, such as the proportion of teenagers in the region, the unemployment rate, local lagged per capita income growth, average firm size, aggregated industry dummies, amongst others. Additionally, more recent papers by Neumark and Wachter (1995) and Farber (2005) control for full industry-year and/or city-year fixed effects in order to better account for relevant omitted variables.

Comparing the specification given by equation 3.26 with our specification (given in equation 3.25), there are strong similarities. In particular, previous studies have included controls similar to the employment rate as well as industry-year effects that are in equation 3.25. The main difference is that union effects show up through their impacts on outside options in our specification but are represented by the simple proportion union variable in equation 3.26. Changes in our outside option variable will be correlated with changes in the proportion union because, as described earlier, declines in the probability of finding a union job translate into declining outside options for nonunion workers. But the second, union wage change, channel will be missed (except to the extent it happens to be correlated with the change in the union proportion). Deriving our estimating regression from theory both tells us what to control for (the local employment rate and industry x time effects) and implies a constructed outside option variable with specific representations of the way unionisation affects nonunion wages. The specification does not contain the union proportion on its own. We will proceed in stages: first showing estimates based on the standard specification then showing what happens to the coefficient on the union proportion when we introduce our rent variable.

Implementations of the standard specification also have not addressed the selection issue as the composition of nonunion firms changes. This is problematic because in the presence of selection of this type, one cannot find an instrument for the union proportion (such as, say, variables related to the cost of unionisation) that is uncorrelated with the error mean term. That is, one cannot untangle the dual effects of a rise in unionisation costs: that it raises the probability of unionisation independently of shocks to productivity (which is what we would hope the instrument would do); and that it implies a change in the composition of nonunion firms.

[15] Or is pooled for union and nonunion workers, with the proportion union interacted with a union dummy.
[16] Typically this metric is calculated either at the national-industry level, or at the city/state level.

3.3 Estimation

In this section we present details and results regarding estimation of the nonunion wage equation 3.25 derived from the model presented in section 2.

3.3.1 Data

We are interested in comparisons across steady states over a medium-long time horizon and as such we consider variation over 10 year periods. For each time period we pool observations across 3 years to reduce statistical noise affecting our wages estimates across industry-city cells. We consider variation across 1980, 1990, 2000 and 2010, using the years 1978-1980, 1988-1990, 1998-2000, and 2008-2010. In our analysis we use data from the Current Population Survey Merged Outgoing Rotation Groups downloaded from the National Bureau of Economic Research (NBER) for 1990-2000. Our data for the years 1988-1990 comes from the CPS May extracts, also downloaded from the NBER. The May extracts importantly record answers to questions regarding union membership/coverage, although due to limitations in the coverage question we define union workers as those who belong to a labor union.

From this data we extract all workers between the ages of 25-65 who do not report being in school either full-time or part-time. We construct potential experience as max(min(age-years of schooling-6, age-16),0), dropping those with negative potential experience. One limitation of the education data in the MORG's is that prior to 1992 education was reported as the number of years completed, but in later years as the highest grade completed. To deal with this issue we convert categories of completed schooling to completed years post-1991 using Park (1994), and years of schooling pre-1992 to education categories based on Jaeger and Page (1996). We define union workers as workers reporting belonging to a labour union.

We follow Lemieux (2006) closely in the construction of our wage data, working with hourly wages. Specifically, wages are based on individuals reporting employment in the reference week as wage and salary workers. We set allocated wages to missing and for workers paid hourly we use hourly earnings multiplied by usual weekly hours worked. For workers not paid hourly, we use edited weekly earnings, multiplying the weekly earnings topcode by 1.4 for topcoded observations. Wages are converted to 2000 dollars using a CPI deflator. We set to missing weekly earnings associated with an hourly wage below 1 and greater than 100 in 1979 dollars. All calculations use the earnings weights provided in the data.

The industry definition we use to segment the data into industry-city cells is an aggregated grouping of industry codes based on the 1980 industrial classification used by the Census Bureau. As industrial codes are not consistent across years we must crosswalk the 1970, 1990, and 2000 industry codes to the 1980 classification. We do this using crosswalks provided by IPUMS

and the Census Bureau.[17] The aggregate industry definition used in this paper consists of 51 industries. Table C.1 shows the relationship between this detailed industry definition and the 1990 industrial classification system used by the Census Bureau.

We are limited in the construction of our city definition by data availability in the CPS and by the changing definitions used to define metropolitan areas over time. In the years for which we are using the May extracts, 44 metropolitan areas (SMSAs) are available. As a result we are limited to working with these 44 areas. The counties included in these metropolitan area definitions are not stable over time. To deal with this issue we create the most consistent definition possible for each of these 44 SMSAs given data limitations. Where possible we make use of the limited number of counties identified in the CPS.[18] Our city definition is reasonably consistent over time, though, despite our best efforts, additional, relatively less populous counties will be added to the definition over time for some cities.[19] In moving from Metropolitan Statistical Areas (MSAs) to Core-Based Statistical Areas (CBSAs) we are unable to separately identify Dallas and Forth-Worth and so our definition is made up of 43 cities. The final geographic definition we use pools data for these 43 cities and the remaining population. Specifically, we create additional regions made up of the remaining state population absent the population living in these 43 cities. In the end, our core geographic measure is composed of 93 areas that are fairly consistently defined over the course of the sample period.

Additionally, we use data on union elections to proxy for the costs of unionisation, λ_{ct} in our model. The idea is that locations where the proportion of union certification elections that result in a certification are more union friendly. To obtain these proportions we use National Labor Relations Review Board (NLRB) case data for the three year periods for which we use CPS data.[20] We focus on certification elections, and cases where a conclusive decision on certification was reached.[21] We use the county of the unit involved in the election to construct our geographic measures, aggregating counties to our definition of cities discussed above.

3.3.2 The Outside Option Term

Central to our empirical work are the outside option terms characterising alternative job prospects in either the union or nonunion sectors. Through the bargaining process, wages will increase if workers' other employment opportunities improve. As defined above, these terms are com-

[17] Available at `https://www.census.gov/topics/employment/industry-occupation/guidance/code-lists.html` and `https://usa.ipums.org/usa/volii/occ_ind.shtml`

[18] IPUMS CPS provide a list of these codes: `https://cps.ipums.org/cps/codes/county_19952004_codes.shtml`

[19] The metropolitan area definition used by the IPUMS identifies this general pattern of expanding metropolitan area definitions over time: `https://usa.ipums.org/usa/volii/county_comp2b.shtml`

[20] Our thanks to Hank Farber for providing this data.

[21] As opposed to the case being dismissed or withdrawn.

posed of the rents a worker would get in expectation when searching for a new job ($\sum_i \eta^u_{ic} w^u_{ic}$ for union jobs and $\sum_i \eta^n_{ic} w^n_{ic}$ for nonunion jobs) and the relative probability of finding work in the union sector (T^u_{jc}).

We face two issues in constructing the expected rent terms. The first is that wage differences that reflect skill differentials are not rents that can be used in bargaining (a janitor cannot use the opening of a new law firm in town as the basis for bargaining a better wage). The second is that our specification as currently written involves regressing the mean wage in a given industry x city cell on the average wage in the city, which implies a standard reflection problem. We address these problems by starting with log wage regressions estimated at the national level, separately for each of our sets of sample years (the 1978-80, 1988-90, 1998-2000, and 2008-2010 sets). We include a complete interaction of education level dummies, a quadratic in age, and gender and race dummy variables. In one version of these regressions, we use only nonunion workers and include a complete set of city by industry dummy variables. We then use the coefficients on these variables as our dependent variable in the regression specification given by (3.25). This removes skill and demographic variation from our wage measure. In a second specification, we work with pooled union and nonunion worker data at the national level, running a log wage equation that includes the same set of skill and demographic variables plus a complete set of industry dummy variables. We interpret the coefficients on the industry dummies as rents.[22] In this specification, we interact the industry dummy variables with a union dummy, which allows industry rents to differ in the two sectors. We then replace the wages, w^u_{ic} and w^n_{ic}, with the industry premia, which we call ν^u_i and ν^n_i, in the outside option expressions. For example, we replace, $\sum_i \eta^u_{ic} w^u_{ic}$ with $\sum_i \eta^u_{ic} \nu^u_i$. Because this uses national level premia instead of local average wages, the direct reflection link is broken. The result is an outside option variable expressed such that workers in cities with a concentration in industries that pay high rents at the national level are able to bargain high wages. Because our specification includes industry fixed effects, we identify the effects of these outside options by comparing workers in the same industry across cities with different industrial and union compositions.

In our model we assume that the relative likelihood of finding work in the union sector differs by city, previous industry of work, and whether the worker was previously unionised. Using matched CPS data, sample sizes are simply not sufficient to warrant a fully flexible characterising of transition paths in this manner. However, using this data we can construct measures of union and industry relative transitions at the national level for each sample period. This variable captures the relative ability for workers to transition into unionised work depending on their past industry of employment and whether they previously worked in a unionised po-

[22] We define the industry dummy variables such that the coefficient values are defined relative to the overall average wage.

sition or not. We assume that local variability in finding unionised work is well captured by the fraction of unionised jobs.[23] We construct our measure as follows:

$$T_{jc}^{u} = \frac{T_i^u P_c}{T_i^u P_c + (1 - T_i^u)(1 - P_c)} \quad T_{jc}^{n} = \frac{T_i^n P_c}{T_i^n P_c + (1 - T_i^n)(1 - P_c)} \tag{3.27}$$

where T_i^n if the probability of transitioning from a nonunion job in industry i to a unionised job in any industry and P_c is the fraction of unionised jobs in the city. This is a simplified version of similar measures constructed by Tschopp (2017) who uses rich data to calculate transitions between industries.[24] As in Tscopp we interpret this as a measure of relative mobility into the union sector.[26] For a given worker it captures their relative ability to find union work over nonunion work. For a worker in Detroit who works in construction it measures the relative likelihood of matching with a union job over a nonunion jobs in any industry. The higher is this measure, the greater the likelihood workers will match with a union job, relative to a nonunion one. If there are many union jobs in Detroit, and construction workers tend to transition into unionised employment, then these workers will have a higher relative mobility measure of finding unionised employment relative to workers in cities with lower rates of unionisation, and who work in industries with relative low transition rates into unionised jobs.

Working with the wage rent variables and the transition rates, we form our measure of the outside option value as:

$$E_{jc}^{n} = \sum_i \eta_{ic}^{n} v_i^{n} + T_{jc}^{n} \left(\sum_i \eta_{ic}^{u} v_i^{u} - \sum_i \eta_{ic}^{n} v_i^{n} \right) = R_c + E_{jc}^{nd} \tag{3.28}$$

Thus, the outside option for a nonunion worker in industry j in city c can be written as the

[23] Though our framework assumes that bargaining effects operate only through the unemployment channel, that is, workers must first transition through unemployment to find a job, this transition measure is constructed, due to data limitations, using transitions between sectors which may, or may not, have included an intervening unemployment spell. There is a sense then in which the union outside option term captures on-the-job search dynamics. This is the case in a very simple on-the-job search framework in which workers use nonunion jobs as a means to move into higher paying union work. Formally modelling on-the-job search, or job laddering is beyond the scope of this paper, and as noted by Beaudry et al. (2012) is not straightforward and is sensitive to modelling of the search process and its relationship to wage determination. Still, it is worth highlighting that given the data restrictions we are facing, our transition measure likely embeds some dynamics arising from on-the-job search.

[24] To create the T_i^u and T_i^n variables, we construct transitions using additional data from IPUMS-CPS. For years 1990, 2000, and 2010, all transition data is constructed using IPUMS data. This is because IPUMS data contains a unique identification variable which allows for easy and accurate tracking of individuals over months of the CPS survey. For 1980, we match IPUMS identification data to the May extracts, as union data is not contained in IPUMS for these years, and there is little available information on how to link the raw data over time. The matching is based on household identifiers and personal characteristics. One limitation of the May extracts is that it is not possible to track individuals for most of 1981 and for all of 1982. The sample size is also notably smaller. To overcome this limitation we extend the range of years used to calculate transitions. Using the May extracts we match individuals from 1977 to 1981[25], and we match individuals from 1983 to 1984 using the MORG data.

[26] Tscopp's specification would take the same form if the economy was composed of two sectors.

sum of the average rent in city c among nonunion jobs, R_c, and the weighted difference between the average rent in union jobs and the average rent in nonunion jobs, E_{jc}^{nd}, with weights given by the probability of a nonunon worker in the j x c cell getting a union job. The R_{ct} variable is the same as the rent term used in BGS but only for nonunion jobs.

Changes in the outside option are driven by five factors. The first is changes in the local composition of nonunion jobs as captured in the η_{ic}^{n}'s. In essence, if high paying jobs such as those in the steel industry are replaced with lower paying service sector jobs then the outside options for all workers in the city is reduced. The second factor is the wage rents in the nonunion sector: even if there is no shift in the industrial composition of nonunion jobs, if the steel industry stops paying higher wages then it no longer offers an attractive outside option for workers in other industries. The third factor is the probability the nonunion worker can get a union job. If union jobs pay, on average, higher rents then a decrease in the probability of getting a union job means lower access to those rents and, therefore, a less valuable outside option. The value of the union option is also altered if there is a shift in the composition of union jobs (given by the η_{ic}^{u}'s) or the wage premia in the union sector (the ν_{i}^{u}'s), which are the fourth and fifth factors. The value of these latter factors are weighted by the probability of getting a union job.

3.3.3 Dealing with Endogeneity

We will estimate our derived specification in first differences in order to eliminate any industry x city time invariant characteristics. Ignoring selection issues for the moment, this means we are considering the regression given by:

$$\Delta w_{ict}^{n} = \gamma_{0it}^{n} + \gamma_{1}^{n}\Delta E_{ict}^{n} + \gamma_{2}^{n}\Delta ER_{ct} + \gamma_{4}^{n}\Delta\epsilon_{ict} \tag{3.29}$$

Given that the specification includes a complete set of industry x time period fixed effects, the relevant identifying variation for the estimated coefficients comes from across-city within-industry variation. That means that concerns about endogeneity centre on the question of whether city-level changes in productivity are correlated with the outside option and employment rate variables. There is clear reason to be concerned about such a correlation for the employment rate variable, which is at the city level of aggregation. In addition, BGS show that local industrial composition captured in the η_{ic}^{u} and η_{ic}^{n}) terms in the outside options value expression, (3.28), can be written as functions of ϵ_{ict}. Whether this implies an endogeneity problem depends on the time series processes of the productivity shocks. If they follow a random walk specification in which the changes in ϵ_{ict} are independent of their levels, then there is no endogeneity issue with this variable. Otherwise, there is reason to treat it as potentially

endogenous.

We address the potential endogeneity issues using Bartik style instruments. As Goldsmith-Pinkham et al. (2018) point out, Bartik instruments are functions of the start of period values for the η_{ict}'s - the local industrial composition - and any combination of those values can be used as an instrument. BGS argue that in our case one can find specific combinations with intuitive appeal within the theory by examining decompositions of the outside option variables. In this spirit, consider simple decompositions of those variables in equation 3.28:

$$\Delta R_{ct} = \sum_i \Delta(\eta^n_{ict}) v^n_{it} + \sum_i \eta^n_{ict+1} \Delta v^n_{it} \tag{3.30}$$

and,

$$\begin{aligned} \Delta E^{nd}_{jct} = T^n_{jct} \sum_i ((\Delta \eta^u_{ict}) v^u_{it} + \eta^n_{ict} v^n_{it}) + T^n_{jct} \sum_i \left(\eta^u_{ict+1} \Delta v^u_{it} - \eta^n_{ict+1} v^n_{it+1}\right) \\ + \left(\Delta T^n_{jct}\right) \left[\sum_i \left(\eta^u_{ict+1} v^u_{it+1} - \eta^n_{ict+1} v^n_{it+1}\right) \right] \end{aligned} \tag{3.31}$$

where we have left out terms related to changes in the nonunion industrial composition and nonunion wages from the expression for ΔE^{nd}_{jct} because they are already in the ΔR_{ct} decomposition. These expressions indicate that we can decompose changes in outside options into terms related to changes in the union transition probability, changes in the industrial composition of work in the union/nonunion sectors, and changes in the wage premia. Based on this, we form five instruments in three sets which rely on distinct sources of exogenous variation.

The first set of instruments we construct isolate variation in the outside option terms coming from changes in the industrial structure. We sever the link between changes in composition and local productivity shocks by using growth in industry employment at the national level to predict composition changes locally. These instruments then take the form:

$$\begin{aligned} IV1^n_{jctu} &= T^n_{jct} \sum_i \left(\left(\hat{\eta}^u_{ict+1} - \eta^u_{ict} \right) v^u_{it} + \eta^n_{ict} v^n_{it} \right) \\ IV1_{ctn} &= \sum_i \left(\hat{\eta}^n_{ict+1} - \eta^n_{ict} \right) v^n_{it} \end{aligned} \tag{3.32}$$

where $\hat{\eta}^u_{ict+1}$ and $\hat{\eta}^n_{ict+1}$ correspond to predicted values of end of period industrial shares in city c for union and nonunion jobs, respectively. We construct the predicted union industrial shares as follows (with the nonunion shares constructed analogously). We first construct predicted employment levels using start of period levels at the city level combined with national level

growth rates for the relevant industry:

$$\hat{N}_{icut+1} = N_{icut} \cdot \left(\frac{N_{iut+1}}{N_{iut}} \right) \tag{3.33}$$

We then form predicted city level employment as $\hat{N}_{cut+1} = \sum_i \hat{N}_{icut+1}$ and, from that, we construct predicted employment shares as $\hat{\eta}^u_{ict+1} = \frac{\hat{N}_{icut+1}}{\hat{N}_{cut+1}}$.

It is worth pausing to consider the conditions under which these instruments are valid. Recall that we include a complete set of time x industry effects and, so, we are working with within-industry, cross-city variation. Note also that the variation in the instruments comes from start of period, cross-city differences in the industrial proportions (the η_{ict}'s) and the start of period, cross-city differences in the union transition variable (T^n_{jct}). Validity of the instrument requires that these are uncorrelated with the relevant variation in the error term: cross-city variation in productivity growth. That is, we require random-walk-like assumptions that the productivity process follows a random walk (since, as BGS show, the $\eta'_{ict}s$ can be written as functions of the ϵ_{ict}'s) and that start of period union transition probabilities are uncorrelated with city-level productivity changes. We can potentially assess these assumptions using over-identification tests.

Our second instrument set is constructed as follows:

$$\begin{aligned} IV2^n_{jctu} &= T^n_{jct} \sum_i \left(\hat{\eta}^u_{ict+1} \Delta v^u_{it} - \hat{\eta}^n_{ict+1} v^n_{it+1} \right) \\ IV2_{ctn} &= \sum_i \hat{\eta}^n_{ict+1} \Delta v^n_{it} \end{aligned} \tag{3.34}$$

and isolates variation in the outside option term coming from changes in the wage premia. The model is built on an explanation for why unionized firms pay higher wages: that collective bargaining allows them to extract more of the match-specific surplus. In the appendix, we provide expressions for the wage equations written in terms of the deep parameters of the model. Based on that, the union premia for each industry can be written as a function of the final good price for that industry weighted by the union bargaining power parameter, the curvature of the production function, the discount rate, and the match death rate. The implication is that unions are better able to hold up firms in higher price/higher rent industries. The factors determining the size of the premia are all determined outside the model or at the national level rather than the local level. On this basis, we treat them as exogenous.

Our final instrument isolates variation related to the probability a nonunion worker in a particular industry x city cell moves into a union job. We construct our measure of this probability, as shown in (3.27), using a combination of national level probabilities of nonunion workers from a given industry finding a union job and the local proportion of workers who

are unionised. The first of these varies at the national industrial level and so, in our regressions including industry fixed effects, it does not represent a problematic source of variation. This is not the case for the locally defined union proportion term. To instrument for this term then, we make clear how the union proportion term is tied to the local industrial structure:

$$P_{ct} = \frac{\sum_i N_{icut}}{\sum_u \sum_i N_{icut}} \tag{3.35}$$

Given this, it is possible to predict changes in the local unionisation using national changes in the composition of work. That is, we use national level predictions in employment in jobs to predict local employment growth. If there are declines in union employment in sectors with high local employment, then this will predict a significant decline in the union proportion. Our predicted union proportion term is calculated as:

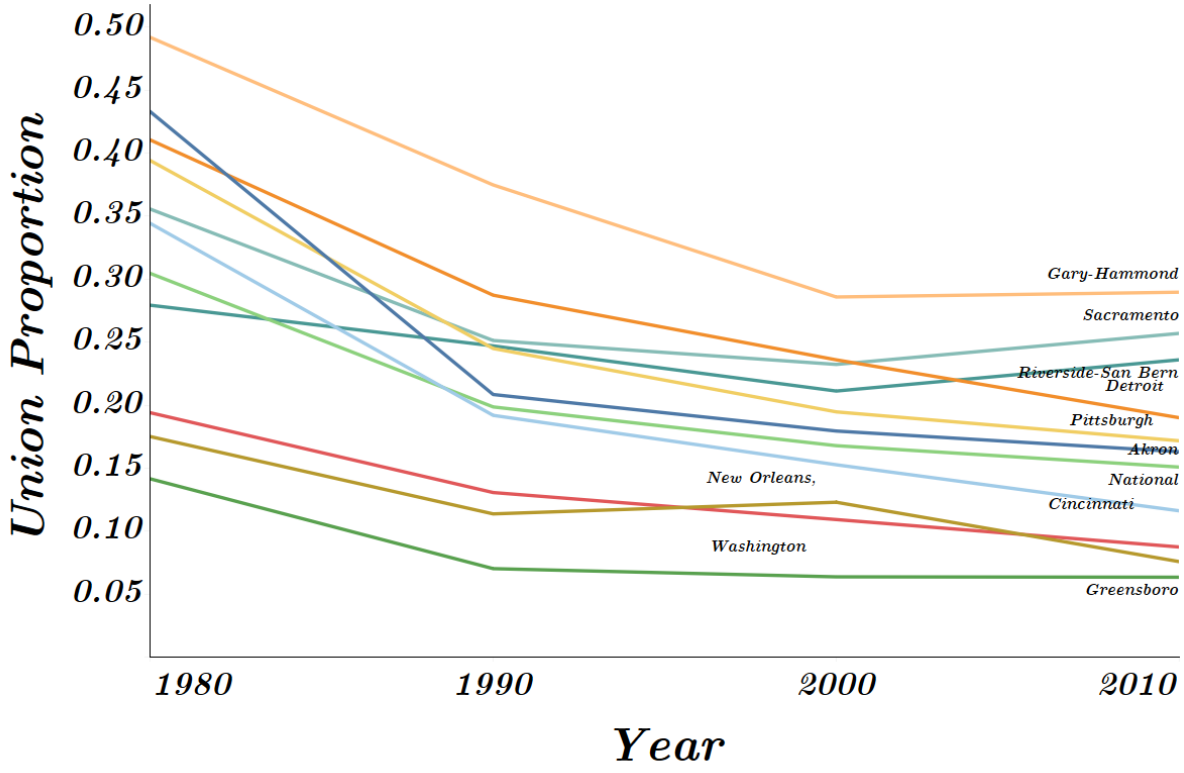
$$\hat{P}_{ct} = \frac{\sum_i \hat{N}_{icut}}{\sum_u \sum_i \hat{N}_{icu}} \tag{3.36}$$

where, we defined $\hat{N}_{icut}$ when describing the first instrument set. Given this, our final instrument is given by:

$$IV3^n_{jct} = \left(\hat{T}^n_{jct+1} - T^n_{jct}\right)\left[\sum_i \left(\hat{\eta}^u_{ict+1} v^u_{it+1} - \hat{\eta}^n_{ict+1} v^n_{it+1}\right)\right] \tag{3.37}$$

Following BGS and our earlier discussion, the condition required for the validity of this instrument is that growth in local productivity is independent of past values of the relative industry-city advantage, and of past values of the proportion unionised. The validity of this assumption relies on the exogeneity of the start of period industry shares and the union proportion. In future work we plan to explore the correlation between these variables and other potential drivers of changes in wage growth, identifying the shares with the largest Rotemberg weights (following Goldsmith-Pinkham et al. (2018) to identify those most important to ensuring exogeneity. Jaeger et al. (2018) further identify potential problematic correlation in Bartik instruments over time, which makes it impossible to separately identify short and long term effects. Though the correlation across time in our instruments is much lower than that found by the authors for the standard shift-share instrument used in the immigration literature, we will devote more attention to this issue in future work.

Figure 3.1: Declining Unionisation Across Selected Cities



Notes: The proportion of unionised workers is plotted from 1980-2010 for selected cities. Cities with the largest, and smallest decline over 1980-2010 are presented.

3.3.4 Descriptive Patterns

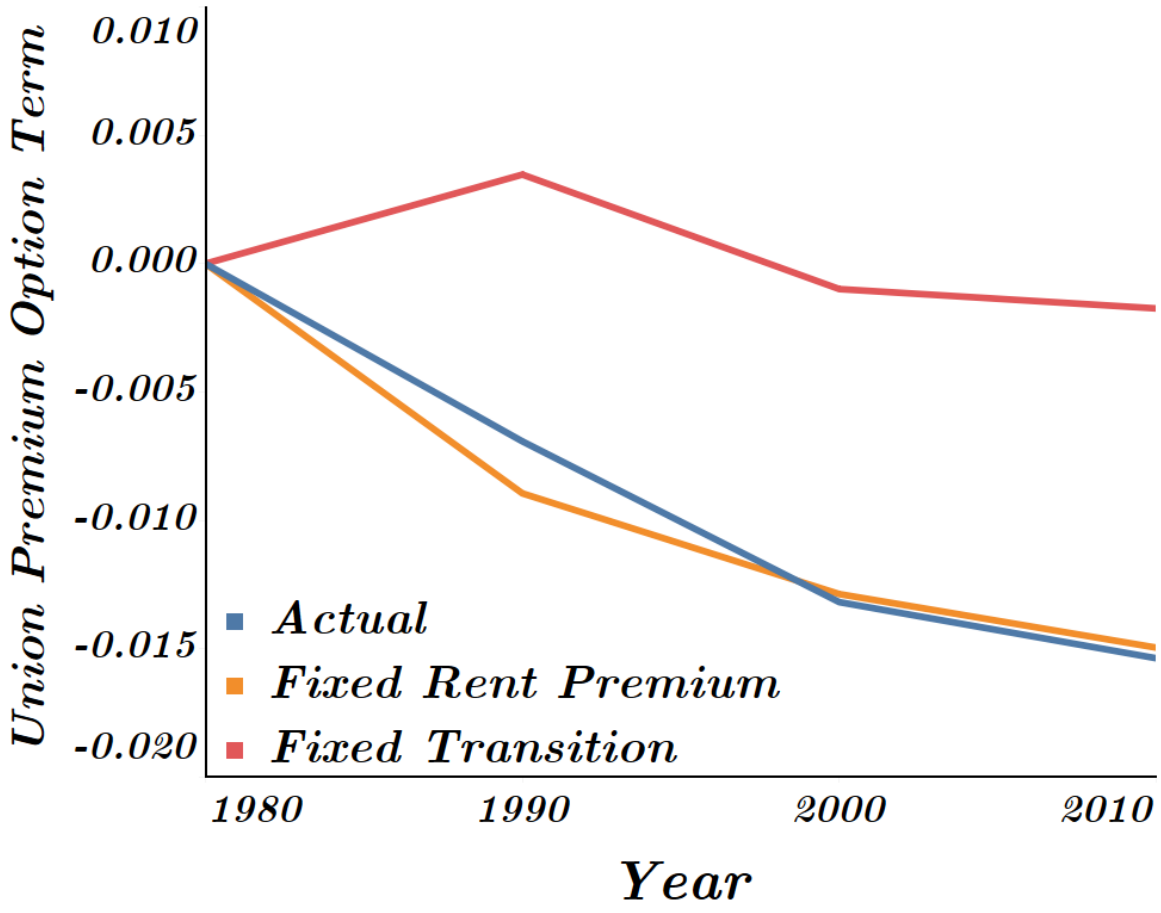
Before turning to estimation we first explore key patterns in unionisation over our sample period as pertains to our framework. That is, we explore patterns in union/nonunion rents, and in the fraction of workers unionised. As is well known, the decline of unionisation in the United States over 1980-2010 (and for other rich world nations over a similar time frame, see Schmitt and Mitukiewicz (2012) and Lesch (2004)), has been remarkable.

In Figure 3.1 we plot the fraction of workers unionised at the city level over 1980-2010. We plot trends for the national average, and for 10 cities that experienced the largest, and smallest, percentage declines over this period. On average, 30% of jobs were unionised at the city level in 1980, but this number declines to 15% by 2010. Small declines are observed in cities with low initial rates of unionisation, such as New Orleans, Washington, and Greensboro. In cities, like Detroit, Gary, and Pittsburgh, where the union sector played a much larger role in the 1980 economy, the declines are much larger: respectively 21, 22, and 22 percentage points. We present the same patterns at the state level in Figure C.1 where we observe especially large declines for Michigan, Pennsylvania, Ohio, and Tennessee. We again see wide variation in the decline in unionisation across states, ranging from an almost 80% decline for Tennessee to just over a 25% decline in Hawaii.

Observed declines in rates of unionisation are large, which naturally has implications for our outside option terms, where weight is attached to nonunion and union work in proportion to the share of jobs in each sector. In panel Figure 3.2 we plot trajectories in the union contribution to workers outside options ($T_{ic} \times (R_c^u - R_c^n)$). There has been a sustained decline in workers unionised outside options over the period 1980-2010 with a decline observed each decade. Given the patterns already presented in declining rates of unionisation over 1980-2010, and especially the rapid decline between 1980-1990, this pattern is hardly surprising. Over 1980-1990 a much greater weight will be placed on nonunion work when individuals bargain with employers, given the relative abundance of nonunion work. The sustained decline between 1980-2000 despite the rapid decline in unionisation rates over the first decade can in part be explained by the pattern outlined in Figure C.2 which presents time trends in the proportion of union jobs and the transition probability of workers into union jobs. Interestingly, the decline in the transition probability is lower than the decline in the proportion of unionised jobs, especially over the 1980-1990 period. From 1990-2010 these measures experience a parallel decline.

To examine what is driving variation in the outside option term we fix in turn the transition probability, and the union rent premium ($R_c^u - R_c^n$) at their 1980 levels to examine counterfactual movements in outside options. This exercise makes clear the importance of the declining union transition over this period. Had it stayed fixed over the sample period, only a very small decline in union outside options would have been observed due to declining union rents relative to

Figure 3.2: Time Trends in Union Outside Option Premium



Notes: Time trends for the union premium component of outside options: $T_{ic} \times (R_c^u - R_c^n)$. Raw averages are calculated by year across city-industry cells and normalised to 1980. Outside options are constructed using our core sample of 50 industries and 93 cities. 'Fixed Transition' is the trajectory for the union premium term if T_{ic} is fixed at its 1980 value. 'Fixed Rent Premium' is the trajectory for this term if the rent premium $(R_c^u - R_c^n)$ is fixed at its value in 1980 for each city.

nonunion rents. In fact, we observed an increase in this counterfactual measure between 1980-1990 coming from an initial increase in union rent premia over this period. This increase is more than offset by declines in the following two decades. In fixing the union rent premia we can trace the decline in outside options driven solely by variation in the likelihood of finding a unionised job. Much of the actual decline in workers outside options is driven by this measure, but increasing union rent premia over 1980-1990 offset some of the initial decline in outside options over this period. Approaches which fail to account for the changing composition of union/nonunion work, then miss an important counterbalancing force over 1980-1990.

3.3.5 Results

We now turn to estimation results based on the model presented in section 2. This search model considers spillover effects operating between the union and nonunion sectors through the bargaining channel: nonunion workers can bargain a higher wage if their outside wage options are higher. In section 2.5 we extended this simple framework to endogenise union formation which is driven by elections at the level of the firm. Incorporating union formation makes clear issues related to selectivity in estimating the effect of outside options for nonunion workers. We begin by presenting results without considering selectivity, focussing on endogeneity issues related to the right hand side variables.

OLS Results

We first present OLS results from estimating nonunion wage equation (3.18) in Table 3.1. In all specifications work in first differences and include industry-year dummies. The variation used to estimate the coefficient on outside options is then operating within-industry, across-city. We cluster standard-errors at the city-year level.

We begin in column (1) by considering a very simple specification which mirrors the empirical exercise of previous research considering spillover effects operating through variation in the size of the union sector. We include changes in the union proportion in an industry-city cell as our core explanatory variable here, finding a positive relationship between wage growth, and growth in unionisation. This is somewhat in line with previous results, although, as mentioned, prior results are sensitive to the specification used and the variation exploited for identification. In exploiting variation across cities our results are comparable to those of Hirsch and Neufeld (1987) and Holzer (1982) amongst others. Hirsch and Neufeld (1987) find a coefficient estimate on the union proportion in the region of .25-.58 for nonunion workers when exploiting industrial variation, although they find little evidence of a spillover effect operating at the local level. Holzer (1982) finds a positive spillover effect using the rate of unionisation at the SMSA for

white males, although his results are somewhat sensitive to the inclusion of supply and demand shifters, and the sample time frame.

In column (2) we control for industry-year fixed effects which leads to a large decrease in the magnitude of the coefficient on the proportion union from .27 to .041. In controlling for industry-year fixed effects our results align more closely to those of Neumark and Wachter (1995) who control separately for industry and year effects. In contrast to our results however they find a negative relationship between wages and the percent organised. Where our estimation differs is in exploiting variation within industries across cities, and controlling fully for industry-year omitted variables. They do not consider regional variation, instead exploiting variation across industries at the national level over time. Farber (2005) exploits variation in the probability of unionisation across states and industries in the cross-section, and in turn controls for state and industry fixed effects. When controlling for industry and state fixed effects Farber finds a coefficient estimate on the probability of unionisation around .18 which declines significantly over time. This result is more similar to ours than that of Neumark and Wachter (1995), but again the source of variation is not directly comparable. In taking first differences we are controlling for any fixed city-industry effects, and we additionally control for common industry trends in wage growth.

In column (3) we include the variables derived from our model which separately capture the nonunion and union contribution to workers' bargaining positions. Once we include these variables the coefficient on the union proportion term declines by a factor of 10 and is no longer statistically significant. This suggests that any threat effects captured by this measure occurs through the bargaining channel reflected in our outside options variables. Our results indicate a strong positive relationship between nonunion rents and wages, and also the union outside option term and wages. In column (4) we additionally control for the employment rate, and in column (5) we drop the union proportion term which has little effect on our estimates. Controlling for changes in the city employment rate serves to reduce our coefficient estimates although the effects are still quite large. It is worth highlighting that the coefficient estimate on our nonunion rent is lower than that found by BGS for the entire population. This can be explained by the fact that their rent term embeds the union sector, and that the sample used for estimation in their paper is substantially different than ours. In particular, BGS are able to use Census data and hence exploit variation across 152 cities and over 100 industries. Our estimates then may represent an underestimate, given our inability to exploit more disaggregated data which better captures local labour market variability.

In the bottom rows of the table we test for a significant difference in the coefficients on the union and nonunion outside option terms and reject the hypothesis that they are equal. This violates the prediction of our model, and suggests instead that in bargaining, union outside

Table 3.1: OLS Results

	(1)	(2)	(3)	(4)	(5)
ΔP_{ic}	0.27***	0.041**	0.0041	0.0080	
	(0.028)	(0.021)	(0.019)	(0.019)	
ΔR_c^n			2.38***	1.99***	1.99***
			(0.29)	(0.30)	(0.30)
ΔE_{ic}^n			3.97***	3.36***	3.38***
			(0.66)	(0.60)	(0.59)
ΔER				1.06***	1.06***
				(0.18)	(0.18)
Observations	6925	6925	6925	6925	6925
R^2	0.023	0.44	0.48	0.49	0.49
Year × Ind.	No	Yes	Yes	Yes	Yes
			Test $\gamma_{11}^n = \gamma_{12}^n$		
p-val:			.02	.03	.02
F-Stat.			5.39	4.92	5.21

Notes Standard errors in parentheses clustered at the city-year level. * denotes significance at the 10% level, ** denotes significance at the 5% level, *** denotes significance at the 1% level. The dependent variable is the change in the regression adjusted average hourly wage of nonunion workers in an industry-city cell. Estimates obtained using decadal changes over 1980-2010 across 50 industries and 93 cities.

options are weighted more heavily. It would fit, for example, with union wages being more salient for all workers and therefore a more potent reference in bargaining. However, given the likely endogeneity in our estimates we cannot conclude that these test statistics are valid. Additionally, as we will make clear in section 4, firms can respond to the union threat though wage emulation whereby they pay workers a premia correlated with the outside options of unionised workers. To the degree that the union contribution term is correlated with wage emulation this will affect our results.

IV Results

Estimates obtained using the instrumental variables outlined in section 3.3 are presented in Table 3.2. These IVs exploit exogenous variation coming from national level changes in industrial composition, or wage premia. One set of IVs corresponds to variation coming from the changing ability to find unionised work. The two other sets correspond to changes in wage premia and changes in the composition of work across industries. Across columns we consider alternative combinations of these instruments, and in the bottom panel of the table we present test statistics and p-values associated with the Sanderson-Windmeijer test (Sanderson and Windmeijer, 2016) for weak instruments in the first stage.[27] In column (1) we include the full set of IVs, including both union and nonunion outside options as our core explanatory variables. The instrument set includes one for the employment rate which is the classic Bartik instrument: the weighted average of national industrial growth rates, using the start of period industrial shares for the city as weights. However, we find that this instrument is highly correlated with our other instruments and results in large estimates of the impact of the employment rate that are out of line with those in BGS. Though the Sanderson-Windmeijer F-statistic is 10.72, indicating a significant first stage, much of the variation used to predict the employment rate is coming from instruments associated with the union contribution to outside options. In the remaining columns of the table, we instrument for the outside option values but not for the employment rate. We follow Stock and Watson (2011) in interpreting the employment rate as a control variable - a variable that is not of direct interest in its own right but is useful for picking up its own effect and those of correlated omitted variables. In our case, we view the employment rate as capturing its own effect plus the impact of general, local demand shifts. This allows us to isolate the outside option effects we care about from demand effects. The estimates for the proportion union and the outside option values in column (2), when we do not instrument for the employment rate, are very similar to those using the employment rate instrument. In comparison to

[27] This test statistic provides a test of weak instruments when there are multiple endogenous variables, and is similar to the Angrist-Pischke test, but includes a modification to ensure the correct asymptotic distribution of the test statistic.

Table 3.2: IV Results

	(1)	(2)	(3)	(4)	(5)
ΔP_{ic}	0.016	0.015			
	(0.021)	(0.021)			
ΔR_c^n	1.45***	1.58***	1.59***	1.58***	1.53***
	(0.54)	(0.54)	(0.54)	(0.55)	(0.56)
ΔE_{ic}^n	2.40**	2.56**	2.59**	2.26**	3.42**
	(1.14)	(1.09)	(1.07)	(1.14)	(1.43)
ΔER	1.71***	1.16***	1.16***	1.17***	1.14***
	(0.56)	(0.21)	(0.21)	(0.21)	(0.21)
Observations	6925	6925	6925	6925	6925
R^2	0.49	0.49	0.49	0.49	0.49
Year × Ind.	Yes	Yes	Yes	Yes	Yes
ERIV	Yes	No	No	No	No
IVs	all	all	all	$IV1$-$IV2$	$IV1^n$-$IV2^n$-$IV3$
F-Stats:					
ΔR_c^n	33.19	39.47	39.75	45.18	51.71
ΔE_{ic}^n	73.24	70.65	71.12	53.26	41.97
ΔER	10.72				
p-val:					
ΔR_c^n	0.00	0.00	0.00	0.00	0.00
ΔE_{ic}^n	0.00	0.00	0.00	0.00	0.00
ΔER	0.00				
			Test $\gamma_{11}^n = \gamma_{12}^n$		
p-val:	.50	.47	.45	.62	.28
F-Stat.	.46	.53	.57	.24	1.19

Notes Standard errors in parentheses clustered at the city-year level. * denotes significance at the 10% level, ** denotes significance at the 5% level, *** denotes significance at the 1% level. The dependent variable is the change in the regression adjusted average hourly wage of nonunion workers in an industry-city cell. Estimates obtained using decadal changes over 1980-2010 across 50 industries and 93 cities.

our OLS estimates the coefficient on the nonunion rent term is about 20% lower while the coefficient associated with the union outside option term declines by about a third. We find that the union proportion term is still not significant here and we drop it for our results in columns (3)-(5).

In column (3) we present our preferred specification which includes all instruments for outside options, excludes the instrument for the employment rate, and excludes the union proportion term. Statistical tests indicate strong first stage estimates for both of our outside option terms indicating we do not suffer from weak instruments. Furthermore, in contrast to our OLS estimates we cannot conclude here that the coefficient estimates on the nonunion rent and the union outside option term are statistically distinguishable. This is in part due to the larger decrease in the union outside option term when using plausible exogenous variation, but also due to increased noise in our estimates.

In columns (4) and (5) we consider alternative subsets of instrumental variables to test the sensitivity of our results. Recall that we have three sets of instruments: a set based on shifts in industrial composition; a set based on changes in wage premia; and an instrument based on shifts in the union proportion. From our discussion of selection, the third instrument is potentially problematic because it could be related to the extent of any selection into the union sector as well as to the direct effects on outside option values. As a result, it is useful to estimate a specification in which we only use the first two instrument sets, freeing up variation related to the proportion union to be used in identifying selection effects. In column (4) we restrict the set of instruments used to that which exploits variation in wage premia or employment shares. In these specifications, we identify the effect of outside options by working with variation in the average rents rather than through the proportion union. Our results are remarkably similar to those in column (3), with a mild decline being observed in the union outside option term. In column (5), we instead include the nonunion industry share and wage premia variables as instruments for R_{ct} but only include IV3, which focuses on the probability of moving into a union job, to instrument for the difference between union and nonunion outside options, $E_c^{nd}t$. Here the coefficient on the union outside option term increases by around 50% but the nonunion outside option term has a very similar estimated effect to the other specifications. Thus, we can use variation only related to industrial composition and rents to identify the outside option effects, getting very similar results to those when we use variation in the union proportion. In section x, we will use the latter variation to identify selection effects. Note that the fact that we get the same results with different subsets of our instruments means that we pass an overidentification test arising from the model. What matters for wage bargaining, according to the model, is changes in the value of outside options, whether those arise from shifts in the local industrial composition, in the rents paid in different industries, or in the probability of getting

a union job. The fact that all of those forms of variation generate the same effects when parsed through our outside option terms is a strong piece of evidence in favour of the model.[28]

Counterfactual Exercise

Our results thus far have indicated a strong, significant relationship between quality job opportunities in both the nonunion and union sectors and wages. From these results, the question arises of how changing patterns of unionisation have affected wage growth over 1980-2010, a period of rapid deunionisation. Descriptive patterns regarding outside option terms and rents suggest that both declining unionisation and stagnating growth in union rents (post-1990) have potentially large wage effects. In this section, we sequentially decompose changes in wages as predicted by our estimated wage equation to examine the counterfactual path for wages had the composition of work and outside options been unchanged over the period 1980-2010.

We start with total average wages at the city level which are the weighted average between nonunion and union earnings at the city level, where the weight is the proportion unionised at the city level:

$$w_{ct} = P_{ct} \times w^{u}_{ct} + (1 - P_{ct}) \times w^{n}_{ct} \tag{3.38}$$

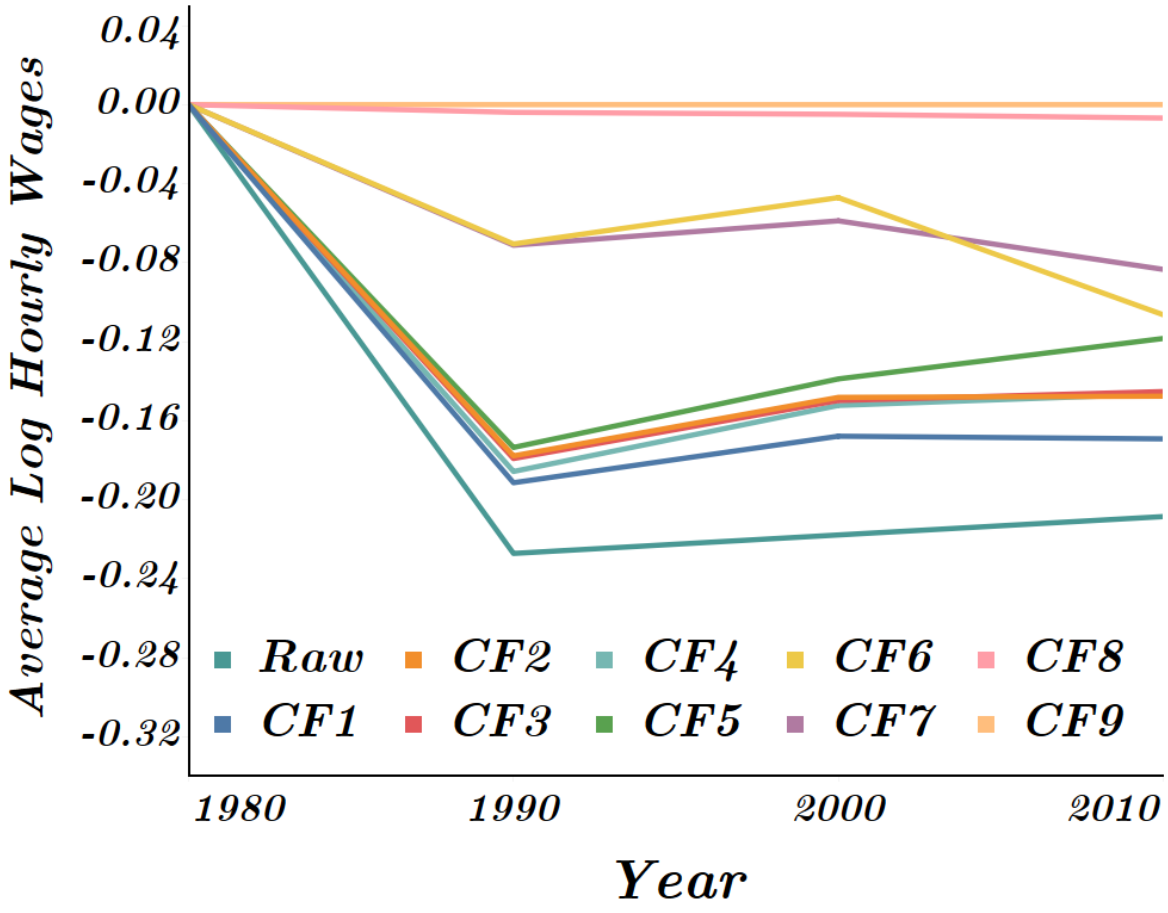
We use the estimated industry-city wages that we employed as the dependent variable in our regressions combined with industrial shares at the local level to create city level wages as follows: $w^{n}_{ct} = \eta^{n}_{ict} w^{n}_{ict}$. To estimate time trends we average across cities each period, using city populations as weights. Figure Figure 3.3 contains the decomposition of this trend, which is the bottom line in the figure. It depicts an overall real wage trend that is strongly decreasing between 1980 and 1990 - falling by approximately 25% in that decade - and then mildly increasing in the following decades.

The first step in our decomposition consists of holding the union proportions, P_{ct} in (3.38) constant. This is the most direct effect of unions, showing what would have happened to the overall average wage if both union and nonunion wages had still pursued their trends but there had been no change in the weights attached to each. This is represented in the CF1 line in the figure. According to that line, this channel for the impact of the unionisation decline generated an 3.9% drop in the mean wage, accounting for about 19% of the overall drop in the mean wage in the 1980s and about a fifth of the drop between 1980 and 2010.

In the remaining steps in our decomposition, we focus on movements in the nonunion wage, i.e., on union spillover effects into the nonunion sector, because that is what we actually

[28]More formally, BGS argue in a similar context that their outside option instruments would be invalid if the change in productivity at the city level were correlated with the start of period industrial composition. The different instruments weight that offending potential correlation differently so that if the correlation is non-zero, they should generate different estimated effects. The same is true in our context and, so, the lack of difference in estimates with the different instrument sets implies that the data fits with our core identifying assumption.

Figure 3.3: Average Wage Decomposition



Notes: We fully decompose average log hourly wages in stages using our estimated wage equation. We use estimates obtained from column (3) of Table 3.2 to examine counterfactual wage movements in nonunion wages. We construct city wages by aggregating estimated industry-city wages to the city level. Using the city level proportion of jobs that are unionized we construct the average wage as $\hat{w}_c = P_c \times \hat{w}_c^u + (1-P_c) \times \hat{w}_c^n$ where $\hat{w}_c^n = \sum_i \eta_{ic}^n \hat{w}_{ic}^n$ and $\hat{w}_c^u$ is similarly constructed. 'Raw' is the actual movement in $\hat{w}_c$ over the sample period. 'CF1' fixes the union proportion P_c at its 1980 level. 'CF2' further fixes the probability of transitioning into a unionised job. 'CF3' additionally fixes the industrial composition of work in the union sector: η_{ic}^u. 'CF4' also fixes the union industry premia over nonunion premium. 'CF5' further fixes the industrial composition of work in the nonunion sector. 'CF6' fixes nonunion wage premia which enter both through outside options and through fixed effects in the estimated wage equation. 'CF7' fixes the Employment rate and 'CF8' additionally fixes the union wage $\hat{w}_c^u$

estimate. We provide a brief discussion of union wage movements at the end. Thus, in the second step in our decomposition, we fix the probability that a nonunion worker can find a union job, T^n, which has its impact through the outside option effect on nonunion wages. This, as reflected in the CF2 line, suggests that the mean wage would have been 0.014 log points higher in 1990 and 0.022 log points higher in 2010 if the ability of nonunion workers to transition into unionised work had remained unchanged at 1980 levels. This amounts to about one-eighth of the long term wage decline. In our variable construction, T^n varies by the initial, nonunion sector industry and the city but not by the destination industry. It could matter whether de-unionisation happens differently in high wage compared to low wage sectors. That is captured in changes in the local industrial composition of union jobs, the η^u_{ict}'s, holding changes in the nonunion composition constant. We generate this effect in our decomposition by allowing the nonunion composition to follow its actual path but hold changes in the union composition constant. That is what is reflected in the CF3 line and it shows little added effect. Next, we hold constant the union wage premia by industry. Because those premia actually increased in the 1980s, this has an offsetting effect on the mean wage compared to the changes in unionisation rates over that decade. In the longer run, union premia have declined and the overall effect is negligible relative to the unionisation rate effects. Overall, fixing all variables related to union outside options leads to a mild increase in nonunion wages in 1990, but a much larger increase over the full sample period. Our estimates indicate that declines in union power reduced non-union wages by about .03 log points over the entire period through their impacts on outside options. This implies a .022 log point decrease in the overall mean wage which, when combined with the direct union proportion effect in CF1, leads to the conclusion that declining union power accounts for 29.8% of the overall mean wage decline between 1980 and 2010.

The next step in the decomposition exercise finds a small impact of the industrial composition of nonunion work on wages over the 1980-90 period, with a larger effect between 2000-2010. Note that this is the effect operating through the outside option term. The direct effect of shifts in nonunion industry premia (seen in the move from CF5 to CF6) are substantial over the 1980s and then decline thereafter. Changes in the employment rate have little effect on wages between 1980-1990 but serve to lower wages in 2000 and increase them in 2010. Declining union wages explain .07 log points of the .22 point decline in wages between 1980 and 1990, as seen in the move from C7 to C8. They then experience a mild increase over 1990-2000 before declining further over 2000-2010. Based on our model, some portion of the decline in union wages over time likely stems from the decline in unionisation affecting the outside options of union workers, as it does for nonunion workers. Given sample size issues, we were not able to estimate the outside option effects for union workers directly but we can get a rough notion of their magnitude by constructing changes in the outside options variables they face and using

them in combination with the impact coefficients we estimated for nonunion workers.

Note that the outside options for union workers are different because their probabilities of transiting to another union job are higher than for a nonunion worker transiting to a union job. Based on this, our rough estimate is that 61% of the change in union wages is due to a change in their outside options induced by the decline in unionisation in the economy as a whole. If we add this to the spillover effects we estimated for nonunion workers and the simple effect through reweighting the mix of union and nonunion wages captured in the first counterfactual, then we get that approximately a 21% of the decline in the mean real wage over the 1980s and a 47% of the decline between 1980 and 2010 can be attributed to the decline in the union sector. In summary, the predictions of our model suggest both an important role for outside options in wage fluctations, and an interesting counterbalancing effect operating between 1980-1990 whereby relative increases in union premia offset the declining ability of nonunion workers to find union jobs.

3.3.6 Alternative Specifications

In this section we consider the sensitivity of our results to excluding the public sector, and to an alternative specification for outside options.

Alternative Transition Measure

We consider an alternative specification which uses the proportion of unionised jobs at the city level as our measure of the ability of workers to transition into unionised jobs. One limit of our transition measure is that we are restricted to capturing transitions at the national level by industry. Though we believe there are clear benefits to incorporating this data, it may serve to mask important local variation. Table C.4 presents results using the proportion union as our transition measure. Our union outside option term is now defined at the city level: E_c^{2n}. The results for this specification are similar to our preferred specification, although the impact of union outside options is larger (though not statistically different). A counterfactual exercise built upon these estimates would yield similar patterns to that in Figure 3.3 but would imply a much greater role for outside options as the union proportion term declines much more rapidly than transitions into the union sector. If declines in local transitions are correlated with movements in the local proportion union term, then our results may serve as a conservative estimate of the impact of union outside options on wages.

Removing the Public Sector

Thus far we have included the public sector, both in the construction of our outside option terms, and as an observation on the left hand side of the wage equation. Card et al. (2018) however outline the marked difference in unionisation between the private and public sectors since 1980 such that unionisation is now 5 times higher in the public sector. Estimates including the public sector then will understate the average decline in union outside options which may affect our estimates. Additionally, to the degree that public sector jobs engage in a unique and distinct wage setting process to the private sector this may bias our results away from finding evidence of wage bargaining effects. In columns 1 and 2 of Table C.5 we present results, dropping $i - c$ observations for industrial jobs in the public sector. Additionally, in columns (3) and (4) we exclude the public sector in the construction of outside option terms. In the first instance our results change very little, though when we exclude the public sector from the construction of rents, we find some evidence of both an increased impact of union outside options, and rents on wages. In general though our results are robust to inclusion of the public sector.

3.3.7 Controlling for Selectivity

Thus far we have presented estimates using IVs to break the linkage between local productivity shocks and growth in outside options. Section 2.5 however makes clear that there is likely selectivity into the union sector based on productivity draws. This arises because productivity shocks are differently weighted in the union and nonunion sector due to alternative methods of wage bargaining. If this were not the case, then selection on productivity would take place indirectly through its effect on local rents. In particular, if outside options are changing over the sample period and if they are weighted differently in the bargaining process between union and nonunion jobs, then the threshold level of productivity necessary to induce unionisation will change. In this way, outside options are linked to changes in selectivity on ϵ draws between periods, and this could bias our regression estimates.

We address selection through a standard, generalized Heckman two-step approach (see Heckman (1979), Dahl (2002), Snoddy (2019)). The idea in this approach is that the error mean term in (3.25), $E(\epsilon_{ic}|I_{icf} \leq 0)$, creates an omitted variables bias that can be addressed by including a fitted error mean term as a covariate in the regression. Further, the error mean can be expressed as a non-linear function of the probability of selection (the probability of being nonunion in our case) or of exogenous variables that drive that probability. As we discussed in Section 2.5.1, the fact that the error mean term is a function of the nonunion probability raises difficulties for identification in the standard specification in which the union proportion is entered as a right hand side variable. However, our instrument set includes instruments that are not a function

Table 3.3: Controlling for Selectivity

	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
ΔR_c^n	1.99***	1.55***	1.99***	1.55***
	(0.30)	(0.55)	(0.30)	(0.55)
ΔE_{ic}^n	3.35***	2.18*	3.45***	2.42**
	(0.60)	(1.14)	(0.58)	(1.14)
ΔER	1.06***	1.18***	1.06***	1.17***
	(0.18)	(0.21)	(0.19)	(0.21)
Observations	6925	6925	6905	6905
R^2	0.49	0.49	0.49	0.49
Year × Ind.	Yes	Yes	Yes	Yes
ERIV		No		No
IVs		IV1-IV2		IV1-IV2
P_{ic} Quartic	Yes	Yes	No	No
Election Vars.	No	No	Yes	Yes
F-Stats:				
ΔR_c^n		46.02		46.75
ΔE_{ic}^n		56.13		57.5
p-val:				
ΔR_c^n		0.00		0.00
ΔE_{ic}^n		0.00		0.00
	Joint Test on Selectivity Controls			
p-val:	.75	.60	.30	.44
F-Stat.	.48	.68	1.22	.97

Notes Standard errors in parentheses clustered at the city-year level. * denotes significance at the 10% level, ** denotes significance at the 5% level, *** denotes significance at the 1% level. The dependent variable is the change in the regression adjusted average hourly wage of nonunion workers in an industry-city cell. Estimates obtained using decadal changes over 1980-2010 across 50 industries and 93 cities.

of the union proportion and we have seen that we obtain very similar results whether or not we include the union proportion and union transition rate variables as instruments. Thus, we can take an approach in which we use the restricted instrument set that does not include the union proportion to identify the effects of our rent variables, using functions of the union proportion or related variables to absorb the selectivity effect.

Given these arguments, we examine potential selection effects using two sets of variables. First, we include a quartic in the change in the proportion of workers in the industry x city x time cell who are unionised. In doing this, we are taking the model very seriously and assuming that the proportion union does not determine nonunion wages directly - any effect it has reflects selection. We present both OLS and IV estimates of our main specification including the quartic in the change in the union proportion. For the IV estimates, we use the restricted IV set that does not include IV3 - the instrument that focuses on variation in the unionisation rates.

Following Fortin et al. (2019), we also estimate a specification in which we proxy for costs of unionisation using National Labor Relations Board (NLRB) data on certification elections.[29] In particular, we calculate the number of workers involved in certification elections in each city over a three year window around the years 1980-1990-2000-2010, divided by the number of nonunion workers in the city. We also construct a second measure: the fraction of certification elections won over the same three year window. The first measure then calculates the ability of workers to hold union elections, even if they are ultimately unsuccessful. In areas with high costs to unionise we would expect relatively few elections to take place. Furthermore, we would expect costs to unionise to be higher in areas where the chance of winning an election is low. The costs necessary to mount a successful campaign are likely greater when the chance of winning is low. In using these two variables to proxy for the costs of unionising we include them as linear controls, and include their interaction which roughly approximates the successful share of nonunion workers involved in certification elections. We include a quadratic in each term.

We present OLS and IV results in columns (1)-(4) of Table 3.3. In both OLS and IV specifications our results change very little when compared to our main results presented above. Our results indicate then that there is little selectivity driving results in our main specifications. We have discussed instances in which this situation may arise, however, we cannot rule out the possibility that our election variables are not good proxies for the costs of forming a union, in particular since the joint test of significance for our full set of selection IVs is not significant. Nevertheless, there are certainly intuitive reasons why the ability to both hold and win elections would proxy for the costs of forming a union.

[29] See section 1.3.1 for details on the data.

Table 3.4: Subsample Analysis - Coefficient Estimates on Outside Options

	OLS	IV	Sample
	Full Sample: Alternative Geography		
City	2.10***	1.71***	6925
	(0.29)	(0.46)	
State	2.55***	2.01***	4855
	(0.39)	(0.47)	
	Estimates by Sex		
Male	1.93***	1.29***	2842
	(0.34)	(0.41)	
Female	2.03***	2.15*	
	(0.38)	(0.74)	
	Estimates by Age		
Age 20-35	2.57***	2.38***	3052
	(0.34)	(0.39)	
Age 36-55	1.73***	2.00***	
	(0.42)	(0.55)	
	Estimates by Education		
$\leq HS$	1.60***	1.21***	3040
	(0.29)	(0.32)	
$> HS$	2.55***	2.20**	
	(0.42)	(0.87)	
	Males Only: Estimates by Age		
Age 20-35	2.23***	1.86***	1954
	(0.36)	(0.44)	
Age 36-55	1.17***	1.51***	
	(0.38)	(0.52)	
	Males Only: Estimates by Education		
$\leq HS$	1.46***	1.08***	1642
	(0.32)	(0.38)	
$> HS$	2.38***	2.73***	
	(0.38)	(0.72)	

Notes Standard errors in parentheses clustered at the state-year level. * denotes significance at the 10% level, ** denotes significance at the 5% level, *** denotes significance at the 1% level. All estimates for subsamples are performed using state-industry level variation. The dependent variable is the change in the regression adjusted average hourly wage of nonunion workers in an industry-state cell. Estimates obtained using decadal changes over 1980-2010 across 50 industries and 50 States. The coefficient estimated presented is on the pooled outside option term $E_{ic}^{4n} = R_c^n + E_{ic}^n$.

3.3.8 Subsample Analysis

In all regression results presented thus far, and in the counterfactual exercise in Figure 3.3 we are considering wage fluctuations for the population as a whole. As is well established in the literature, historically unionised jobs were concentrated among low-skilled men (although as highlighted by Card et al. (2018), there has been a remarkable rise in the share of unionised jobs held by women), and so it is worthwhile to consider both whether estimates of the impact of declining union outside options affects different groups of workers similarly, and what the implications of this decline are, given the marked difference in the prevalence of union jobs across subgroups in 1980 and the different job opportunities available to these groups. One limitation of estimating effects separately by subpopulations is in the number of $i - c$ cells available to identify effects. To increase the size of cells we estimate our effects at the State level, which ensures a greater number of cells are well populated. This restriction is not sufficient to allow us to separately identify the effects of nonunion rents and union outside options. To obtain estimates with sufficient precision on which to base counterfactual experiments we impose the restriction that the coefficient on nonunion rents and union outside options is equal and estimate a single coefficient on the sum of the two terms. An imposition of this restriction is somewhat supported by the estimates of our estimates obtained for the entire sample. We employ the same set of instruments as before constructed separately for each subsample. To aid comparison to our main OLS and IV results we present estimates of the pooled outside option term using city and state level variation respectively in rows 1 and 2 of Table 3.4. Unsurprisingly the coefficient estimates fall between the coefficient estimate on nonunion rents and union outside options. Results obtained using states as our geographic measure are slightly higher than when using variation across cities.

Comparing Men and Women

We first obtain results mirroring those in Tables 3.1 and 3.2 for men and women separately. These results are presented in rows 3 and 4 of Table 3.4, with OLS results indicating little difference between the coefficient estimate on outside options between subgroups. IV results however indicate that the return to outside options is larger for women, though this difference is not statistically significant. Though we cannot rule out the hypothesis that coefficient estimates are the same, it is worth mentioning that the main explanation for different coefficient estimates embedded in our model is differences in bargaining power. This could vary with the nature of the work conducted by the subgroup.

In Table 3.5 we decompose the nonunion wage to estimate the impact of union wage spillovers on earnings. We report separately the contribution of the various components driving variation

in union outside options, and the total change. The results of this exercise for men and women are presented in columns 1 and 2. Wages over 1980-2010 declined by 0.016-0.014 percentage points as a result of worsening union outside options. The effect of declining unionisation is felt similarly by both subgroups although we estimate a coefficient estimate of twice the size for women. It is worth noting that if this result is due to noise in estimates, and the true underlying coefficient is equal between the two groups, the estimated impact of declining union outside options for men would be much larger than that of women. For males union wage spillovers explain 10.6% of wage declines over this period, but for women they explain 17.8%. This results from a steeper decline in wages for men over the sample period. For men the impact of a declining transition rate is larger than for women, and for both subgroups this effect is partially offset by increases in the union wage premia over 1980-1990. For women this offsetting effect disappears by 2010, though it continues to operate for men. This variation in premia serves to somewhat equalise the overall impact of declining union outside options on wages between men and women.

Comparing the Young and the Old

In rows 5 and 6 of Table 3.4 we present our estimates comparing workers aged between 20 and 35, and those aged between 36 and 55. Our IV estimates indicate very similar coefficient estimates on outside options between the two subgroups. In Table 3.5 we perform the same decomposition as for men and women, finding that the impact of declining unionisation lowered wages for younger workers to a greater extent than for older workers. In part this is explained by a much greater decline in the transition probability between 1980 and 1990. For both sets of workers increasing union premia served to offset declining transitions, although for older workers this effect served to largely immunise workers against declining outside options between 1980-1990. For younger workers this offset is only partial and wages still declined 0.014 log points from 1980-1990 due to declining outside options. For both sets of workers declining union premia between 1990-2010 offset the positive impact of the initial increase, and ultimately, variation in union premia tends to lower wages over the period.

Comparing Education Classes

We now turn to comparing the effects of declining unionisation on wages operating through outside options for less educated workers with a high school education or less, and for more highly educated workers with more than a high school education. Our regression results in rows 7 and 8 of Table 3.4 provide some evidence that the coefficient on outside options is larger for the more highly educated grouping, although the difference is not significant in our IV estimates due to increased standard errors. The results of our decomposition exercise presented

Table 3.5: Outside Options Contribution to Changing Wages - Subsample Analysis

	Male	Female	Age 20-35	Age 36-55	$\leq HS$	$> HS$
			1980-1990			
Raw	-0.182	-0.114	-0.164	-0.170	-0.166	-0.144
Change in T	-0.012	-0.008	-0.024	-0.013	-0.016	-0.007
Change in η^u	0.000	0.001	0.000	0.001	0.001	0.007
Change in $v^u - v^n$	0.006	0.005	0.010	0.009	0.009	0.005
Total	-0.006	-0.002	-0.014	-0.004	-0.006	0.005
Total/Raw	0.034	0.016	0.084	0.022	0.035	-0.038
			1980-2010			
Raw	-0.154	-0.076	-0.161	-0.184	-0.193	-0.115
Change in T	-0.020	-0.011	-0.033	-0.024	-0.028	-0.014
Change in η^u	0.001	0.000	-0.002	-0.001	0.001	0.010
Change in $v^u - v^n$	0.003	-0.002	0.004	0.005	0.008	0.003
Total	-0.016	-0.014	-0.030	-0.021	-0.019	-0.002
Total/Raw	0.106	0.178	0.187	0.111	0.100	0.013

Notes We fully decompose average log hourly nonunion wages in stages using our estimated wage equation. We use IV estimates presented in Table 3.4 to calculate the contribution of changing components of union outside options. 'Raw' is the actual movement in $\hat{w}_{ic}^n$ over the sample period. In turn, we fix the transition rate into the union sector (T), union sector industrial composition, and union wage premia, to calculate their effect on wages over the sample periods considered. 'Total' is the total change in nonunion wages coming from changes in union outside options. 'Total/Raw' is the ratio of the union outside option contribution to the complete wage change.

in Table 3.5 are more stark than for previous subsample breakdowns. In particular, declining transitions into the union sector served to lower the wages of low educated workers in the nonunion sector by .028 log points, while the corresponding effect for more highly educated workers is just .014 log points over the entire sample period. This result is observed despite the increased importance of union outside options in driving wages, hence reflecting very different rates of decline in the union contribution of outside options between subgroups.

For less educated workers, fixing the composition of union work has little impact on earnings, but remarkably, for highly educated workers, fixing the industrial composition serves to offset the impact of declining transitions into the union sector. That is, over 1980-2010, the composition of union work performed by more highly educated workers tended to shift towards high paying industries. Absent other changes this would increase wages, and over the sample period this serves to largely offset the impact of declining union job availability. As in our results for the entire population, the same pattern of increasing wage premium over 1980-1990 followed by a decline is observed. For highly educated workers this subsequent decline offsets the initial increase while for less educated workers this offset is only partial and serves to lower wages. In general, changing outside options led to large declines in the wages of less educated workers, but had no real effect on the earnings of more highly educated workers. This is in large part due to the shifting of union work towards high paying industries for this subsample of workers.

Comparing Young and Old Men

Given the historic concentration of unionised jobs among male workers we now turn to focuses on the wages of men between 1980-2010. We first compare young and old men with our regression results in rows 9 and 10 of Table 3.4 finding some evidence that the impact of outside options on wages is larger for younger workers, although this difference is observed only in our OLS estimates. The decomposition of wages presented in columns 1 and 2 of Table 3.6 is very similar to that presented in columns 3 and 4 of Table 3.5 for the entire population, finding a larger impact of worsening union outside options on younger workers. We find that between 1980-1990 increasing union premia almost completely offsets the negative effect of declining transitions into union jobs for older workers. There is no subsequent decline in the union premia for older workers, while there is some evidence of a decline for younger workers between 1990 and 2000. Overall the nonunion wages of younger male workers would be .027 log points higher in 2010 had outside options been static, whereas, the wages of older workers would be just 0.014 log points higher. The impact of outside options on wages is therefore twice as large for younger male workers.

Table 3.6: Outside Options Contribution to Changing Wages - Males Only

	Age 20-35	Age 36-55	$\leq HS$	$> HS$
		1980-1990		
Raw	-0.224	-0.156	-0.170	-0.132
Change in T	-0.024	-0.010	-0.022	-0.002
Change in η^u	0.003	0.001	0.002	0.008
Change in $v^u - v^n$	0.008	0.008	0.011	0.002
Total	-0.013	-0.001	-0.009	0.008
Total/Raw	0.057	0.004	0.051	-0.060
		1980-1990		
Raw	-0.127	-0.170	-0.180	-0.114
Change in T	-0.035	-0.026	-0.040	-0.008
Change in η^u	0.002	0.003	0.004	0.013
Change in $v^u - v^n$	0.006	0.009	0.010	0.006
Total	-0.027	-0.014	-0.026	0.010
Total/Raw	0.215	0.085	0.143	-0.087

Notes We fully decompose average log hourly nonunion wages in stages using our estimated wage equation, focusing on male subgroups by education and age. We use IV estimates presented in Table 3.4 to calculate the contribution of changing components of union outside options. 'Raw' is the actual movement in $\hat{w}_{ic}^n$ over the sample period. In turn, we fix the transition rate into the union sector (T), union sector industrial composition, and union wage premia, to calculate their effect on wages over the sample periods considered. 'Total' is the total change in nonunion wages coming from changes in union outside options. 'Total/Raw' is the ratio of the union outside option contribution to the complete wage change.

Comparing Men by Education Class

We now turn to comparing male workers by education. Rows 11 and 12 of Table 3.4 present our estimated results for these subgroups indicating that the coefficient on outside options for more highly educated workers is larger than for the less educated and this difference is statistically significant in both OLS and IV estimates. This may reflect higher bargaining power on the part of more highly educated workers. The results of our decomposition exercise in the last two columns of Table 3.6 are similar, but more stark than those in the last two columns of Table 3.5. Specifically, the overall impact of worsening outside options on less educated workers is larger than before at 0.026 log points. This reflects a decline in the transition probability which, all else fixed, would lower wages by .04 log points. Increased union premia between 1980-1990 are responsible for attenuating the negative effect of outside options on wages.

For highly educated workers, remarkably, this exercise indicates that fixing outside options serves to lower wages. Outside options for highly educated males served to increase wages over the sample period, in marked contrast to less educated workers. Specifically, had union outside options been fixed, wages would have been .01 log points lower for this group. This is explained by the relatively small decline in transitions into the union sector for this subgroup, and the substantive impact of industrial composition on the average earnings of workers in the union sector. In particular, for these workers, over 1980-2010 union work concentrated more heavily in industries paying a relatively high wage. This effect is large and outweighs the impact of declining transitions into high paying union jobs. In contrast to less educated workers, there is only a small increase in union premia between 1980-1990 although highly educated workers see a notable increase between 1990-2010.

Overall, our results indicate that highly educated male workers wages increased over the sample period as a result of improved outside options. For this subgroup, the impact of declining unionisation was offset by a shift in the composition of union work towards relatively high paying sectors, and to a lesser extent by increasing union wage premia. For less educated workers, increased union premia partially offsets the impact of declining unionisation, but these workers still observe a substantive decline in wages of .026 log points between 1980-2010.

3.4 The Firm Response to the Union Threat

So far we have considered firms to be passive players in the process of union certification and our results reflect this restriction. We now move to a more realistic setting in which firms at risk of being unionised can respond to forestall unionisation. In TD, firms respond by hiring more skilled workers who they know will vote against unionisation. While this response is possible, it seems to us to likely be of second order importance relative to more direct re-

sponses. In real world descriptions of firm reactions to the threat of unionisation[30], firms adopt some combination of three possible responses: 1) paying higher wages to reduce the net benefit of unionising; 2) providing better working conditions to match what workers would get in a union setting (in the context of our model where only union workers get the amenities, ψ_{icf}, nonunion firms would provide similar amenities to their workers); and 3) intimidation (which, in our model, would correspond to increasing the cost of unionisation, λ_c). In this section, we consider the implications of each of these possible responses for nonunion wage determination and firm selection into union versus nonunion status.

Before discussing each response separately, note that firms do not need to consider employing any response if their workers are happily nonunion. That is, if the costs of unionising, the amenities that would be available to the workers at this firm if they unionise, the union wage they would get if they organise, and the nonunion wage they get if they don't are such that $I_{icf} < 0$ then there is no reason for the firm to bear costs to incentivize its workers not to form a union. We are interested in the set of firms that are unionised (i.e., for which $I_{icf} > 0$) but near the margin of being so in our first, non-responsive firm model. Recall that the fixed cost of forming a union depends on the legal climate in each state while the non-wage benefits of unionising are firm specific. As a result, not all firms in the same industry and city will have the same worker preferences about unionisation.

3.4.1 The Wage Response

We first consider the possibility that firms respond to their worker's desire to unionise by offering the workers a wage that just offsets the benefit of unionising. In particular, based on the discussion of worker preferences on whether to unionise underlying the index function (3.23), the wage the firm would need to offer to make workers indifferent about forming a union is:

$$w^*_{icf} = w^u_{icf} - \lambda_c + \psi_{icf} \tag{3.39}$$

Now consider worker preferences about unionising when they consider the non-emulation nonunion wage we derived earlier (w^n_{icf}), the cost of unionising and the union wage. In this situation, we can define a $\psi^*_{icf} = \lambda_c - (w^u_{icf} - w^n_{icf})$ - the value for amenities such that workers at this firm are indifferent between whether they organise or not. For $\psi \leq \psi^*_{icf}$, workers will not organise, and the firm will pay the regular nonunion wage, w^n_{icf}. For $\psi > \psi^*$, workers will prefer to unionise if offered the regular nonunion wage and firms will consider the value

[30] See https://www.theguardian.com/film/2018/aug/23/pay-a-living-wage-bernie-sanders-accuses-disney-of-dodging-fair-pay, https://www.theguardian.com/sustainable-business/target-anti-union-video-cheesy-but-effective, and https://www.theguardian.com/us-news/2019/jun/12/delta-workers-pro-union-report-threats-management

of offering the emulation wage, w^*_{icf}, instead and forestalling unionisation. We show in the appendix that the firm will be willing to pay the higher, emulation wage until the point where the costs and benefits to unionisation overlap, which happens when $\psi_{icf} = \lambda_c$. At that point, the emulation wage equals the union wage. Beyond it, the wage required to prevent union formation exceeds the union wage, and to fight the union would lower firm profits. For all $\psi_{icf} > \lambda_c$, then, the firm will not fight the union and will become unionised. Thus, we can characterize the firm union status as follows:

- $\psi_{icf} < \psi^*_{icf}$: f is nonunion and pays w^n_{icf}
- $\psi^*_{icf} \leq \psi_{icf} < \lambda$: f is nonunion but emulates unionised firms and pays w^*_{icf}
- $\psi_{icf} \geq \lambda$: f is union and pays w^u_{icf}

Note that both cut-offs rise with λ_c and, so, states that implement policies that raise the cost of unionising will have more nonunion firms. As the unionising costs rise, the actual nonunion wage will not change. However, the observed nonunion wage will decline because the fraction of nonunion workers who are paid the higher, emulation wage, will decline.[31] Moreover, the emulation wage w^*_{icf}, will also decline. Thus, we would observe a decline in unionisation combined with an increase in the observed union wage differential. This pattern might seem to imply that de-unionisation is arising because of union rigidity on wages in the face of declining demand in sectors. But this can happen even without unions being rigid (i.e., even though union wages will decline with declines in ϵ). The pattern of declining unionisation with an increasing union wage differential is what we observe during the 1980s.

Based on this discussion, de-unionisation has two effects on nonunion wages. The first is the bargaining effect: the outside option for nonunion workers captured in our rent variable declines because individual workers cannot point to the possibility of getting a high paying union job if they were to break off bargaining. The second is the emulation effect: as the threat of being unionised becomes less prominent both the proportion of firms that pay an emulation wage declines and the emulation wage needed at any firm paying it also declines.

To obtain an empirical specification related to the model including union emulation, first note that the log-linearized version of the union emulation wage is given by,

$$w^*_{icf} = \gamma^u_0 + \gamma^u_1 \tilde{E}^n_{ict} + \gamma^u_2 ER_c + (1-\gamma^u_3)\psi_{icf} - \lambda_{ct} + \gamma^u_4 \epsilon_{ic} \tag{3.40}$$

[31] To see this note that the cut-off for the lower end of the emulation range, $\psi_u = \lambda_{ct} - (w^u - w^n)$, will rise faster than the upper end (which is λ_{ct}) as λ increases. This can be shown by noting that $\frac{\partial \psi_u}{\partial \lambda_{ct}} = 1 - \frac{\partial w^u}{\partial \lambda_{ct}}$ and that $\frac{\partial w^u}{\partial \lambda_{ct}} < 0$ since union wages are lower when union related amenities are higher.

There is one additional complication that we have yet to address, which is that for certain values of ϵ there may be no range over which workers wish to unionise. In terms of the thresholds defined above this occurs when $\psi^* = \lambda_c - (w^u - w^n) \geq \lambda_c$, which occurs when the nonunion wage exceeds the union wage. In this instance, for all values of ψ, workers will prefer to remain nonunionised. We define the productivity draw that equalises the costs and benefits of unionisation as ϵ^*. For productivity draws above this threshold the union wage will exceed the nonunion wage and there will be some set of ψ value over which unionisation is preferred.

The observed mean log nonunion wage in an $i-c$ cell is given by a weighted average of the actual nonunion wage, and the emulation wage. We define this observed wage as $\bar{w}^n_{icf}$:

$$\begin{aligned} E(\bar{w}^n_{ic}) = & \frac{Pr(I_{icf} < 0)}{Pr(I_{icf} < 0) + Pr(I_{icf} > 0, \psi_{icf} < \lambda_c))} E(w^n_{icf} | I_{icf} < 0) \\ & + \frac{Pr(I_{icf} > 0, \psi_{icf} < \lambda_c)}{Pr(I_{icf} < 0) + Pr(I_{icf} > 0, \psi_{icf} < \lambda_c))} E(w^*_{icf} | I_{icf} > 0, \psi_{icf} < \lambda_c) \end{aligned} \tag{3.41}$$

The weights are the probability of firms being of each type conditional on being observed as a nonunion firm. To form this regression equation, we first need to specify the probabilities that make up the weights:

$$Pr(I_{icf} < 0) = \int_0^\infty \int_{-\infty}^{\frac{\lambda_c - \Delta - \alpha_4 \epsilon}{(1-\gamma_3)}} f(\psi) g(\epsilon) d\psi d\epsilon \tag{3.42}$$

and,

$$Pr(I_{icf} > 0, \psi_{icf} < \lambda_c) = \int_{\epsilon *}^\infty \int_{\frac{\lambda_c - \Delta_{ic} - \alpha_4 \epsilon}{(1-\gamma_3)}}^{\lambda_c} f(\psi) g(\epsilon) d\psi d\epsilon \tag{3.43}$$

where, the first probability is the probability that workers would choose to be nonunion and the second is the probability that firms are nonunion but emulate union wage setting, $\Delta = \alpha_{0i} + \gamma^u_1 E^u_{ic}(w) - \gamma^n_1 E^n_{ic}(w) + \alpha_2 ER_c$, the union benefit threshold is $\psi^* = \frac{\lambda_c - \Delta - \alpha_4 \epsilon}{(1-\gamma_3)}$, and recalling that the range for the productivity shock, ϵ_{ic} is $[0, \infty]$.

The conditional mean wage expressions are given by:

$$E(w^n_{icf} | I_{icf} \leq 0) = \gamma^n_{0i} + \gamma^n_1 R^n_c + \gamma^n_1 E^{nd}_{ic} + \gamma^n_2 ER_{ic} + \gamma^n_4 E(\epsilon_{ic} | I_{icf} \leq 0) \tag{3.44}$$

and,

$$\begin{aligned} E(w^*_{icf} | I_{icf} > 0, \psi_{icf} < \lambda_c) = & \gamma^u_0 + \gamma^u_1 R^n_c + \gamma^u_1 E^{ud}_{ic} + \gamma^u_2 ER_c - \lambda_c \\ & + E(\gamma^u_4 \epsilon_{ic} + (1 - \gamma^u_3) \psi_{icf} | I_{icf} > 0, \psi_{icf} < \lambda_c) \end{aligned} \tag{3.45}$$

We discuss the two approaches we take to estimation of this wage response model below.

3.4.2 The Amenity Response

Firms could respond to the threat of unionisation through means other than raising wages. The first possibility is that they respond by increasing amenities for the workers. Since this is one of the things workers get out of a union, a direct response of this type seems possible. In our specification, worker utility on a union job is a linear function of the value of union amenities available to him if the firm is unionized with the value being expressed in dollar equivalents. In order for firms to want to provide amenities rather than wages as a means of resisting unionisation, the cost of providing the amenity must be less than or equal to the dollar valuation that the worker gets from the amenity. Otherwise, wages (increasing which costs the firm a dollar and gives the worker a dollar in value) will be a more cost-effective response. Assume, in particular, that providing amenities has an increasing and convex marginal cost function with the marginal cost of the initial units provided being below a dollar for one dollar worth of amenities as valued by the worker. In that case, nonunion firms would want to use amenities as a response to a union threat until the point where the marginal cost of a dollar's worth of amenities rises to one dollar. After that, they would respond through wage emulation.

However, if it is cost effective for a nonunion firm to use amenities to respond to a union threat, it would also be cost effective for it to pay in amenities instead of wages even in the absence of such a threat. Thus, if a threat emerges, the nonunion firm will already be providing amenities up to the point where their marginal cost equals a dollar. In that case, there is no room for the firm to respond to a union threat using amenities. Instead, it will respond through wage emulation.

3.4.3 The Intimidation response

As a third potential response we allow firms to increase the costs to unionise for workers. Recall that workers in city c face a fixed cost of unionising, λ_c. Firms can increase that cost at a cost to themselves. For example, they could lock out the workers and either not produce or hire scabs who are less productive than the actual workers. The firm could also take legal action to delay the union vote, imposing more costs on the workers. In the appendix, we set out the value function for a firm that employs intimidation and compare it to the value functions when the firm chooses the wage emulation response. We show that, because of differences in timing between when the firm carries out the intimidation and when it hires workers, intimidation is a less efficient response for some values of ψ. We show that under reasonable assumptions there is a new cut-off value, ψ^b_{icf} such that, $\psi^*_{icf} < \psi^b_{icf} < \lambda$. Firms with $\psi^*_{icf} < \psi_{icf} \leq \psi^b_{icf}$

will use intimidation and, as a result, will remain nonunion and pay the nonunion wage. Firms with $\psi^b_{icf} < \psi_{icf} \leq \lambda$ will be nonunion but pay the emulation wage. As before, firms with ψ_{icf} below ψ^*_{icf} will be nonunion, paying the nonunion wage, and firms with ψ_{icf} above λ will be unionised. Thus, introducing intimidation serves to expand the region over which unions are nonunion and pay the simple nonunion wage to include the range over which firms use intimidation. We do not have a way to separately identify ψ^b_{icf} from ψ^*_{icf} and so cannot distinguish between a model in which there is no intimidation and the relevant cut-off for the nonunion region is ψ^*_{icf} and one in which there is intimidation and the relevant cut-off is ψ^b_{icf}. Since both cut-offs are functions of the same variables - λ, the expected rents, and the employment rate - there is essentially no impact on our empirical specifications of including or not including intimidation. We will proceed as if there is no intimidation in order to simplify the exposition.

3.4.4 Regression Analysis

We turn now to estimating the conditional mean wage expression (3.41) in order to investigate whether incorporating firm emulation responses affects our earlier conclusions about the impacts of unions on nonunion wages. We linearize (3.41) around the main driving forces from the model: ΔR^n_c, changes in average nonunion rents; ΔE^{nd}_{ic}, changes in the difference between union and nonunion rents weighted by the probability a nonunion worker finds a union job; ΔE^{ud}_{ic}, changes in the difference between union and nonunion rents weighted by the probability a union worker's next job is a union job; ER_c, the employment rate; and λ_c, costs of unionising. ΔE^{ud}_{ic} is a new term compared to our earlier expression when we did not consider firm responses to unionisation threats. It is the most direct reflection of the inclusion of the emulation channel. Recall that λ_c affects the observed mean nonunion wage both because it affects the probability that a firm is an emulator and the emulation wage, and both effects work in the same direction: the larger is λ_c, the lower is the emulation wage (because firms don't need to pay as much to keep workers from unionising when the cost of unionising is higher) and, as discussed earlier, as λ_c increases, the probability of a firm being a union emulator decreases. Note that in contrast to the case in which firms do not respond to unionisation threats, λ_c is a direct determinant of the mean observed nonunion wage and so unionisation costs are not available to identify selection effects as they were before. We either need to assume that there is no selection effect or that the coefficients on our proxies for λ_c are a combination of the direct cost effects and the selection effects.

The rent terms also affect both the underlying mean wages conditional on being either a simple nonunion firm or being a union emulator and the weights applied to each in the overall mean. Their predicted signs in the linearized regression depend on the relative size of rent effects on simple nonunion wages and on the union wages that some firms are emulating. Both

R_c and E_{ic}^{nd} have positive predicted effects on the underlying conditional mean wages but uncertain or negative effects on the probability of a firm being a union emulator. For that reason, their ultimate impact on the overall nonunion mean wage is uncertain. On the other hand, improvements in the relative value of union sector rents for union workers, E_{ic}^{ud}, both increases emulation wages and the weight on those wages, so its predicted effect is positive.

Note that in comparison to the specification derived from the simpler model, we now have a new term, E_{ic}^{u}, the outside option a worker could expect to access if he was unionised. Apart from this, the specification is the same as the one from the simpler model - it includes as right hand side variables, E_{ic}^{n}, ER_c and then other, non-linear terms involving λ_c. If the E_{ic}^{u} term enters significantly, that would be evidence in favour of the emulation based model being relevant.

Results from this specification are presented in Table 3.7. Note that the included variables are the same as in our simpler specification when we did not allow for firm responses apart from the inclusion of the union worker's outside option term, E_{ic}^{ud}, and the fact that the proxies for the cost of unionising bear a double interpretation. Compared to the earlier estimates from the simpler model, the coefficient on ΔR_c is about 50% larger. In the context of the model, this would arise because that rent has positive effects on both the pure nonunion wage and the emulation wage. It must be taking a higher value because of the added control for the union worker's outside option, which the more complete theory indicates belongs in the regression. In contrast, in the IV estimation, the coefficient on the nonunion worker's value of the relative outside options in the union and nonunion sectors is one-third its size in the simpler equation. This is offset by a significant effect of the union worker's outside option value. Interpreting strictly from the model, the implication is that what we thought was a union related bargaining effect, showing up through E_{ic}^{nd} was really capturing the emulation effect, represented by the union worker's outside option variable. Put another way, nonunion workers cannot use union job options as a means of bargaining higher wages, but union workers can, and to the degree this increases union wages it will affect nonunion wages through the emulation channel associated with the firm responding to the union threat.

Taken together, the estimates imply that direct bargaining effects related to shifts in outside options have strong, positive effects on wage setting as evidenced by the coefficient on ΔR_c. But the bargaining channel is not the main way in which unions affect nonunion wages. Instead, they have their effect mainly through getting firms to pay higher wages in order to emulate unionised workplaces, keeping unions out.

Table 3.7: Including Wage Emulation Effects

	OLS	IV	OLS	IV
	(1)	(2)	(3)	(4)
ΔR_c^n	2.55***	2.25***	2.52***	2.25***
	(0.33)	(0.57)	(0.33)	(0.55)
ΔE_{ic}^n	0.46	0.87	0.50	0.87
	(0.57)	(0.88)	(0.56)	(0.88)
ΔE_{ic}^u	1.36***	1.00**	1.33***	1.08**
	(0.28)	(0.44)	(0.27)	(0.43)
ΔER	0.81***	0.92***	0.82***	0.90***
	(0.19)	(0.22)	(0.19)	(0.22)
$\Delta \frac{EL_c}{N_c}$			-0.069	-0.11
			(0.61)	(0.61)
$\Delta \frac{EL_c^{win}}{EL_c}$			-0.011	-0.0098
			(0.020)	(0.020)
$\Delta \frac{EL_c}{N_c} \times \Delta \frac{EL_c^{win}}{EL_c}$			-5.07*	-5.09*
			(2.95)	(3.03)
Observations	6569	6569	6549	6549
R^2	0.50	0.50	0.50	0.50
Year × Ind.	Yes	Yes	Yes	Yes
ERIV		No		No
IVs		all		all
F-Stats:				
ΔR_c^n		36.85		30.92
ΔE_{ic}^n		139.36		112.79
ΔE_{ic}^u		36.70		29.32
p-val:				
ΔR_c^n		0.00		0.00
ΔE_{ic}^n		0.00		0.00
ΔE_{ic}^u		0.00		0.00

Notes Standard errors in parentheses clustered at the city-year level. * denotes significance at the 10% level, ** denotes significance at the 5% level, *** denotes significance at the 1% level. The dependent variable is the change in the regression adjusted average hourly wage of nonunion workers in an industry-city cell. Estimates obtained using decadal changes over 1980-2010 across 50 industries and 93 cities.

3.5 Conclusion

In this paper we provide new estimates of union wage spillovers operating through a novel channel. By formalising job finding, wage setting, and union formation we derive a specification grounded in theory, which makes clear the channel through which equilibrium effects are affecting wages. Moreover, our model informs our identification strategy by making clear what is included in our wage equation error term. In the extended version of our model, we endogenise union threat effects, finding in our regression estimates that threat effects dominate bargaining effects. However, to fully untangle these two channels and uncover the true effects of bargaining on wage setting, we need to estimate the model structurally. We will turn to structural estimation in future work.

Our results indicate an important role played by union wage spillovers in lowering wages over the 1980-2010 period. Much of this decline is due to the reduced ability of workers to find union jobs. The industrial composition of work plays an important role, and somewhat counterbalances the negative impact of declining union job opportunities. For the entire population we find that had union outside options been fixed over 1980-2010, wages would have been 2% points higher. We also find substantial heterogeneity in our results across subgroups of workers. In particular, we find that more highly educated workers are more than able to insulate themselves against declining unionisation, while the same is not true for less educated workers.

Conclusion

This dissertation studies important topics in the economics of local labour markets: how much wages vary across regions, why these wages vary, and how best to estimate these patterns and effects.

In Chapter 1, I outline and present evidence in support of an improved methodology to correct for selection bias. This method is generalizable to myriad empirical settings and has wide applicability beyond the local labour market literature. I show how a Post-Double-Lasso selection procedure can be used to reliably select key terms capturing selectivity from amongst a high dimensional set. This model selection method overcomes the dimensionality problem inherent in these settings, and allows for more flexible selection correction relative to more traditional methods. Numerical evidence confirms the efficacy of this method, showing that across a wide range of reasonable settings, bias is reduced by around 75-99%.

The benefits to this method are further explored in an empirical setting in Chapter 2. In this chapter I leverage the improved performance of this methodology to derive improved estimates of the returns to education across States using the 1990 US Census. I find that estimates obtained using this novel procedure differ significantly from more traditional estimates. Though I confirm the general upward bias in the returns to a college education in OLS estimates, my results suggest that traditional estimates overstate the degree of upward bias in corrected estimates. Conversely, my results suggest bias is understated in estimates of the return to an advanced degree, or the penalty associated with a less than high school education.

In Chapter 3 I present evidence in favour of union wage spillovers operating at the city level. In this chapter I examine a novel channel of spillovers operating through wage bargaining whereby workers with improved job prospects can negotiate higher wages with their employer. I build a novel model which incorporates wage bargaining, union formation, endogenises the union wage premium, and allows for firm responses to the union threat. Using the unique insights generated by this framework I estimate the response of non-union wages to improved job opportunities in the union sector using an instrumental variables strategy, which employs Bartik style shift-share instruments. Averaging the effect of wage spillovers across cities I estimate that average hourly wages would be 2 percentage points higher in 2010 had the size, and

composition, of the union sector been fixed over the sample period 1980-2010. Moreover, I find substantial heterogeneity across subgroups of workers in the impact of declining unionisation on nonunion wages over this period.

Bibliography

Abowd, John M and Henry S Farber (1982) "Job queues and the union status of workers," *ILR Review*, Vol. 35, pp. 354–367.

Abraham, Katharine G and Henry S Farber (1988) "Returns to seniority in union and nonunion jobs: a new look at the evidence," *ILR Review*, Vol. 42, pp. 3–19.

Abraham, Katharine G and James L Medoff (1984) "Length of service and layoffs in union and nonunion work groups," *ILR Review*, Vol. 38, pp. 87–97.

——— (1985) "Length of service and promotions in union and nonunion work groups," *ILR Review*, Vol. 38, pp. 408–420.

Acemoglu, Daron and William B Hawkins (2014) "Search with multi-worker firms," *Theoretical Economics*, Vol. 9, pp. 583–628.

Açıkgöz, Ömer Tuğrul and Barış Kaymak (2014) "The rising skill premium and deunionization," *Journal of Monetary Economics*, Vol. 63, pp. 37–50.

Addison, John T and Barry T Hirsch (1989) "Union effects on productivity, profits, and growth: Has the long run arrived?" *Journal of Labor Economics*, Vol. 7, pp. 72–105.

Ahn, Hyungtaik and James L Powell (1993) "Semiparametric estimation of censored selection models with a nonparametric selection mechanism," *Journal of Econometrics*, Vol. 58, pp. 3–29.

Athey, Susan (2017) "The impact of machine learning on economics," in *Economics of Artificial Intelligence*: University of Chicago Press.

Athey, Susan and Guido W Imbens (2015) "Machine learning methods for estimating heterogeneous causal effects," *stat*, Vol. 1050.

Athey, Susan, Guido W Imbens, and Stefan Wager (2018) "Approximate residual balancing: debiased inference of average treatment effects in high dimensions," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 80, pp. 597–623.

Bayer, Patrick, Nathaniel Keohane, and Christopher Timmins (2009) "Migration and hedonic valuation: The case of air quality," *Journal of Environmental Economics and Management*, Vol. 58, pp. 1–14.

Beaudry, Paul, David A Green, and Benjamin Sand (2012) "Does industrial composition matter for wages? A test of search and bargaining theory," *Econometrica*, Vol. 80, pp. 1063–1104.

Beaudry, Paul, David A Green, and Benjamin M Sand (2014) "Spatial equilibrium with unemployment and wage bargaining: Theory and estimation," *Journal of Urban Economics*, Vol. 79, pp. 2–19.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen (2012) "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica*, Vol. 80, pp. 2369–2429.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen (2014) "Inference on treatment effects after selection among high-dimensional controls," *The Review of Economic Studies*, Vol. 81, pp. 608–650.

Belloni, Alexandre, Victor Chernozhukov, and Lie Wang (2011) "Square-root lasso: pivotal recovery of sparse signals via conic programming," *Biometrika*, Vol. 98, pp. 791–806.

Bertoli, Simone, J Fernández-Huertas Moraga, and Francesc Ortega (2013) "Crossing the border: Self-selection, earnings and individual migration decisions," *Journal of Development Economics*, Vol. 101, pp. 75–91.

Björkegren, Daniel and Darrell Grissen (2017) "Behavior revealed in mobile phone usage predicts loan repayment," *Working Paper*.

Bombardini, Matilde, Giovanni Gallipoli, and Germán Pupato (2012) "Skill dispersion and trade flows," *American Economic Review*, Vol. 102, pp. 2327–48.

Borjas, George J, Stephen G Bronars, and Stephen J Trejo (1992) "Self-selection and internal migration in the United States," *Journal of Urban Economics*, Vol. 32, pp. 159–185.

Borjas, GJ (1987) "Self-selection and the earnings of immigrants.," *The American Economic Review*, Vol. 77, pp. 531–53.

Boström, Henrik (2008) "Calibrating random forests," in *2008 Seventh International Conference on Machine Learning and Applications*, pp. 121–126, IEEE.

Bourguignon, François, Martin Fournier, and Marc Gurgand (2007) "Selection bias corrections based on the multinomial logit model: Monte Carlo comparisons," *Journal of Economic Surveys*, Vol. 21, pp. 174–205.

Breiman, Leo (2001) "Random forests," *Machine Learning*, Vol. 45, pp. 5–32.

Card, David (2001) "The effect of unions on wage inequality in the US labor market," *ILR Review*, Vol. 54, pp. 296–315.

Card, David, Thomas Lemieux, and W Craig Riddell (2004) "Unions and wage inequality," *Journal of Labor Research*, Vol. 25, pp. 519–559.

——— (2018) "Unions and Wage Inequality: The Roles of Gender, Skill and Public Sector Employment," NBER working paper w25313.

Carneiro, Pedro and Sokbae Lee (2011) "Trends in quality-adjusted skill premia in the United States, 1960-2000," *American Economic Review*, Vol. 101, pp. 2309–49.

Chichignoud, Michaël, Johannes Lederer, and Martin J Wainwright (2016) "A practical scheme and fast algorithm to tune the lasso with optimality guarantees," *The Journal of Machine Learning Research*, Vol. 17, pp. 8162–8181.

Dahl, Gordon B (2002) "Mobility and the return to education: Testing a Roy model with multiple markets," *Econometrica*, Vol. 70, pp. 2367–2420.

Davies, Paul S, Michael J Greenwood, and Haizheng Li (2001) "A conditional logit approach to US state-to-state migration," *Journal of Regional Science*, Vol. 41, pp. 337–360.

Denil, Misha, David Matheson, and Nando De Freitas (2014) "Narrowing the gap: Random forests in theory and in practice," in *International conference on machine learning*, pp. 665–673.

dHaultfoeuille, Xavier and Arnaud Maurel (2013) "Inference on an extended Roy model, with an application to schooling decisions in France," *Journal of Econometrics*, Vol. 174, pp. 95–106.

Diamond, Rebecca (2016) "The Determinants and Welfare Implications of US Workers' Diverging Location Choices by Skill: 1980-2000," *American Economic Review*, Vol. 106, pp. 479–524.

Dickens, William and Lawrence F Katz (1986) "Interindustry wage differences and industry characteristics."

DiNardo, John, Nicole M Fortin, and Thomas Lemieux (1996) "Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach," *Econometrica*, Vol. 64, pp. 1001–1044.

DiNardo, John and Thomas Lemieux (1997) "Diverging male wage inequality in the united states and ganada, 1981–1988: Do institutions explain the difference?" *ILR Review*, Vol. 50, pp. 629–651.

Dolton, Peter J, Gerald H Makepeace, and Wilbert Van der Klaauw (1989) "Occupational choice and earnings determination: The role of sample selection and non-pecuniary factors," *Oxford Economic Papers*, Vol. 41, pp. 573–594.

Donsimoni, Marie-Paule (1981) "Union power and the American labour movement," *Applied Economics*, Vol. 13, pp. 449–464.

Dubin, Jeffrey A and Daniel L McFadden (1984) "An econometric analysis of residential electric appliance holdings and consumption," *Econometrica*, Vol. 52, pp. 345–362.

Eisenhauer, Philipp, James J Heckman, and Edward Vytlacil (2015) "The generalized Roy model and the cost-benefit analysis of social programs," *Journal of Political Economy*, Vol. 123, pp. 413–443.

Farber, Henry S, Daniel Herbst, Ilyana Kuziemko, and Suresh Naidu (2018) "Unions and inequality over the twentieth century: New evidence from survey data," NBER working paper w24587.

Farber, Henrys (2005) "Nonunion wage rates and the threat of unionization," *ILR Review*, Vol. 58, pp. 335–352.

Firpo, Sergio, Nicole Fortin, and Thomas Lemieux (2018) "Decomposing wage distributions using recentered influence function regressions," *Econometrics*, Vol. 6, p. 28.

Fortin, Nicole M, Thomas Lemieux, and Neil Lloyd (2019) "Labor Market Institutions and the Distribution of Wages: The role of Spillover Effects," *Working Paper*.

Freeman, Richard B (1980) "Unionism and the Dispersion of Wages," *ILR Review*, Vol. 34, pp. 3–23.

Freeman, Richard B and James L Medoff (1981) "The impact of the percentage organized on union and nonunion wages," *The Review of Economics and Statistics*, pp. 561–572.

French, Eric and Christopher Taber (2011) "Identification of models of the labor market," in *Handbook of Labor Economics*, Vol. 4: Elsevier, pp. 537–617.

Glaeser, Edward L, Scott Duke Kominers, Michael Luca, and Nikhil Naik (2018) "Big data and big cities: The promises and limitations of improved measures of urban life," *Economic Inquiry*, Vol. 56, pp. 114–137.

Goldsmith-Pinkham, Paul, Isaac Sorkin, and Henry Swift (2018) "Bartik instruments: What, when, why, and how," NBER working paper w24408.

Gosling, Amanda and Thomas Lemieux (2001) "Labour market reforms and changes in wage inequality in the United Kingdom and the United States," NBER working paper w8413.

Hastie, Trevor, Robert Tibshirani, and Martin Wainwright (2015) *Statistical learning with sparsity: the lasso and generalizations*: CRC press.

Heckman, J and C. Taber (2008) "Roy model," in Durlauf S.N. and Blume L.E. eds. *The Oxford Handbook of Innovation*: Palgrave Macmillan.

Heckman, James J (1979) "Sample Selection Bias as a Specification Error," *Econometrica*, Vol. 47, pp. 153–161.

Heckman, James J, Bo E Honore et al. (1990) "The Empirical Content of the Roy Model," *Econometrica*, Vol. 58, pp. 1121–1149.

Heckman, James J and Edward J Vytlacil (2007) "Econometric evaluation of social programs, part I: Causal models, structural models and econometric policy evaluation," *Handbook of Econometrics*, Vol. 6, pp. 4779–4874.

Hirsch, Barry T and Albert N Link (1984) "Unions, productivity, and productivity growth," *Journal of Labor Research*, Vol. 5, pp. 29–37.

Hirsch, Barry T and John L Neufeld (1987) "Nominal and real union wage differentials and the effects of industry and SMSA density: 1973-83," *The Journal of Human Resources*, Vol. 22, pp. 138–148.

Holzer, Harry J (1982) "Unions and the labor market status of white and minority youth," *ILR Review*, Vol. 35, pp. 392–405.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006) "Unbiased recursive partitioning: A conditional inference framework," *Journal of Computational and Graphical statistics*, Vol. 15, pp. 651–674.

Jaeger, David A and Marianne E Page (1996) "Degrees matter: New evidence on sheepskin effects in the returns to education," *The Review of Economics and Statistics*, pp. 733–740.

Jaeger, David A, Joakim Ruist, and Jan Stuhler (2018) "Shift-share instruments and the impact of immigration," NBER working paper w24285.

Kahn, Lawrence M (1980) "Union spillover effects on organized labor markets," *The Journal of Human Resources*, Vol. 15, pp. 87–98.

Kennan, John and James R Walker (2011) "The effect of expected income on individual migration decisions," *Econometrica*, Vol. 79, pp. 211–251.

Kirkeboen, Lars J, Edwin Leuven, and Magne Mogstad (2016) "Field of study, earnings, and self-selection," *The Quarterly Journal of Economics*, Vol. 131, pp. 1057–1111.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015) "Prediction policy problems," *American Economic Review*, Vol. 105, pp. 491–95.

Krusell, Per and Leena Rudanko (2016) "Unions in a frictional labor market," *Journal of Monetary Economics*, Vol. 80, pp. 35–50.

Lee, Lung-Fei (1983) "Generalized Econometric Models with Selectivity," *Econometrica*, Vol. 51, pp. 507–12.

Lemieux, Thomas (2006) "Increasing residual wage inequality: Composition effects, noisy data, or rising demand for skill?" *American Economic Review*, Vol. 96, pp. 461–498.

Lesch, Hagen (2004) "Trade union density in international comparison," in *CESifo Forum*, Vol. 5, pp. 12–18.

Lewis, H Gregg (1963) *Unionism and relative wages in the United States: an empirical inquiry*: University of Chicago press.

Moretti, Enrico (2013) "Real wage inequality," *American Economic Journal: Applied Economics*, Vol. 5, pp. 65–103.

Mullainathan, Sendhil and Jann Spiess (2017) "Machine learning: an applied econometric approach," *Journal of Economic Perspectives*, Vol. 31, pp. 87–106.

Murphy, Kevin M and Robert H Topel (1985) "Estimation and inference in two-step econometric models," *Journal of Business & Economic Statistics*, Vol. 3, pp. 370–379.

Murphy, Kevin M and Finis Welch (2001) "Wage Differentials in the 1990s: Is the Glass Half-full or Half-empty?" *The Causes and Consequences of Increasing Inequality*, Vol. 2.

Neumark, David and Michael L Wachter (1995) "Union effects on nonunion wages: Evidence from panel data on industries and cities," *ILR Review*, Vol. 49, pp. 20–38.

Niculescu-Mizil, Alexandru and Rich Caruana (2005) "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, ACM.

Oaxaca, Ronald (1973) "Male-female wage differentials in urban labor markets," *International Economic Review*, pp. 693–709.

Park, Jin Heum (1994) *Estimation of sheepskin effects and returns to schooling using the old and the new CPS measures of educational attainment*, No. 338: Industrial Relations Section, Princeton University.

Pearce, James E. (1990) "Tenure, Unions, and the Relationship between Employer Size and Wages," *Journal of Labor Economics*, Vol. 8, pp. 251–269, URL: `http://www.jstor.org/stable/2535098`.

Pinkse, CAP (1993) "On the computation of semiparametric estimates in limited dependent variable models," *Journal of Econometrics*, Vol. 58, pp. 185–205.

Pissarides, Christopher A (1986) "Trade unions and the efficiency of the natural rate of unemployment," *Journal of Labor Economics*, Vol. 4, pp. 582–595.

Podgursky, Michael (1986) "Unions, establishment size, and intra-industry threat effects," *ILR Review*, Vol. 39, pp. 277–284.

Ransom, Tyler (2016) "Selective Migration, Occupational Choice, and the Wage Returns to College Majors," Mimeo, Available: https://tyleransom.github.io/research/roymajors2016Dec27.pdf.

Rosen, Sherwin (1969) "Trade union power, threat effects and the extent of organization," *The Review of Economic Studies*, Vol. 36, pp. 185–196.

Rothschild, Casey and Florian Scheuer (2013) "Redistributive taxation in the Roy model," *The Quarterly Journal of Economics*, Vol. 128, pp. 623–668.

Roy, Andrew Donald (1951) "Some thoughts on the distribution of earnings," *Oxford Economic Papers*, Vol. 3, pp. 135–146.

Sanderson, Eleanor and Frank Windmeijer (2016) "A weak instrument F-test in linear IV models with multiple endogenous variables," *Journal of Econometrics*, Vol. 190, pp. 212–221.

Sarah Flood, Renae Rodgers Steven Ruggles, Miriam King and J. Robert Warren. (2014) "Integrated Public Use Microdata Series, Current Population Survey: Version 6.0 [dataset]. Minneapolis, MN: IPUMS, 2018.," Permanent url `https://doi.org/10.18128/D030.V6.0`.

Schmitt, John and Alexandra Mitukiewicz (2012) "Politics matter: changes in unionisation rates in rich countries, 1960–2010," *Industrial Relations Journal*, Vol. 43, pp. 260–280.

Snoddy, Iain G (2019) "Learning About Selection: An Improved Correction Procedure," *Working Paper*.

Stock, James H and Mark M Watson (2011) "Introduction to Econometrics, 3rd international edition."

Taschereau-Dumouchel, Mathieu (2017) "The Union Threat," *Working Paper*.

Tibshirani, Robert (1996) "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, pp. 267–288.

Tschopp, Jeanne (2017) "Wage Formation: Towards Isolating Search and Bargaining Effects from the Marginal Product," *The Economic Journal*, Vol. 127, pp. 1693–1729.

Vella, Francis (1998) "Estimating Models with Sample Selection Bias: A Survey," *The Journal of Human Resources*, Vol. 33, pp. 127–169.

Willis, Robert J and Sherwin Rosen (1979) "Education and self-selection," *Journal of Political Economy*, Vol. 87, pp. S7–S36.

Zabek, Mike (2018) "Local Ties in Spatial Equilibrium," Mimeo, Available: https://doi.org/10.31235/osf.io/rpq5z.

Appendix A

Appendix to Chapter 1

A.1 A Brief Discussion of Artifical Neural Networks

An Artificial Neural Network (ANN) is a system designed to mimic how the brain processes information. This system maps inputs to outputs through interconnected nodes (neurons). An example of a feedforward multi-layer perceptron ANN with a single hidden layer is presented in Figure A.1. An ANN takes inputs (variables) which connect to nodes in a hidden layer as shown in the figure. Each connected line between input and node in the hidden layer is assigned a weight. Weights can be positive or negative. At each node, a new variable is created, being the weighted sum of inputs. From the hidden layer, the weighted variable at each hidden node is then mapped to the output layer, which may contain a single node, or a collection of nodes. From the output layer, a second activation/transfer function maps to actual outputs from the network.

The ANN system functions according to the weights assigned to the lines connecting inputs to hidden nodes, and on the form of the transfer function which maps inputs to outputs. To train the Neural Network weights are adjusted using back-propagation to determine the error contribution of a change in the weights. This is done by calculating the gradient of the loss function. Error is fed backwards through the system and weights are adjusted accordingly. Error is estimated using a hold-out data-set and so the model is typically trained on a subset of the data.

In principal, if the NN is well estimated it should be relatively stable with regards to the training data used. Typically, non-linear transfer functions are chosen to map weighted inputs to outputs.[1] A NN with just one hidden layer and a single node functions in a manner similar to a logit model, with weights on each line connecting inputs to the hidden node representing

[1]Transfer functions can be binary, linear, but are often non-linear and sigmoidal. The log and tanh sigmoid functions are popular transfer functions.

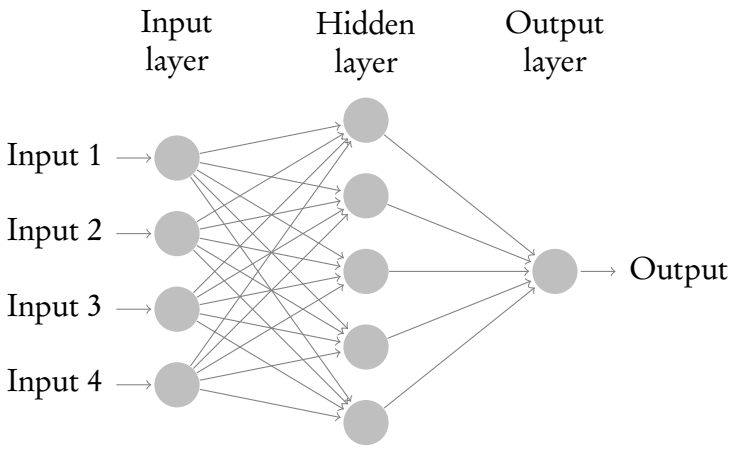


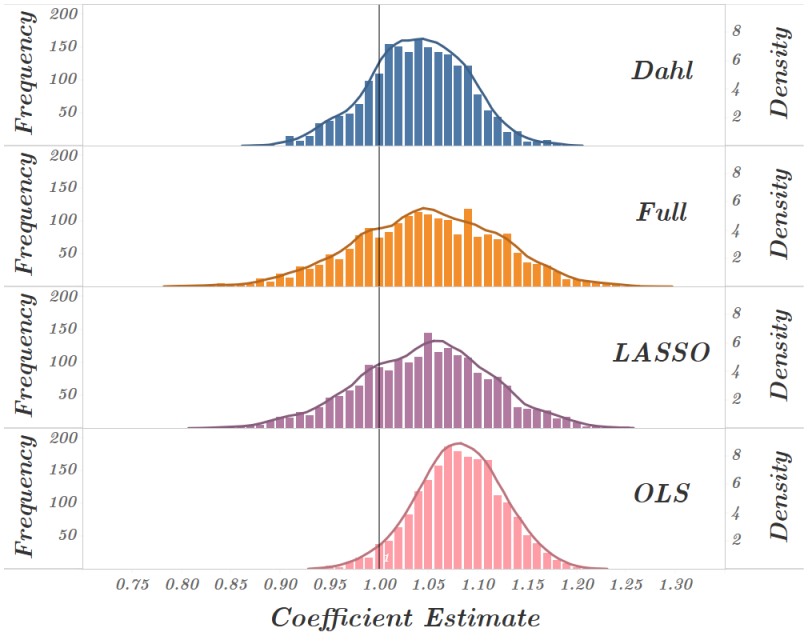
Figure A.1: Feedforward Neural Network (FNN) with 4 inputs, 1 hidden layer with 5 nodes, and a single output

regression coefficients. With multiple nodes, the NN allows for more complex interactions between inputs, and with more hidden layers, the NN allows for more complex non-linear relationships between input variables and outputs.

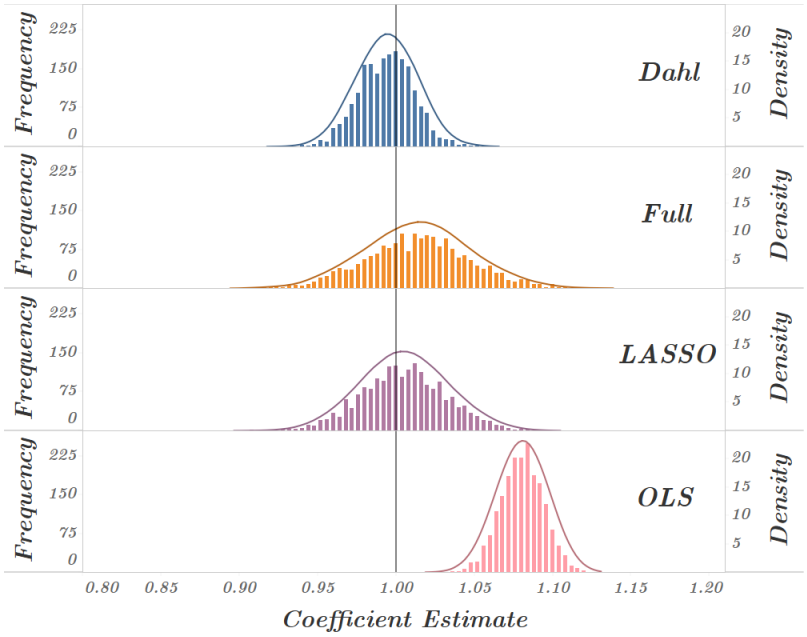
A.2 Monte Carlo Figures and Tables

Figure A.2: Distribution of Monte Carlo Estimates: 5 Locations, SISA Holds Violated

(a) $N = 1000$



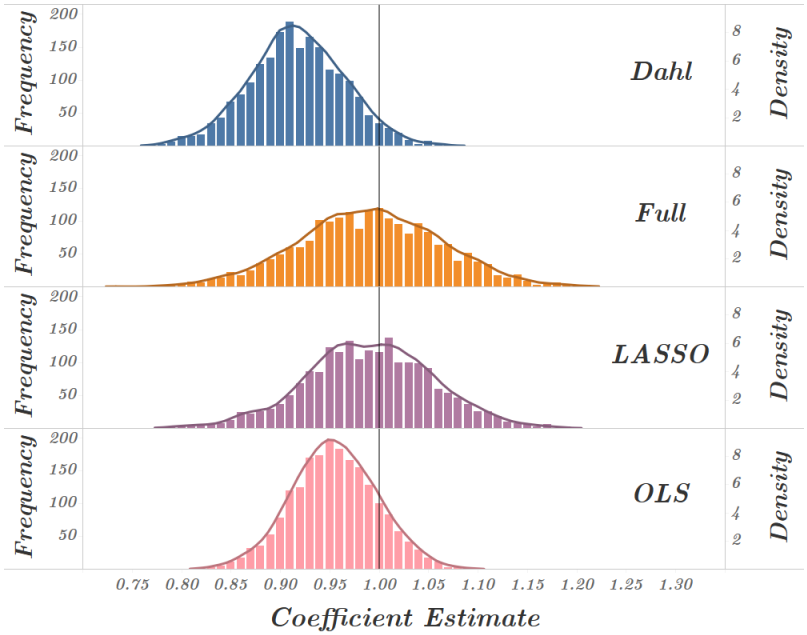
(B) $N = 10000$



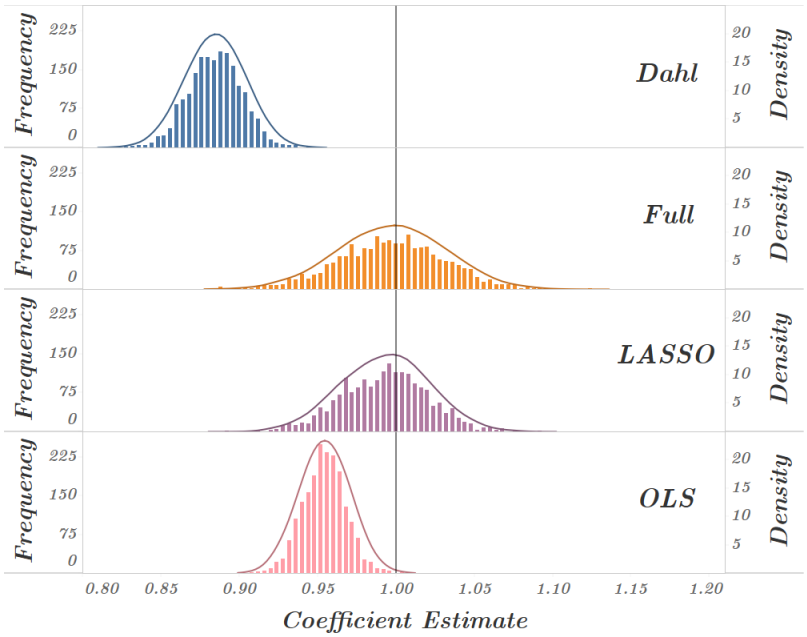
Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Figure A.3: Distribution of Monte Carlo Estimates: 5 Locations, SISA Strongly Violated

(a) $N = 1000$



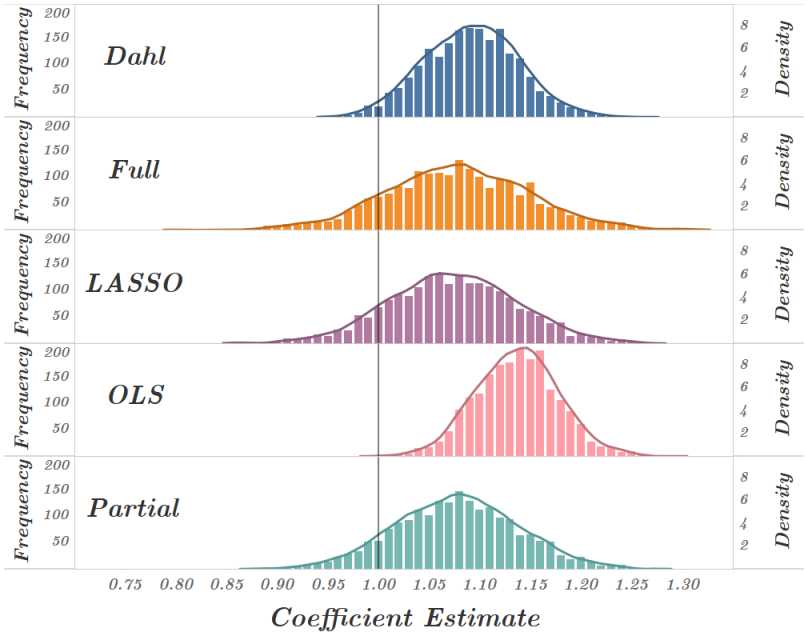
(B) $N = 10000$



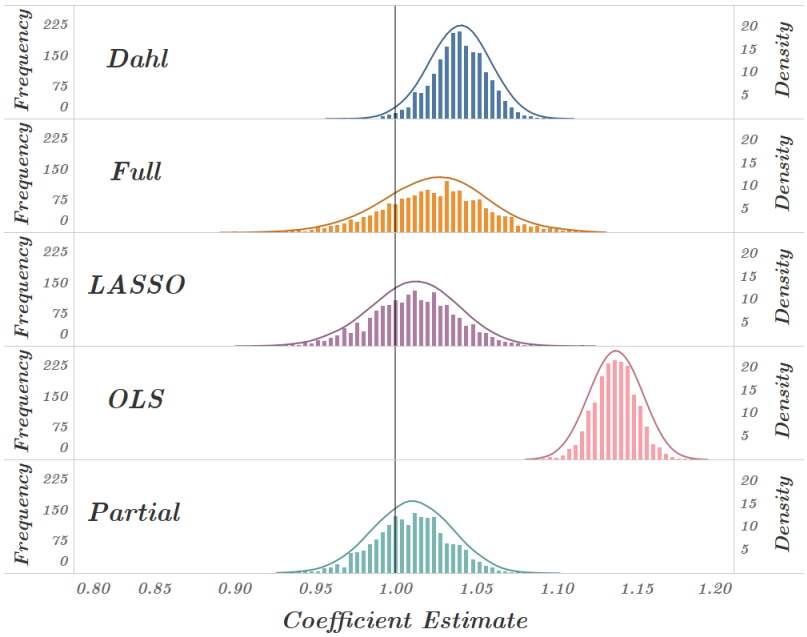
Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Figure A.4: Distribution of Monte Carlo Estimates: 5 Locations, SISA Weakly Violated

(a) $N = 1000$



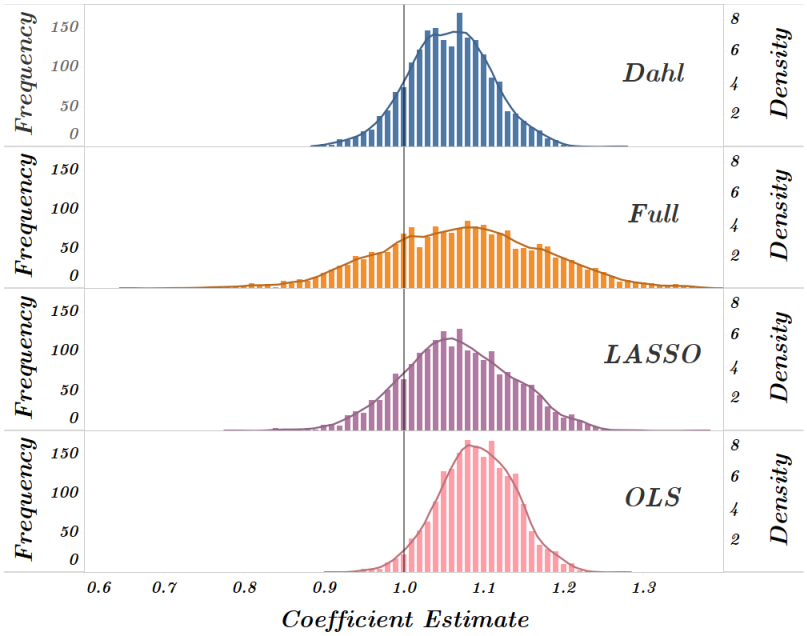
(B) $N = 10000$



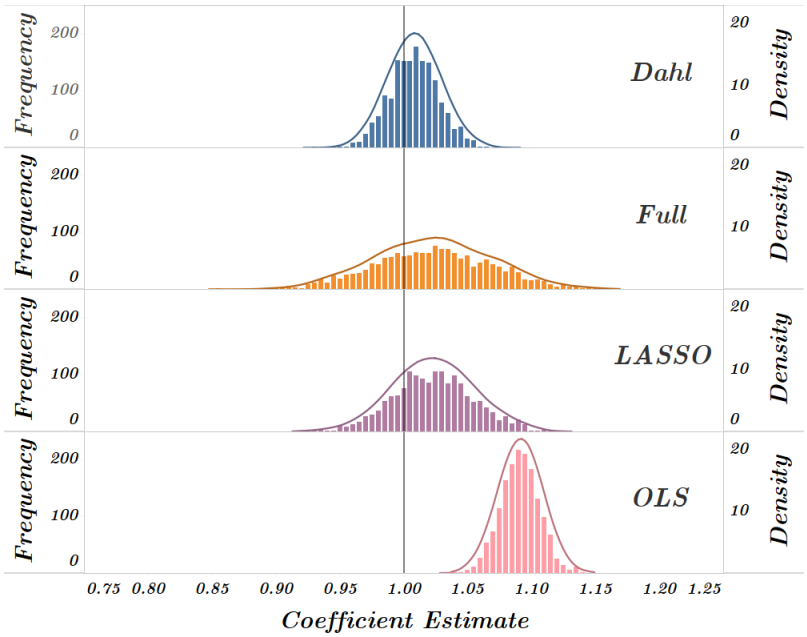
Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Figure A.5: Distribution of Monte Carlo Estimates: 10 Locations, SISA Holds

(a) $N = 1000$



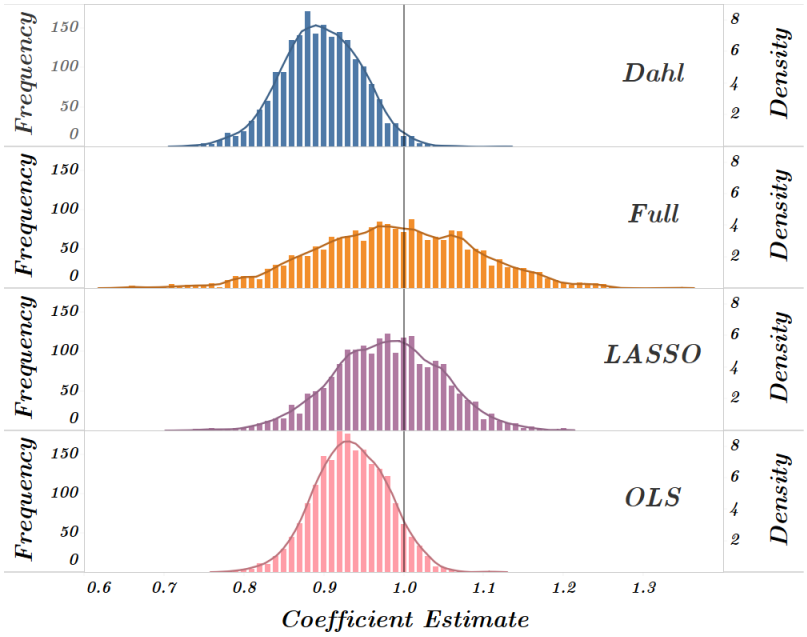
(B) $N = 10000$



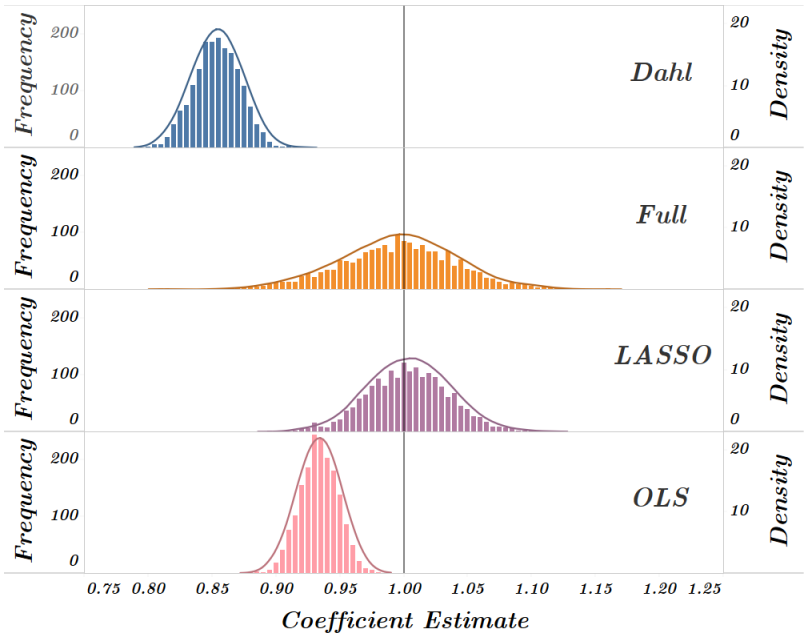
Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Figure A.6: Distribution of Monte Carlo Estimates: 10 Locations, SISA Strongly Violated

(a) $N = 1000$



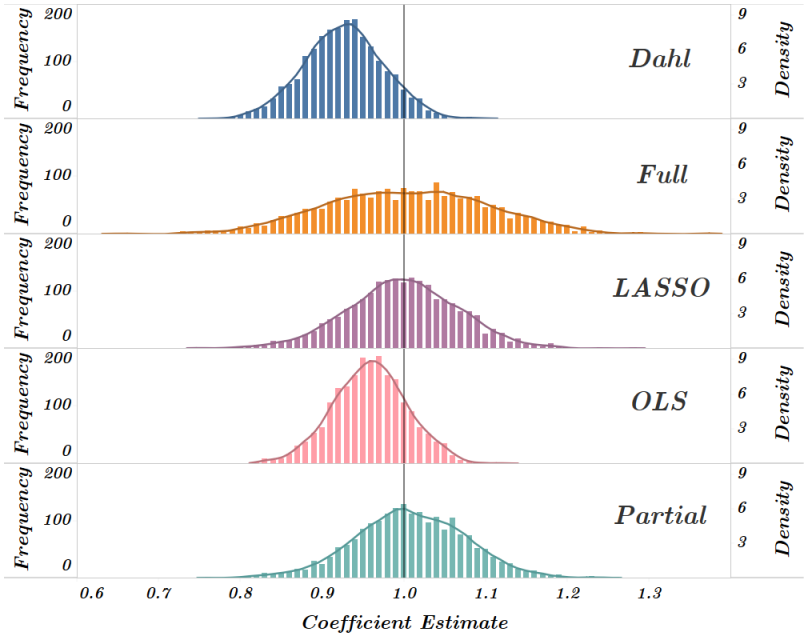
(B) $N = 10000$



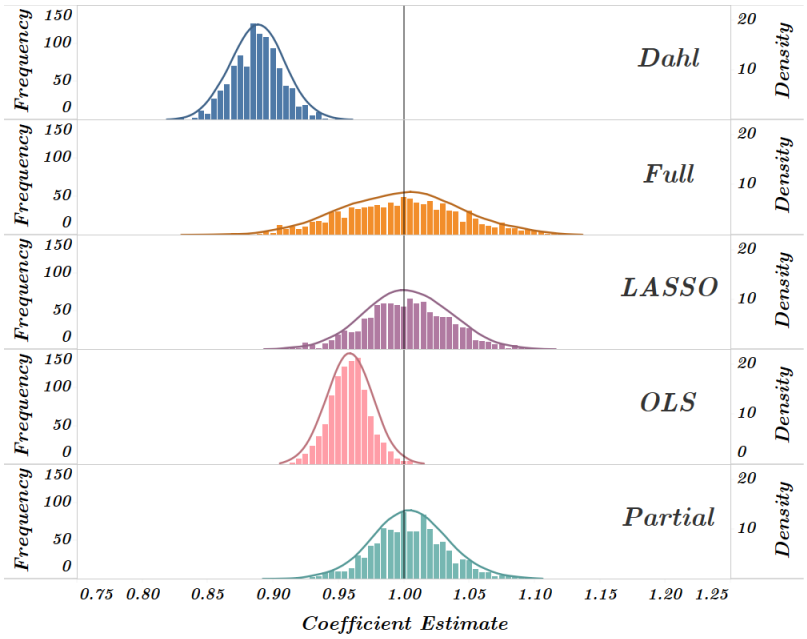
Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Figure A.7: Distribution of Monte Carlo Estimates: 10 Locations, SISA Weakly Violated

(a) $N = 1000$



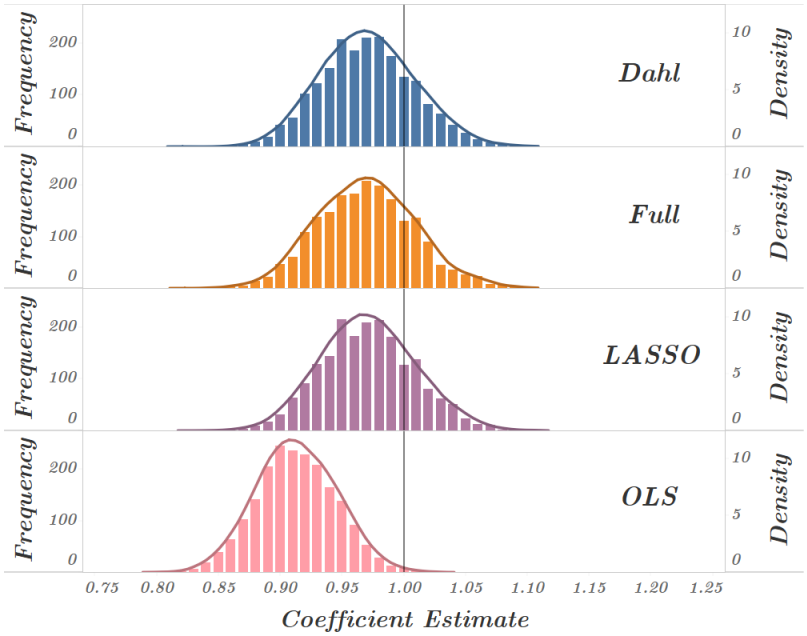
(B) $N = 10000$



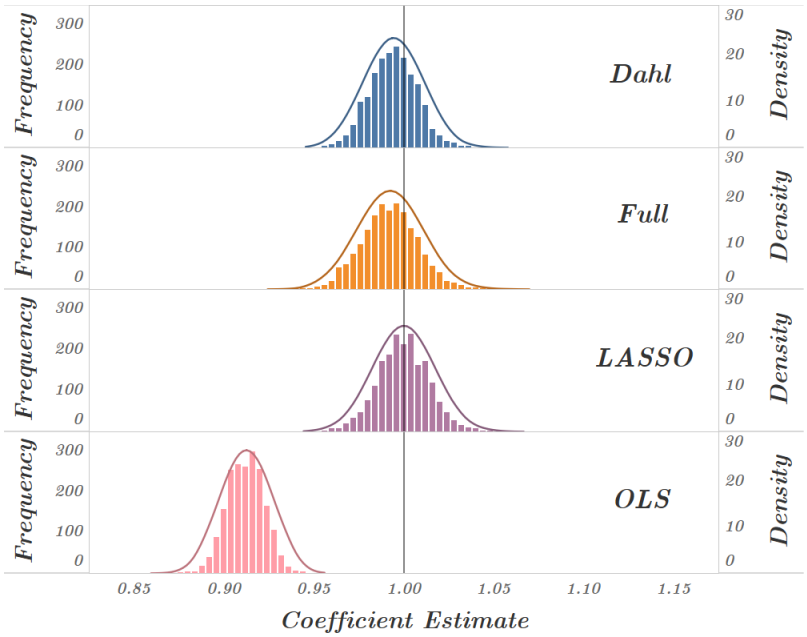
Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Figure A.8: Distribution of Monte Carlo Estimates: 3 Locations, SISA Holds

(a) $N = 1000$



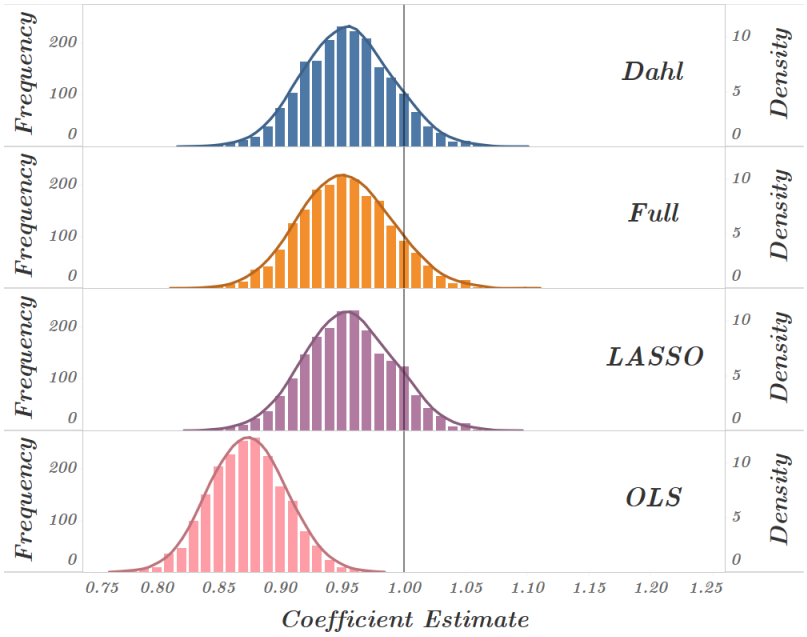
(B) $N = 10000$



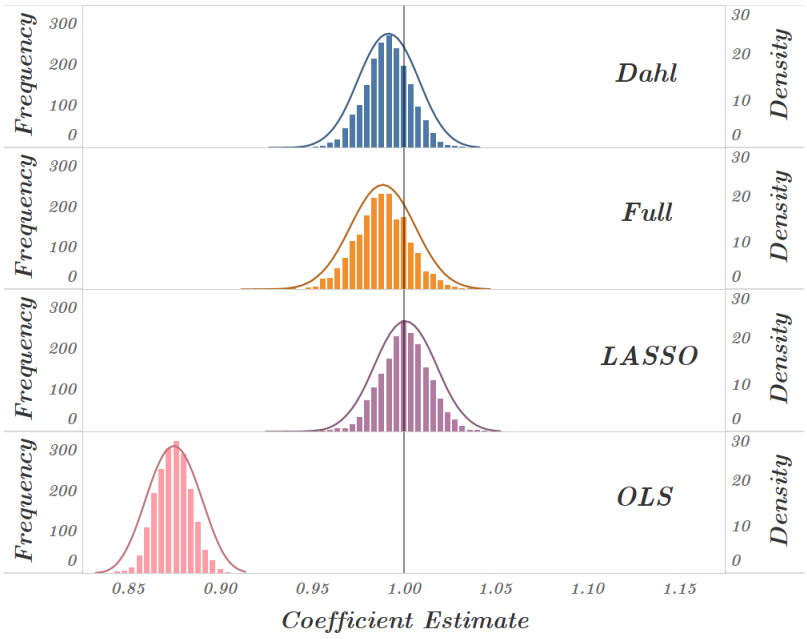
Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Figure A.9: Distribution of Monte Carlo Estimates: 3 Locations, SISA Strongly Violated

(a) $N = 1000$



(B) $N = 10000$



Notes: Histograms and smoothed kernel density plots of coefficient estimates obtained from the Monte Carlo experiment in 1.4. Each panel summarises the distribution of estimates across replications for a particular estimation method. The frequency of estimates is plotted on the left y-axis, and the density of estimates corresponding to the smoothed kernel is on the right y-axis. The vertical line at 1 corresponds to the true coefficient value.

Appendix B

Appendix to Chapter 2

B.1 Data Appendix

Table B.1: Geographic Divisions & Regions

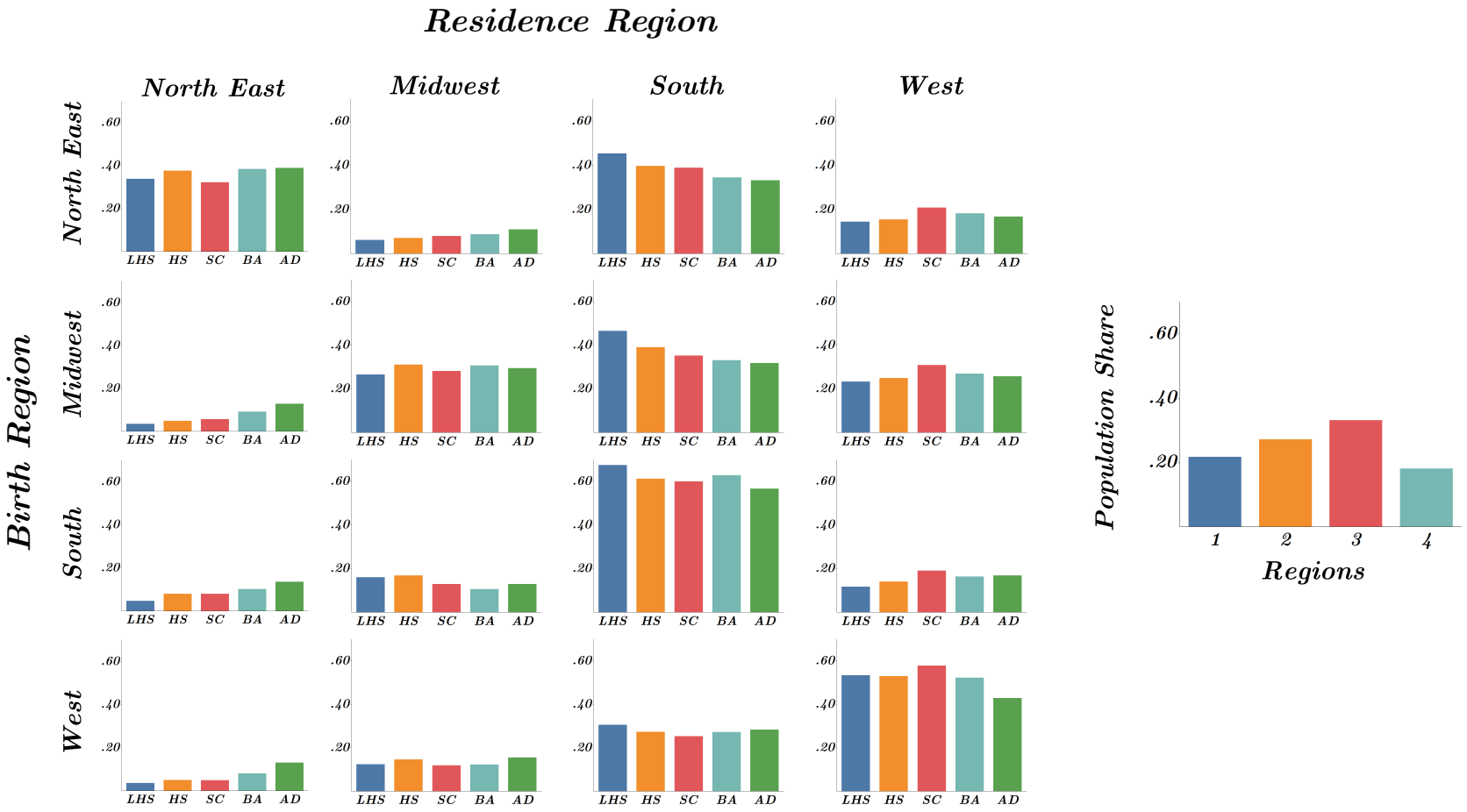
Region: NORTHEAST	
Division: New England	Division: Mid-Atlantic
Connecticut	New Jersey
Maine	New York
Massachusetts	Pennsylvania
New Hampshire	
Rhode Island	
Vermont	

Region: MIDWEST	
Division: East North Central	Division: West North Central
Illinois	Iowa
Indiana	Kansas
Michigan	Minnesota
Ohio	Missouri
Wisconsin	Nebraska
	North Dakota
	South Dakota

Region: SOUTH		
Division: West South Central	Division: South Atlantic	Division: East South Central
Arkansas	Delaware	Alabama
Louisiana	District of Columbia	Kentucky
Oklahoma	Florida	Mississippi
Texas	Georgia	Tennessee
	Maryland	
	North Carolina	
	South Carolina	
	Virginia	
	West Virginia	

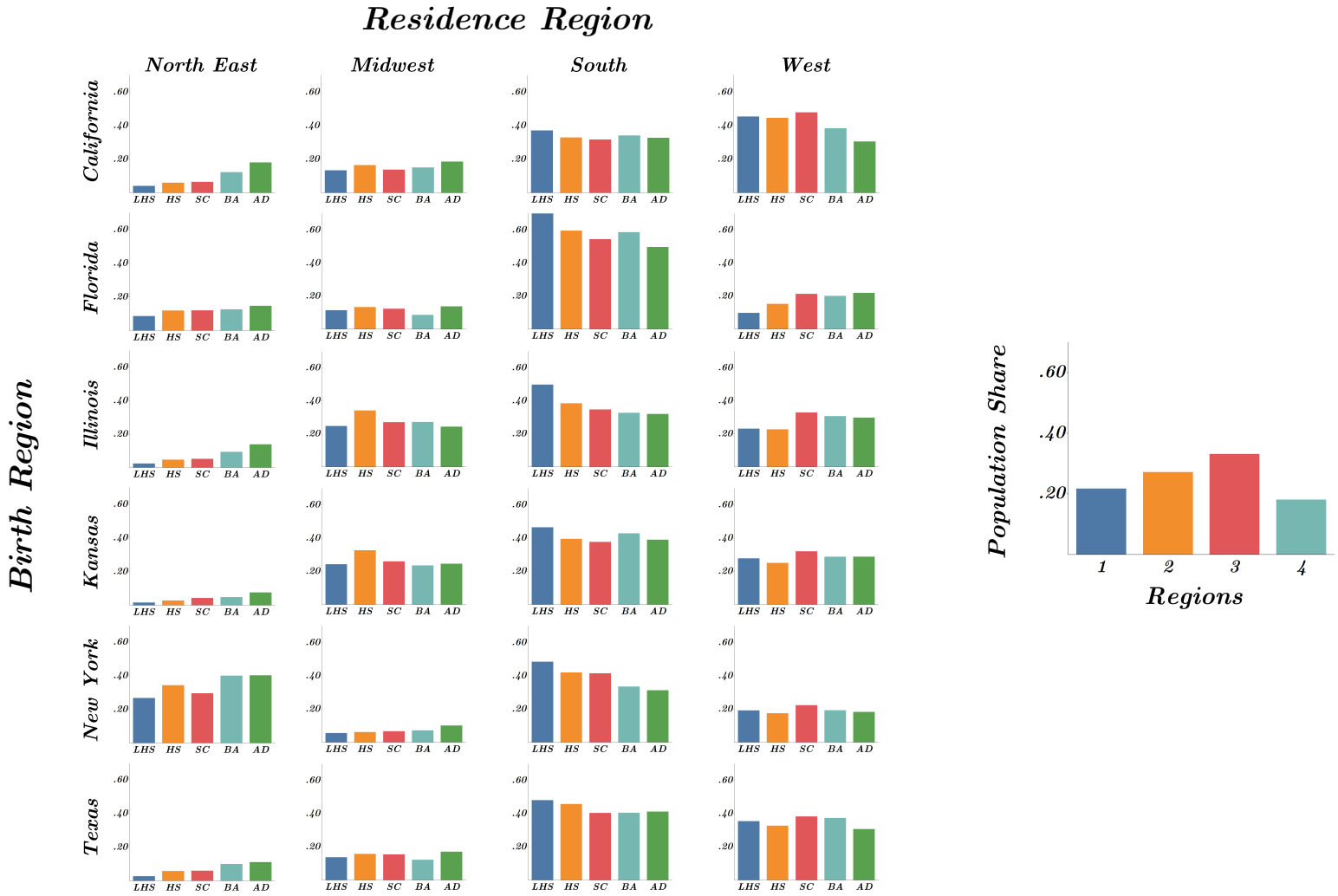
Region: WEST	
Division: Mountain	Division: Pacific
Arizona	Alaska
Colorado	California
Idaho	Hawaii
Montana	Oregon
Nevada	Washington
New Mexico	
Utah	
Wyoming	

Figure B.1: Transition Matrix between Regions



Notes: Raw transition probabilities between regions are calculated using the entire sample of white males aged 24-35. See Table B.1 for the list of states in each region. To put these migration paths in context, the panel to the right of the matrix shows the share of the population living in each region.

Figure B.2: Transition Matrix between Selected States and Regions



Notes: Raw transition probabilities between states and regions are calculated using the entire sample of white males aged 24-35. Stayers are dropped from the sample so migration paths are calculated conditional on moving. See Table B.1 for the list of states in each region. To put these migration paths in context, the panel to the right of the matrix shows the share of the population living in each region.

Table B.2: Aggregated Industry Definitions

Category	Code	1990 Industry Codes
Agriculture Service	1	12, 20, 21 , 30
Other Agriculture	2	10 - 11
Mining	3	40 - 50
Construction	4	60
Lumber and Wood Products, except Furniture	5	230 - 241
Furniture and Fixtures	6	242
Stone Clay, Glass, and Concrete Product	7	250 - 262
Primary Metals	8	270 - 280
Fabricated Metal	9	281 - 300
Not Specified Metal Industries	10	301
Machinery, except Electrical	11	310 - 332
Electrical Machinery, Equipment, and Supplies	12	340 - 350
Motor Vehicles and Equipment	13	351
Aircraft and Parts	14	352
Other Transportation Equipment	15	360 - 370
Professional and Photographic Equipment, and Watches	16	371 - 382
Toys, Amusements, and Sporting Goods	17	390
Miscellaneous and Not Specified Manufacturing Industries	18	391 - 392
Food and Kindred Products	19	100 - 122
Tobacco Manufactures	20	130
Textile Mill Products	21	132 - 150
Apparel and Other Finished Textile Products	22	151 - 152
Paper and Allied Products	23	160 - 162
Printing, Publishing and Allied Industries	24	171 - 172
Chemicals and Allied Products	25	180 - 192
Petroleum and Coal Products	26	200 - 201
Rubber and Miscellaneous Plastics Products	27	210 - 212
Leather and Leather Products	28	220 - 222
Transportation	29	400 - 432
Communications	30	440 - 442
Utilities and Sanitary Services	31	450 - 452, 460 - 472
Wholesale Trade	32	500 - 571
Retail Trade	33	580 - 691
Banking and Other Finance	34	700 - 710
Insurance and Real Estate	35	711 - 712
Private Household Services	36	761
Business Services	37	721, 722, 731 - 750, 892
Repair Services	38	751 - 760
Personal Services, except Private Household	39	762 - 791
Entertainment and Recreation Services	40	800 - 802, 810
Hospitals	41	831
Health Services, except Hospitals	42	812 - 830, 832 - 840
Educational Services	43	842 - 860
Social Services	44	861 - 871
Other Professional Services	45	730, 841, 872 - 891, 893
Forestry and Fisheries	46	31 - 32
Justice, Public Order and Safety	47	910
Administration Of Human Resource Programs	48	922
National Security and Internal Affairs	49	932
Other Public Administration	50	900, 901, 921, 930, 931

Notes: List of aggregated industries and corresponding 1990 codes used by the US Census Bureau.

B.1.1 Data Sources for Roy Model Estimation

In this section, I present additional data sources used for estimation in Tables 2.18 - 2.19.

Unemployment Rate

The 5 Year average unemployment rate is calculated using data from the Bureau of Labor Statistics which calculates unemployment from the Current Population Survey

Quality of Life Variables

Data on population density in 1990 by state is taken from the US census Bureau: https://www.census.gov/population/www/censusdata/density.html. I include the average teacher salary by state using data from the National Center for Education Statistics: https://nces.ed.gov/programs/digest/d01/dt079.asp. From the same source, I use data on average educational expenditure per pupil. I include variables on the crime rate, the violent crime rate and the incarceration rate by state from http://www.disastercenter.com/ which reports data from the Bureau of Justice Statistics.

Climate Variables

All climate data are take from the U.S. National Oceanic and Atmospheric Administration: https://www.ncdc.noaa.gov/ghcn/comparative-climatic-data. I use data on the average, maximum, and minimum daily temperature for all states, the total amount of rain yearly, the number of sunny and rainy days, the average level of humidity in the afternoon, and average wind speed.

State Budget Data

Data on state spending and taxation is collected using American FactFinder on the website of the US Census Bureau. I include state spending on education, health, highways, and public welfare as well as miscellaneous spending. On the revenue side, I use the average sales tax and the average state income state.

B.2 Results Tables and Figures

Table B.3: Corrected Estimates for Selected States using Dahl's Approach and Random Forest Estimated Probabilities: Separate Control Function for Stayers and Movers

	CA	FL	IL	KS	NY	TX
LHS	-0.153***	-0.147***	-0.190***	-0.194***	-0.186***	-0.176***
	(0.010)	(0.011)	(0.012)	(0.023)	(0.011)	(0.010)
Some College	0.129***	0.113***	0.086***	0.033**	0.134***	0.132***
	(0.006)	(0.008)	(0.008)	(0.015)	(0.007)	(0.007)
College	0.417***	0.420***	0.309***	0.322***	0.393***	0.481***
	(0.008)	(0.011)	(0.011)	(0.024)	(0.010)	(0.010)
Advanced	0.557***	0.623***	0.462***	0.467***	0.539***	0.624***
	(0.011)	(0.017)	(0.016)	(0.037)	(0.014)	(0.015)
Exper	0.083***	0.049***	0.073***	0.038	0.083***	0.048***
	(0.011)	(0.015)	(0.014)	(0.029)	(0.012)	(0.013)
$Exper^2$	-0.004***	-0.001	-0.003**	0.000	-0.004***	0.001
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$Exper^3 \times 100$	0.006	-0.001	0.005	-0.004	0.008*	-0.008*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.004)
Married	0.085***	0.149***	0.153***	0.157***	0.146***	0.177***
	(0.006)	(0.007)	(0.007)	(0.013)	(0.006)	(0.006)
SMSA	0.179***	0.111***	0.255***	0.236***	0.235***	0.124***
	(0.011)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
R-Squared	0.145	0.169	0.195	0.181	0.196	0.194
Obs	46727	25579	26533	6234	36765	34890

Notes: Bootstrapped standard errors (500 replications) in parentheses to control for variability in two-step estimation. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

Table B.4: Corrected Estimates for Selected States using Dahl's Approach and Random Forest Estimated Probabilities: Single Control Function

	CA	FL	IL	KS	NY	TX
LHS	-0.155***	-0.151***	-0.192***	-0.194***	-0.187***	-0.181***
	(0.010)	(0.011)	(0.012)	(0.023)	(0.011)	(0.010)
Some College	0.123***	0.120***	0.089***	0.035**	0.136**	0.138**
	(0.006)	(0.008)	(0.008)	(0.015)	(0.007)	(0.007)
College	0.420***	0.441***	0.323***	0.333***	0.403***	0.496***
	(0.008)	(0.010)	(0.011)	(0.022)	(0.010)	(0.009)
Advanced	0.566***	0.650***	0.485***	0.483***	0.558***	0.650***
	(0.011)	(0.016)	(0.015)	(0.034)	(0.013)	(0.014)
Exper	0.082***	0.049***	0.075***	0.039	0.084***	0.049***
	(0.011)	(0.015)	(0.014)	(0.029)	(0.012)	(0.013)
$Exper^2$	-0.003***	-0.001	-0.003**	0.000	-0.004***	0.001
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$Exper^3 \times 100$	0.005	-0.001	0.005	-0.003	0.009**	-0.008*
	(0.004)	(0.005)	(0.005)	(0.010)	(0.004)	(0.004)
Married	0.117***	0.151***	0.156***	0.157***	0.148***	0.174***
	(0.005)	(0.006)	(0.007)	(0.013)	(0.006)	(0.006)
SMSA	0.179***	0.113***	0.256***	0.237***	0.235***	0.125***
	(0.011)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
R-Squared	0.143	0.169	0.194	0.180	0.195	0.194
Obs	46727	25579	26533	6234	36765	34890

Notes: Bootstrapped standard errors (500 replications) in parentheses to control for variability in two-step estimation. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

Table B.5: Corrected Estimates for Selected States using post-double-Lasso and Cell Estimates of Migration Probabilities

	CA	FL	IL	KS	NY	TX
LHS	-0.162***	-0.134***	-0.151***	-0.185***	-0.141***	-0.181***
	(0.013)	(0.014)	(0.021)	(0.034)	(0.020)	(0.014)
Some College	0.091***	0.086***	0.083***	0.080***	0.104***	0.110***
	(0.010)	(0.013)	(0.016)	(0.027)	(0.017)	(0.011)
College	0.382***	0.346***	0.240***	0.412***	0.379***	0.467***
	(0.017)	(0.021)	(0.031)	(0.049)	(0.027)	(0.018)
Advanced	0.498***	0.498***	0.303***	0.520***	0.450***	0.591***
	(0.026)	(0.033)	(0.046)	(0.076)	(0.037)	(0.032)
Exper	0.083***	0.044***	0.058***	0.045	0.073***	0.051***
	(0.012)	(0.015)	(0.014)	(0.031)	(0.012)	(0.013)
$Exper^2$	-0.004***	-0.001	-0.002	-0.001	-0.004***	0.000
	(0.001)	(0.002)	(0.002)	(0.003)	(0.001)	(0.001)
$Exper^3 \times 100$	0.007*	-0.002	0.003	-0.001	0.007	-0.008*
	(0.004)	(0.005)	(0.005)	(0.011)	(0.004)	(0.005)
Married	0.140***	0.144***	0.162***	0.136***	0.134***	0.162***
	(0.008)	(0.009)	(0.013)	(0.019)	(0.012)	(0.009)
SMSA	0.178***	0.108***	0.255***	0.236***	0.231***	0.124***
	(0.011)	(0.008)	(0.007)	(0.013)	(0.006)	(0.006)
# Cells	338	304	285	152	277	306
R-Squared	0.152	0.175	0.206	0.191	0.203	0.202
Obs	46727	25579	26533	6234	36765	34890

Notes: Standard errors in parentheses. Significance levels: * 10%, ** 5%, *** 1%. Model estimated separately for each state using the 1990 US Census for white males aged 25-34. The dependent variable is log hourly wages.

Table B.6: Corrected Estimates versus OLS

	CA	FL	IL	KS	NY	TX
	Dahl 2 Control Function - Random Forest					
LHS	0.02	1.33***	-0.40***	0.01	0.74***	1.01***
	(0.07)	(22.91)	(45.88)	(0.08)	(212.51)	(77.93)
Some College	1.08***	-1.22***	-1.19***	-0.64	-1.23***	-1.38***
	(229.14)	(48.68)	(215.64)	(2.43)	(76.82)	(85.89)
College	-1.19***	-3.07***	-6.01***	-2.45	-4.73***	-3.59***
	(38.60)	(30.73)	(138.49)	(2.64)	(77.37)	(75.14)
Advanced	-2.97***	-3.92***	-8.24***	-3.00	-6.47***	-5.97***
	(94.88)	(29.29)	(116.10)	(2.36)	(78.11)	(74.31)
	Dahl 1 Control Function - Random Forest					
LHS	-0.12	0.94***	-0.62***	0.01	0.63***	0.57***
	(2.38)	(29.12)	(107.07)	(0.02)	(370.17)	(103.36)
Some College	0.53***	-0.48***	-0.91***	-0.38	-1.02***	-0.82***
	(175.46)	(25.53)	(172.86)	(1.66)	(60.67)	(67.59)
College	-0.87***	-0.95***	-4.56***	-1.39	-3.72***	-2.10***
	(74.51)	(21.43)	(114.88)	(1.66)	(56.85)	(63.53)
Advanced	-2.08***	-1.14***	-5.98***	-1.41	-4.54***	-3.39***
	(155.46)	(31.98)	(107.55)	(1.45)	(50.90)	(69.19)

Notes: Corrected estimates compared to OLS estimates for the returns to education. Presented values are corrected estimates, less OLS estimate, multiplied by 100. Hausman test F-statistic in brackets. Hausman test significance: * 10%, ** 5%, *** 1%. Estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34.

Table B.7: Lasso Estimates versus Dahl

	CA	FL	IL	KS	NY	TX
	Lasso versus Dahl (1 Control Functions, Cell Probabilities)					
LHS	2.83	2.75	5.87	0.18	0.61	2.30
Some College	-1.83	-1.04	3.73	-0.31	0.77	-1.82
College	-3.90	-8.40	1.03	0.60	-1.73	-2.10
Advanced	-7.33	-13.76	-2.95	-3.65	-7.48	-2.64
	Lasso versus Dahl (2 Control Functions, Cell Probabilities)					
LHS	2.82	2.93	6.20	0.28	0.82	2.31
Some College	-1.76	1.41	4.92	0.04	1.63	-0.70
College	-3.04	-4.38	3.49	2.18	-0.24	0.02
Advanced	-6.31	-9.32	0.12	-0.27	-5.09	0.37
	Lasso versus Dahl (1 Control Functions, RF Probabilities)					
LHS	3.17	2.38	3.91	-0.20	-0.63	2.00
Some College	-1.55	-1.54	2.06	-0.57	-0.71	-2.50
College	-3.90	-8.54	0.13	0.86	-2.30	-2.63
Advanced	-7.66	-13.59	-3.47	-3.25	-7.50	-3.19
	Lasso versus Dahl (2 Control Functions, RF Probabilities)					
LHS	3.03	1.99	3.69	-0.20	-0.73	1.56
Some College	-2.10	-0.80	2.34	-0.32	-0.50	-1.93
College	-3.58	-6.42	1.59	1.92	-1.29	-1.14
Advanced	-6.77	-10.82	-1.22	-1.66	-5.57	-0.62

Notes: PDL corrected estimates compared to estimates corrected using Dahl's approach. Presented values are PDL estimates, less Dahl estimate, multiplied by 100. Estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34.

Table B.8: PDL Estimates versus Uncorrected Estimates for all States: Less than High School Premia

State	Corrected	Uncorrected	Hausman	State	Corrected	Uncorrected	Hausman
AL	-0.197	-0.221	1.052	MT	-0.143	-0.150	0.161
AK	-0.155	-0.141	0.232	NE	-0.176	-0.147	1.106
AZ	-0.192	-0.199	0.585	NV	-0.150	-0.146	0.685
AR	-0.191	-0.205	0.913	NH	-0.146	-0.143	0.020
CA	-0.123	-0.154	53.589*	NJ	-0.124	-0.157	4.381*
CO	-0.170	-0.173	0.074	NM	-0.145	-0.152	1.540
CT	-0.177	-0.156	0.720	NY	-0.193	-0.194	0.000
DE	-0.128	-0.134	0.054	NC	-0.096	-0.148	13.366*
DC	0.286	0.273	0.809	ND	-0.210	-0.187	0.251
FL	-0.127	-0.160	29.617*	OH	-0.220	-0.209	0.645
GA	-0.168	-0.188	2.597	OK	-0.193	-0.245	11.821*
ID	-0.236	-0.177	6.667*	OR	-0.150	-0.161	0.723
IL	-0.153	-0.186	4.796*	PA	-0.153	-0.170	1.877
IN	-0.224	-0.216	0.227	RI	-0.113	-0.167	2.520
IA	-0.131	-0.144	0.364	SC	-0.128	-0.169	3.766*
KS	-0.196	-0.194	0.012	SD	-0.078	-0.123	6.697*
KY	-0.200	-0.257	11.880*	TN	-0.236	-0.223	0.923
LA	-0.226	-0.250	0.983	TX	-0.161	-0.186	7.089*
ME	-0.201	-0.147	2.444	UT	-0.263	-0.222	4.340*
MD	-0.106	-0.159	22.981*	VT	-0.169	-0.179	0.153
MA	-0.185	-0.181	0.072	VA	-0.143	-0.193	11.167*
MI	-0.174	-0.155	0.746	WA	-0.136	-0.141	0.214
MN	-0.158	-0.150	0.137	WV	-0.198	-0.218	0.513
MS	-0.179	-0.195	0.293	WI	-0.121	-0.176	5.356*
MO	-0.213	-0.212	0.007	WY	-0.201	-0.233	1.561

Notes: Hausman F-test presented in columns 4 and 8. * indicates a significant difference between OLS and corrected estimates at a 5% level. Estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. Corrected estimates use Random Forest estimated migration probabilities.

Table B.9: PDL Estimates versus Uncorrected Estimates for all States: Some College Premia

State	Corrected	Uncorrected	Hausman	State	Corrected	Uncorrected	Hausman
AL	0.044	0.090	5.489*	MT	-0.045	-0.027	0.443
AK	0.066	0.054	10.341*	NE	0.115	0.092	0.990
AZ	0.133	0.135	0.221	NV	0.072	0.068	1.613
AR	0.073	0.088	0.709	NH	0.109	0.116	0.227
CA	0.108	0.118	6.252*	NJ	0.066	0.115	18.102*
CO	0.120	0.129	1.482	NM	0.112	0.129	2.290
CT	0.077	0.080	0.055	NY	0.129	0.146	2.236
DE	0.123	0.126	0.003	NC	0.094	0.133	13.981*
DC	0.180	0.194	0.659	ND	0.120	0.110	0.129
FL	0.105	0.125	21.573*	OH	0.097	0.128	4.179*
GA	0.110	0.123	1.384	OK	0.077	0.084	0.401
ID	0.061	0.061	0.000	OR	0.086	0.057	6.344*
IL	0.110	0.098	1.247	PA	0.053	0.128	29.793*
IN	0.088	0.102	0.775	RI	0.070	0.064	0.068
IA	0.039	0.074	2.155	SC	0.083	0.116	5.282*
KS	0.030	0.039	0.509	SD	0.104	0.092	0.213
KY	0.074	0.113	2.795*	TN	0.111	0.129	1.193
LA	0.060	0.100	3.368*	TX	0.113	0.146	37.819*
ME	0.097	0.114	0.417	UT	0.088	0.078	0.497
MD	0.095	0.115	2.280	VT	0.150	0.139	0.156
MA	0.067	0.074	0.438	VA	0.064	0.105	21.497*
MI	0.115	0.132	1.664	WA	0.069	0.059	1.320
MN	0.081	0.099	1.081	WV	0.226	0.151	5.712*
MS	0.102	0.127	1.507	WI	0.066	0.114	6.812*
MO	0.065	0.084	1.527	WY	0.047	0.047	0.011

Notes: Hausman F-test presented in columns 4 and 8. * indicates a significant difference between OLS and corrected estimates at a 5% level. Estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. Corrected estimates use Random Forest estimated migration probabilities.

Table B.10: PDL Estimates versus Uncorrected Estimates for all States: College Premia

State	Corrected	Uncorrected	Hausman	State	Corrected	Uncorrected	Hausman
AL	0.370	0.478	7.919*	MT	0.254	0.276	0.189
AK	0.364	0.372	1.248	NE	0.350	0.329	0.235
AZ	0.498	0.493	0.121	NV	0.368	0.366	0.040
AR	0.478	0.396	5.936*	NH	0.322	0.366	2.305
CA	0.381	0.429	26.799*	NJ	0.313	0.403	14.065*
CO	0.424	0.463	3.619*	NM	0.444	0.446	0.009
CT	0.366	0.348	0.352	NY	0.380	0.440	6.345*
DE	0.332	0.373	1.135	NC	0.416	0.451	2.200
DC	0.374	0.407	1.403	ND	0.404	0.430	0.127
FL	0.356	0.451	48.316*	OH	0.343	0.415	5.839*
GA	0.377	0.430	5.655*	OK	0.357	0.399	2.069
ID	0.370	0.312	2.631	OR	0.366	0.311	5.888*
IL	0.325	0.369	3.573*	PA	0.246	0.392	33.329*
IN	0.374	0.362	0.115	RI	0.325	0.366	1.092
IA	0.271	0.325	0.980	SC	0.384	0.412	1.027
KS	0.341	0.346	0.027	SD	0.398	0.405	0.013
KY	0.310	0.419	6.066*	TN	0.417	0.455	1.397
LA	0.419	0.433	0.127	TX	0.469	0.517	10.181*
ME	0.442	0.367	2.790*	UT	0.373	0.344	0.699
MD	0.289	0.412	30.041*	VT	0.384	0.360	0.327
MA	0.277	0.321	4.789*	VA	0.352	0.424	12.091*
MI	0.372	0.393	0.438	WA	0.352	0.350	0.007
MN	0.306	0.320	0.134	WV	0.560	0.453	2.682
MS	0.400	0.390	0.049	WI	0.245	0.328	3.315*
MO	0.350	0.354	0.017	WY	0.287	0.272	1.104

Notes: Hausman F-test presented in columns 4 and 8. * indicates a significant difference between OLS and corrected estimates at a 5% level. Estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. Corrected estimates use Random Forest estimated migration probabilities.

Table B.11: PDL Estimates versus Uncorrected Estimates for all States: Advanced Degree Premia

State	Corrected	Uncorrected	Hausman	State	Corrected	Uncorrected	Hausman
AL	0.362	0.530	6.462*	MT	0.492	0.568	0.749
AK	0.505	0.572	2.778*	NE	0.261	0.368	2.059
AZ	0.677	0.701	0.622	NV	0.404	0.469	5.740*
AR	0.632	0.585	0.879	NH	0.413	0.520	3.237*
CA	0.489	0.586	32.265*	NJ	0.409	0.566	18.679*
CO	0.601	0.646	1.375	NM	0.704	0.729	0.400
CT	0.441	0.494	1.212	NY	0.483	0.604	11.394*
DE	0.543	0.647	3.188*	NC	0.521	0.591	3.256*
DC	0.667	0.743	3.216*	ND	0.481	0.579	0.648
FL	0.514	0.662	42.901*	OH	0.355	0.547	18.499*
GA	0.524	0.595	3.667*	OK	0.503	0.607	3.705*
ID	0.654	0.599	0.698	OR	0.395	0.356	0.922
IL	0.450	0.545	5.784*	PA	0.390	0.560	17.076*
IN	0.454	0.508	1.029	RI	0.279	0.480	7.796*
IA	0.446	0.454	0.008	SC	0.396	0.508	5.527*
KS	0.450	0.497	0.666	SD	0.479	0.472	0.005
KY	0.375	0.593	12.022*	TN	0.500	0.594	3.693*
LA	0.497	0.597	2.027	TX	0.618	0.684	5.549*
ME	0.520	0.389	2.352	UT	0.613	0.538	1.788
MD	0.337	0.544	33.882*	VT	0.378	0.352	0.119
MA	0.380	0.447	3.886*	VA	0.458	0.588	14.204*
MI	0.354	0.492	6.548*	WA	0.434	0.465	0.633
MN	0.428	0.478	0.536	WV	0.530	0.525	0.005
MS	0.382	0.543	5.312*	WI	0.247	0.471	9.362*
MO	0.470	0.471	0.000	WY	0.334	0.330	0.019

Notes: Hausman F-test presented in columns 4 and 8. * indicates a significant difference between OLS and corrected estimates at a 5% level. Estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. Corrected estimates use Random Forest estimated migration probabilities.

Table B.12: PDL Estimates versus Dahl Estimates for all States: Less than High School Premia

State	PDL	DAHL	State	PDL	DAHL
AL	-0.197	-0.210	MT	-0.143	-0.156
AK	-0.155	-0.161	NE	-0.176	-0.145
AZ	-0.192	-0.203	NV	-0.150	-0.139
AR	-0.191	-0.207	NH	-0.146	-0.144
CA	-0.123	-0.152	NJ	-0.124	-0.167
CO	-0.170	-0.180	NM	-0.145	-0.157
CT	-0.177	-0.179	NY	-0.193	-0.202
DE	-0.128	-0.142	NC	-0.096	-0.138
DC	0.286	0.183	ND	-0.210	-0.196
FL	-0.127	-0.157	OH	-0.220	-0.249
GA	-0.168	-0.175	OK	-0.193	-0.242
ID	-0.236	-0.187	OR	-0.150	-0.166
IL	-0.153	-0.218	PA	-0.153	-0.186
IN	-0.224	-0.236	RI	-0.113	-0.171
IA	-0.131	-0.173	SC	-0.128	-0.138
KS	-0.196	-0.199	SD	-0.078	-0.084
KY	-0.200	-0.271	TN	-0.236	-0.224
LA	-0.226	-0.251	TX	-0.161	-0.183
ME	-0.201	-0.157	UT	-0.263	-0.220
MD	-0.106	-0.163	VT	-0.169	-0.178
MA	-0.185	-0.194	VA	-0.143	-0.154
MI	-0.174	-0.184	WA	-0.136	-0.138
MN	-0.158	-0.175	WV	-0.198	-0.220
MS	-0.179	-0.175	WI	-0.121	-0.213
MO	-0.213	-0.233	WY	-0.201	-0.203

Notes: All estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. PDL estimates use Random Forest estimated migration probabilities and a single control function for each birth state. DAHL estimates use cell estimated probabilities and a separate control function for movers and stayers.

Table B.13: PDL Estimates versus Dahl Estimates for all States: Some College Premia

State	PDL	DAHL	State	PDL	DAHL
AL	0.044	0.062	MT	-0.045	-0.056
AK	0.066	0.039	NE	0.115	0.087
AZ	0.133	0.120	NV	0.072	0.063
AR	0.073	0.067	NH	0.109	0.095
CA	0.108	0.125	NJ	0.066	0.063
CO	0.120	0.106	NM	0.112	0.095
CT	0.077	0.035	NY	0.129	0.111
DE	0.123	0.089	NC	0.094	0.108
DC	0.180	0.116	ND	0.120	0.103
FL	0.105	0.090	OH	0.097	0.053
GA	0.110	0.087	OK	0.077	0.078
ID	0.061	0.090	OR	0.086	0.047
IL	0.110	0.057	PA	0.053	0.038
IN	0.088	0.032	RI	0.070	0.030
IA	0.039	0.019	SC	0.083	0.095
KS	0.030	0.031	SD	0.104	0.062
KY	0.074	0.048	TN	0.111	0.084
LA	0.060	0.068	TX	0.113	0.116
ME	0.097	0.108	UT	0.088	0.075
MD	0.095	0.056	VT	0.150	0.105
MA	0.067	0.050	VA	0.064	0.061
MI	0.115	0.092	WA	0.069	0.047
MN	0.081	0.041	WV	0.226	0.134
MS	0.102	0.107	WI	0.066	0.014
MO	0.065	0.010	WY	0.047	0.065

Notes: All estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. PDL estimates use Random Forest estimated migration probabilities and a single control function for each birth state. DAHL estimates use cell estimated probabilities and a separate control function for movers and stayers.

Table B.14: PDL Estimates versus Dahl Estimates for all States: College Premia

State	PDL	DAHL	State	PDL	DAHL
AL	0.370	0.430	MT	0.254	0.225
AK	0.364	0.340	NE	0.350	0.302
AZ	0.498	0.450	NV	0.368	0.363
AR	0.478	0.362	NH	0.322	0.328
CA	0.381	0.410	NJ	0.313	0.326
CO	0.424	0.413	NM	0.444	0.377
CT	0.366	0.268	NY	0.380	0.377
DE	0.332	0.328	NC	0.416	0.399
DC	0.374	0.257	ND	0.404	0.411
FL	0.356	0.399	OH	0.343	0.288
GA	0.377	0.371	OK	0.357	0.383
ID	0.370	0.316	OR	0.366	0.291
IL	0.325	0.277	PA	0.246	0.272
IN	0.374	0.256	RI	0.325	0.301
IA	0.271	0.232	SC	0.384	0.384
KS	0.341	0.319	SD	0.398	0.385
KY	0.310	0.334	TN	0.417	0.389
LA	0.419	0.376	TX	0.469	0.460
ME	0.442	0.381	UT	0.373	0.341
MD	0.289	0.315	VT	0.384	0.322
MA	0.277	0.281	VA	0.352	0.341
MI	0.372	0.310	WA	0.352	0.324
MN	0.306	0.230	WV	0.560	0.412
MS	0.400	0.339	WI	0.245	0.183
MO	0.350	0.255	WY	0.287	0.310

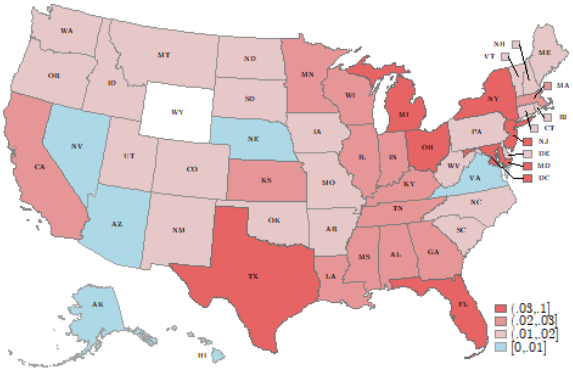
Notes: All estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. PDL estimates use Random Forest estimated migration probabilities and a single control function for each birth state. DAHL estimates use cell estimated probabilities and a separate control function for movers and stayers.

Table B.15: PDL Estimates versus Dahl Estimates for all States: Advanced Degree Premia

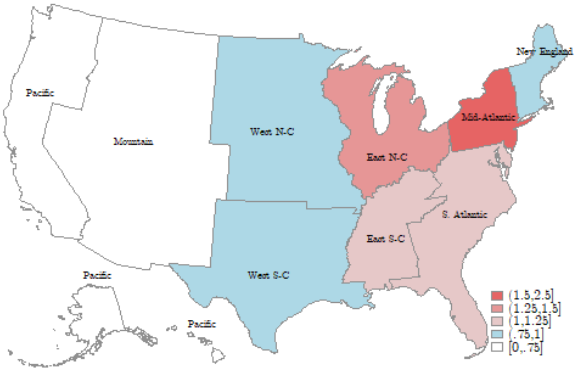
State	PDL	DAHL	State	PDL	DAHL
AL	0.362	0.466	MT	0.492	0.505
AK	0.505	0.566	NE	0.261	0.321
AZ	0.677	0.650	NV	0.404	0.450
AR	0.632	0.545	NH	0.413	0.464
CA	0.489	0.557	NJ	0.409	0.478
CO	0.601	0.593	NM	0.704	0.648
CT	0.441	0.416	NY	0.483	0.531
DE	0.543	0.606	NC	0.521	0.535
DC	0.667	0.567	ND	0.481	0.584
FL	0.514	0.609	OH	0.355	0.393
GA	0.524	0.532	OK	0.503	0.576
ID	0.654	0.592	OR	0.395	0.345
IL	0.450	0.432	PA	0.390	0.424
IN	0.454	0.400	RI	0.279	0.423
IA	0.446	0.343	SC	0.396	0.506
KS	0.450	0.456	SD	0.479	0.466
KY	0.375	0.501	TN	0.500	0.521
LA	0.497	0.525	TX	0.618	0.605
ME	0.520	0.451	UT	0.613	0.555
MD	0.337	0.427	VT	0.378	0.346
MA	0.380	0.397	VA	0.458	0.493
MI	0.354	0.400	WA	0.434	0.452
MN	0.428	0.395	WV	0.530	0.449
MS	0.382	0.463	WI	0.247	0.346
MO	0.470	0.368	WY	0.334	0.380

Notes: All estimates obtained from models run separately for each state, using the 1990 US Census for white males aged 25-34. PDL estimates use Random Forest estimated migration probabilities and a single control function for each birth state. DAHL estimates use cell estimated probabilities and a separate control function for movers and stayers.

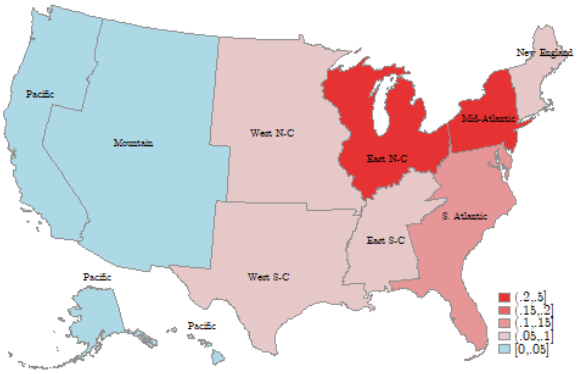
Figure B.3: Florida: Selected Terms and Migration Patterns



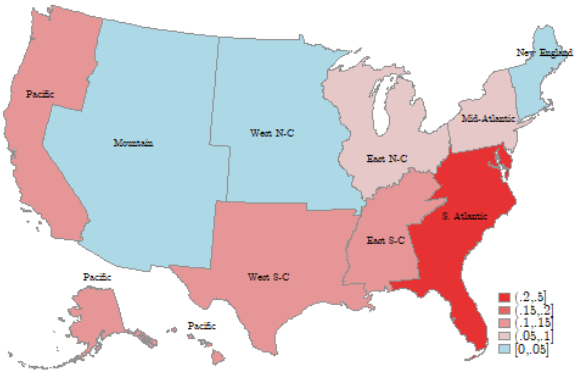
(a) Selected Terms by State



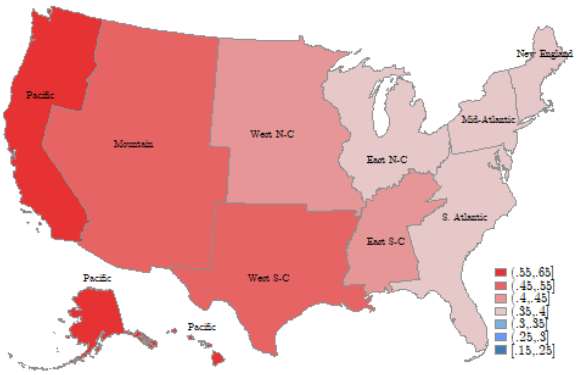
(b) Selected Terms by Division



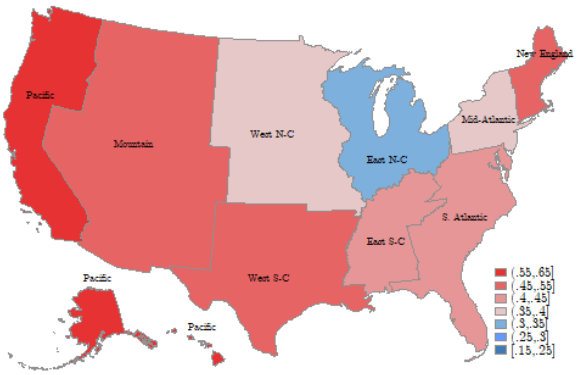
(c) Birthplace Share of Inmigrants



(d) Destination Share of Outmigrants



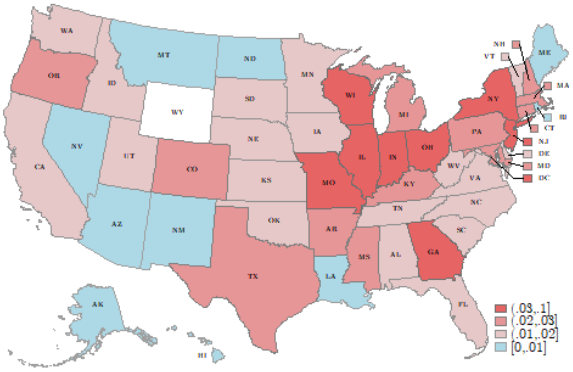
(e) BA share of Inmigrants by Birthplace



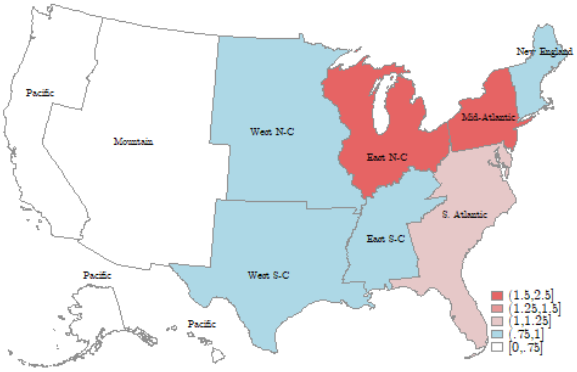
(f) BA share of Outmigrants by Destination

Notes: Panel (a) displays the share of control terms selected by PDL which are calculated using the migration probability from each state. Panel (b) calculates this share at the division level (excluding Florida) and scales by the total share of states in this division. Panel (c) displays the share of immigrants to Florida from each birthplace. Panel (d) shows the destination share of migrants out of Florida. Panel (e) plots the share of immigrants entering Florida from each division who have a College Education. Panel (f) displays the College Educated share of those entering each division who originated in Florida.

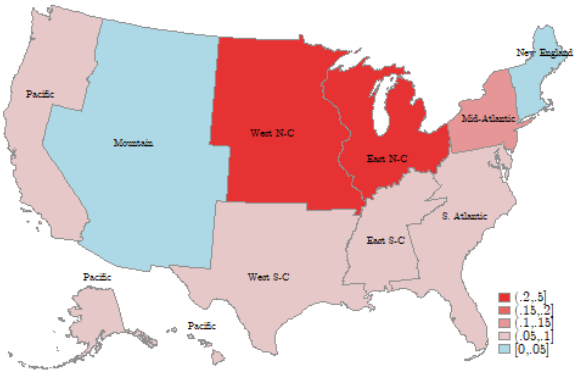
Figure B.4: Illinois: Selected Terms and Migration Patterns



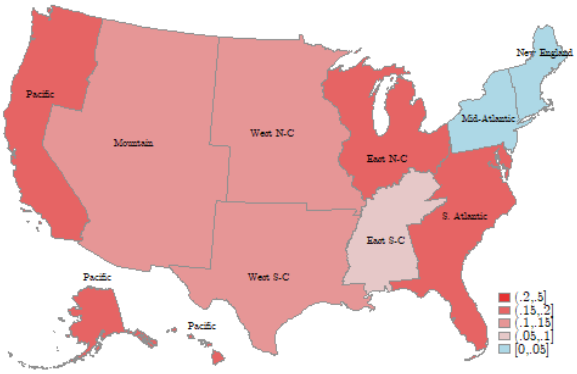
(a) Selected Terms by State



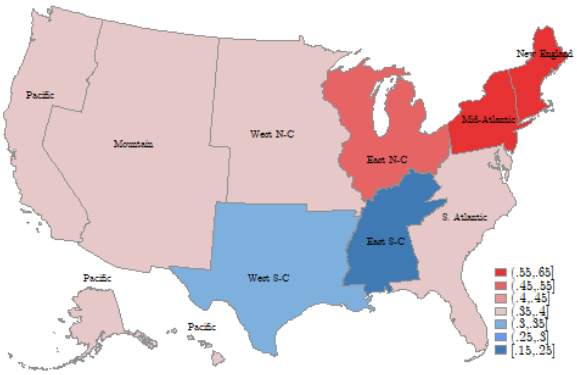
(b) Selected Terms by Division



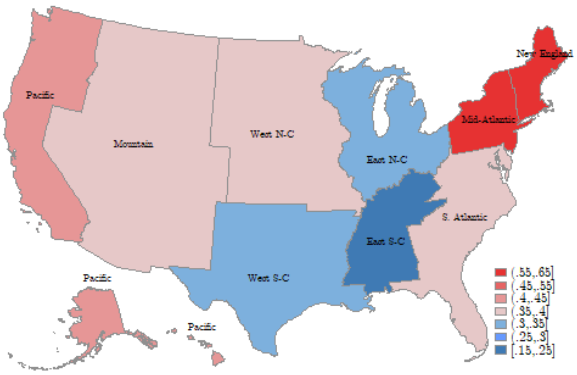
(c) Birthplace Share of Inmigrants



(d) Destination Share of Outmigrants



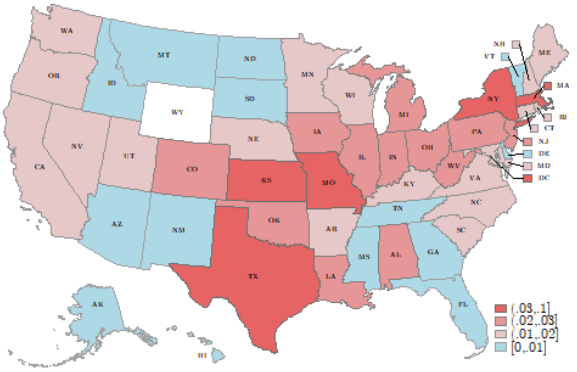
(e) BA share of Inmigrants by Birthplace



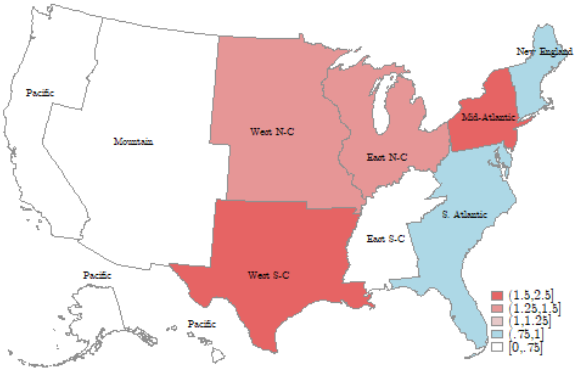
(f) BA share of Outmigrants by Destination

Notes: Panel (a) displays the share of control terms selected by PDL which are calculated using the migration probability from each state. Panel (b) calculates this share at the division level (excluding Illinois) and scales by the total share of states in this division. Panel (c) displays the share of immigrants to Illinois from each birthplace. Panel (d) shows the destination share of migrants out of Illinois. Panel (e) plots the share of immigrants entering Illinois from each division who have a College Education. Panel (f) displays the College Educated share of those entering each division who originated in Illinois.

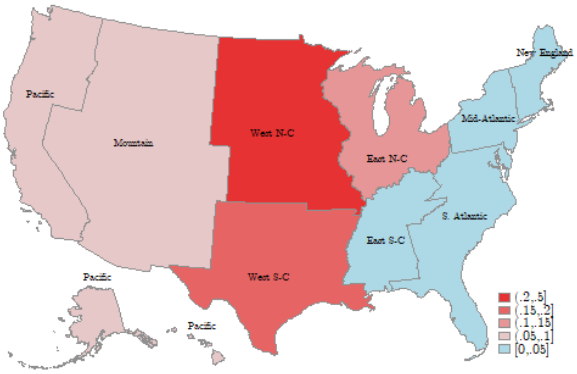
Figure B.5: Kansas: Selected Terms and Migration Patterns



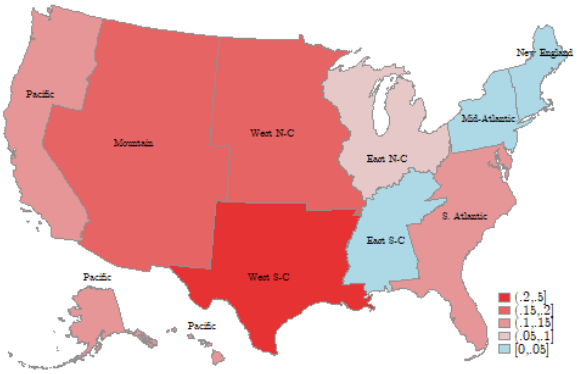
(a) Selected Terms by State



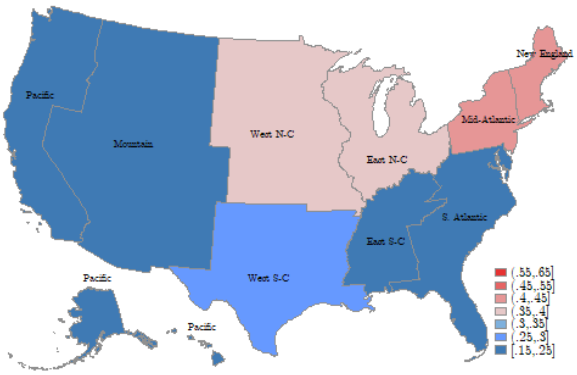
(b) Selected Terms by Division



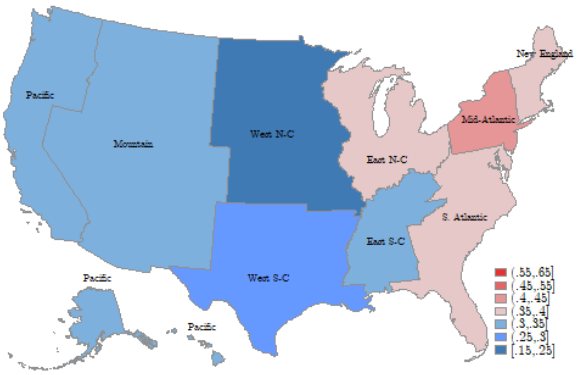
(c) Birthplace Share of Inmigrants



(d) Destination Share of Outmigrants



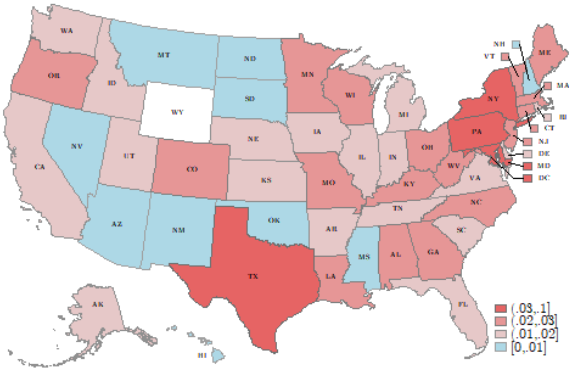
(e) BA share of Inmigrants by Birthplace



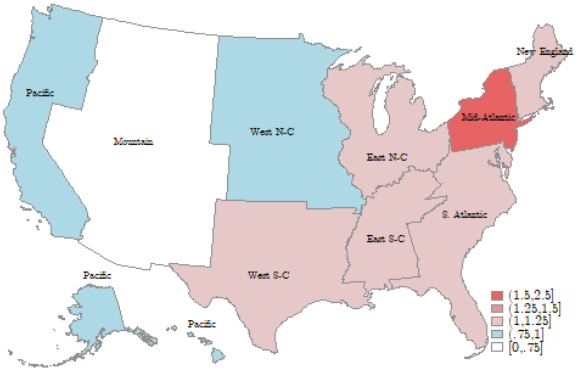
(f) BA share of Outmigrants by Destination

Notes: Panel (a) displays the share of control terms selected by PDL which are calculated using the migration probability from each state. Panel (b) calculates this share at the division level (excluding Kansas) and scales by the total share of states in this division. Panel (c) displays the share of immigrants to Kansas from each birthplace. Panel (d) shows the destination share of migrants out of Kansas. Panel (e) plots the share of immigrants entering Kansas from each division who have a College Education. Panel (f) displays the College Educated share of those entering each division who originated in Kansas.

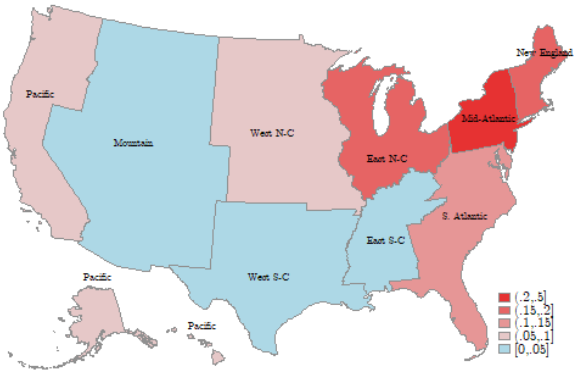
Figure B.6: New York: Selected Terms and Migration Patterns



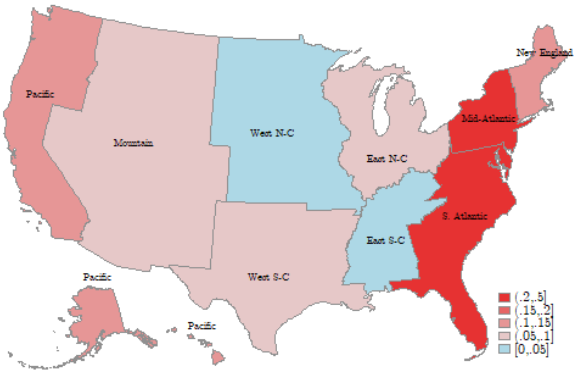
(a) Selected Terms by State



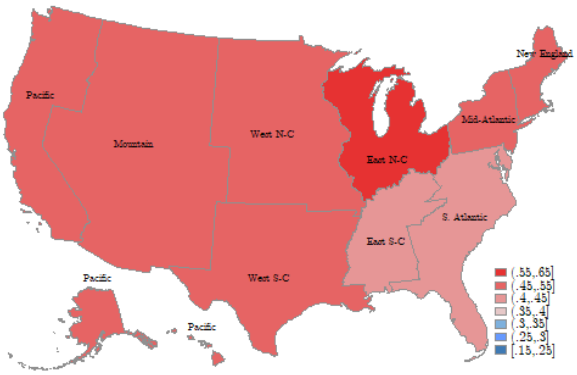
(b) Selected Terms by Division



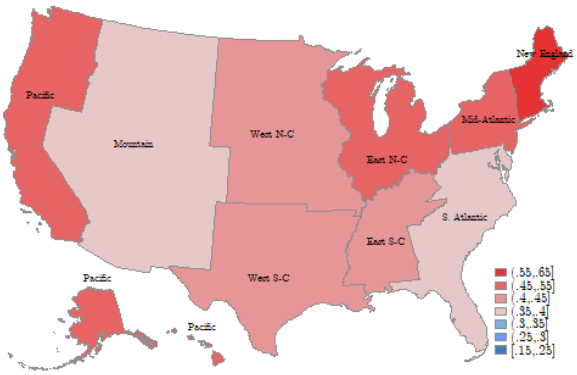
(c) Birthplace Share of Inmigrants



(d) Destination Share of Outmigrants



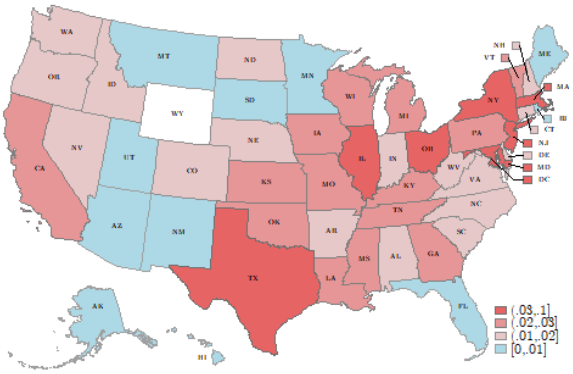
(e) BA share of Inmigrants by Birthplace



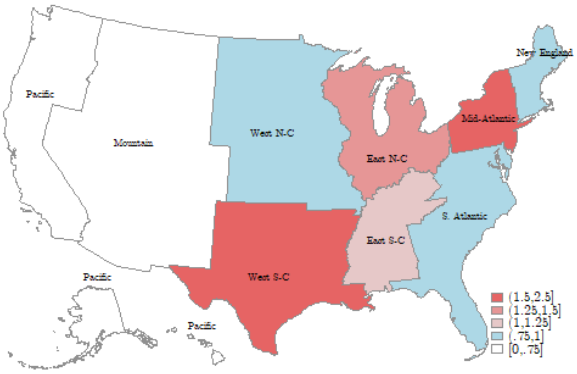
(f) BA share of Outmigrants by Destination

Notes: Panel (a) displays the share of control terms selected by PDL which are calculated using the migration probability from each state. Panel (b) calculates this share at the division level (excluding New York) and scales by the total share of states in this division. Panel (c) displays the share of immigrants to New York from each birthplace. Panel (d) shows the destination share of migrants out of New York. Panel (e) plots the share of immigrants entering New York from each division who have a College Education. Panel (f) displays the College Educated share of those entering each division who originated in New York.

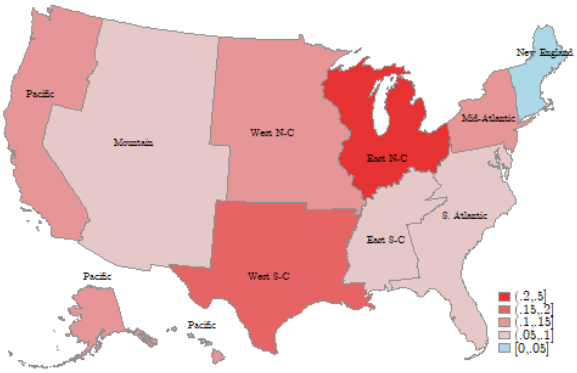
Figure B.7: Texas: Selected Terms and Migration Patterns



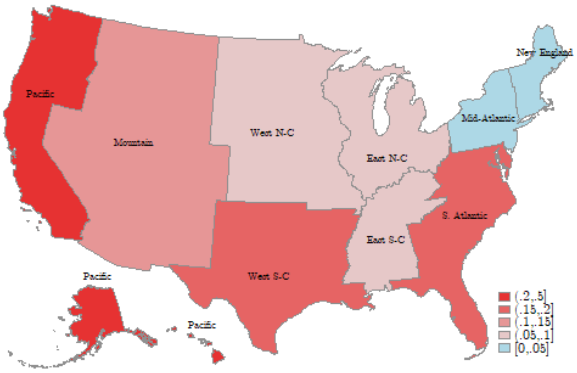
(a) Selected Terms by State



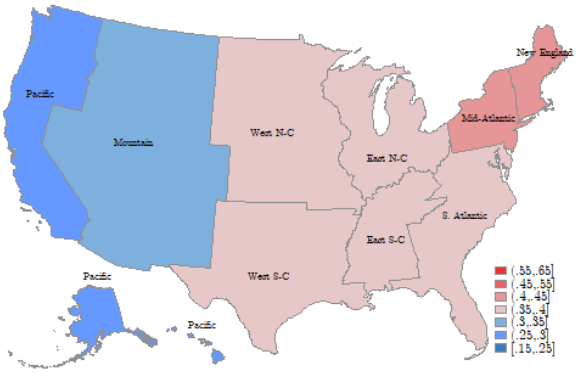
(b) Selected Terms by Division



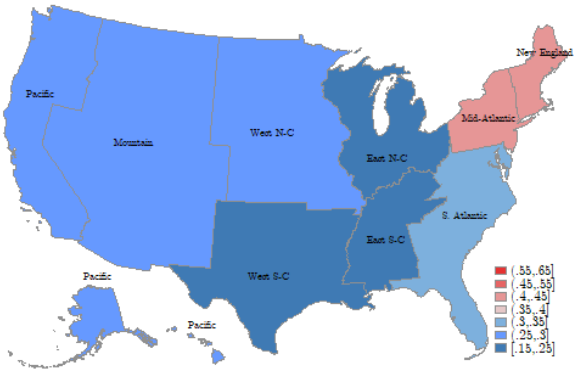
(c) Birthplace Share of Inmigrants



(d) Destination Share of Outmigrants



(e) BA share of Inmigrants by Birthplace



(f) BA share of Outmigrants by Destination

Notes: Panel (a) displays the share of control terms selected by PDL which are calculated using the migration probability from each state. Panel (b) calculates this share at the division level (excluding Texas) and scales by the total share of states in this division. Panel (c) displays the share of immigrants to Texas from each birthplace. Panel (d) shows the destination share of migrants out of Texas. Panel (e) plots the share of immigrants entering Texas from each division who have a College Education. Panel (f) displays the College Educated share of those entering each division who originated in Texas.

Appendix C

Appendix to Chapter 3

C.1 Mathematical Appendix

C.1.1 Derivation of the Firm Surplus

Here we provided details regarding the derivation of the firm surplus under collective bargaining, and individual bargaining.

Collective Bargaining

As noted in text, the surplus from a successful bargain with a union is given by the difference between producing this period with n_u workers (the optimal number of workers with a bargained union wage) and not producing this period along with the cost of rehiring the entire workforce the next period. Noting that the firm has already hired its replacements for workers lost due to normal turnover at the time of the bargaining, this is given by

$$S^u = \left(p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf}(n^u_{icf}) + \rho J^u_{icf}(n^u_{icf})\right) - \left(\pi(0) + \rho J^u_{icf}(0)\right) \tag{C.1}$$

Where $\pi(0) = 0$ corresponds to earning zero profits and $J^u_{icf}(0)$ is the value of a union firm starting the period with no workers. Due to the linear hiring costs, firms will hire back their optimal number of workers,[1] n^u_{ic}, every period, and, as a result, the expression for the value with no workers is:

$$J^u_{icf}(0) = p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf}(n^u_{icf}) - \varkappa \frac{n^u_{icf}}{q_{vc}} + \rho J^u_{icf}(n^u_{icf}) \tag{C.2}$$

[1] Acemoglu and Hawkins (2014) develop a search model with quadratic hiring costs that leads to implications for the firm size distribution and dynamics that we do not consider here.

Substituting this into S^u yields,

$$\begin{aligned} S^u &= \Big(p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf}(n^u_{icf}) + \rho J^u_{icf}(n^u_{icf})\Big) \\ &\quad - \rho\left(p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf}(n^u_{icf}) - \varkappa \frac{n^u_{icf}}{q_{vc}} + \rho J^u_{icf}(n^u_{icf})\right) \\ &= (1-\rho) p_i y_{icf}(n^u_{icf}) - (1-\rho) n^u_{icf} w^u_{icf}(n^u_{icf}) + \rho\varkappa \frac{n^u_{icf}}{q_{vc}} + \rho(1-\rho) J^u_{icf}(n^u_{icf}) \end{aligned} \tag{C.3}$$

Using the definition for $J^u_{icf}(n_u)$, this can be written,

$$\begin{aligned} S^u = (1-\rho) p_i y_{icf}(n^u_{icf}) - (1-\rho) n^u_{icf} w^u_{icf}(n^u_{icf}) + \rho\varkappa \frac{n^u_{icf}}{q_{vc}} \\ + \rho(p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf}(n^u_{icf}) - \varkappa \frac{n^u_{icf}\delta}{q_{vc}}) \end{aligned} \tag{C.4}$$

Simple algebra then yields:

$$S^u = \Big(p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf}(n^u_{icf})\Big) + \rho(1-\delta)\frac{\varkappa n^u_{icf}}{q_{vc}} \tag{C.5}$$

Individual Bargaining

Following TD, we approach the problem of solving for non-union wages in this bargaining environment by first calculating the effect of losing h marginal units of labour and then sending h to zero in order to get the marginal contribution of a single worker. In doing this, we make use of the expression $n-h$ to refer to the removal of h workers from the number of hires n and write the firms surplus from having versus removing h workers as

$$S^n(h) = p_i y_{icf}(n^n_{icf}) - n^n_{icf} w^n_{icf}(n^n_{icf}) + \rho J^n_{icf}(n^n_{icf}) - [\pi(n^n_{icf} - h) + \rho J^n_{icf}(n^n_{icf} - h)] \tag{C.6}$$

with,

$$\pi(n_n - h) = p_i y_{icf}(n^n_{icf} - h) - (n^n_{icf} - h) w^n_{icf}(n^n_{icf} - h) \tag{C.7}$$

$$J^n_{icf}(n^n_{icf} - h) = p_i y_{icf}(n^n_{icf}) - n^n_{icf} w^n_{icf}(n^n_{icf}) - \varkappa \frac{(n^n_{icf} - h)\delta}{q_{vc}} + \rho J^n_{icf}(n^n_{icf}) - \varkappa \frac{h}{q_{vc}} \tag{C.8}$$

Substituting and rearranging yields,

$$S^n(h) = \Big(p_i y_{icf}(n^n_{icf}) - n^n_{icf} w^n_{icf}(n^n_{icf})\Big) - (p_i y_{icf}(n^n_{icf} - h) - (n^n_{icf} - h) w^n_{icf}(n^n_{icf} - h)) + \frac{h\rho(1-\delta)\varkappa}{q_{vc}} \tag{C.9}$$

Dividing by h and taking the limit $\lim_{h\to 0}$ yields the following expression for the firm surplus

$$\lim_{h\to 0} \frac{S^n}{h} = S^n = p_i \frac{\partial y_{icf}(n^n_{icf})}{\partial n} - w^n_{icf}(n^n_{icf}) - n^n_{icf} \frac{\partial w^n_{icf}(n^n_{icf})}{\partial n} + \frac{\rho(1-\delta)\varkappa}{q_{vc}} \tag{C.10}$$

C.1.2 Nonunion Wage Equation Derivation

Plugging the worker and firm surplus into the Nash bargaining condition gives the following:

$$\beta\left(p_i \frac{\partial y_{icf}(n^n_{icf})}{\partial n} - w^n_{icf}(n^n_{icf}) - n^n_{icf} \frac{\partial w^n_{icf}(n^n_{icf})}{\partial n} + \frac{\rho(1-\delta)\varkappa}{q_{vc}}\right) = (1-\beta)\left(w^n_{icf}(n^n_{icf}) + \frac{\rho(1-\delta)w'}{1-\rho(1-\delta)} + \frac{(\rho-1)b}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} + \frac{(\rho-1)\rho q_{uc}}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} E^n_{ic}\right) \tag{C.11}$$

which yields a simple differential equation in wages

$$\begin{aligned} w^n_{icf}(n^n_{icf}) + \beta n^n_{icf} \frac{\partial w^n_{icf}(n^n_{icf})}{\partial n} &= \beta p_i \frac{\partial y_{icf}(n^n_{icf})}{\partial n} + \frac{\beta\rho(1-\delta)\varkappa}{q_{vc}} \\ &\quad - \frac{(1-\beta)\rho(1-\delta)w'_{icf}}{1-\rho(1-\delta)} + \frac{(1-\beta)(1-\rho)b}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} \\ &\quad + \frac{(1-\beta)(1-\rho)\rho q_{uc}}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} E^n_{ic} \end{aligned} \tag{C.12}$$

Solving this expression yields the following wage equation for non-union wages:

$$\begin{aligned} w^n_{icf}(n^n_{icf}) &= \frac{\beta p_i}{1+\beta} \frac{\partial y_{icf}(n^n_{icf})}{\partial n} + \frac{\beta^2}{1+\beta} p_i \epsilon_{ic} + \frac{\beta\rho(1-\delta)\varkappa}{q_{vc}} - \frac{(1-\beta)\rho(1-\delta)w'_{icf}}{1-\rho(1-\delta)} \\ &\quad + \frac{(1-\beta)(1-\rho)b}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} + \frac{(1-\beta)(1-\rho)\rho q_{uc}}{(1-\rho(1-\delta))(1-\rho(1-q_{uc}))} E^n_{ic} \end{aligned} \tag{C.13}$$

In steady state this becomes

$$w^n_{icf}(n^n_{icf}) = \frac{1-\rho(1-\delta)}{1-\beta\rho(1-\delta)}\frac{\beta p_i}{1+\beta}\left(\frac{\partial y_{icf}(n^n_{icf})}{\partial n} + \beta\epsilon_{ic}\right) + \frac{\beta\rho(1-\delta)(1-\rho(1-\delta))}{(1-\beta\rho(1-\delta))}\frac{\varkappa}{q_{vc}} + \frac{(1-\beta)(1-\rho)}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))}b + \frac{(1-\rho)\rho(1-\beta)q_{uc}}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))}E^n_{ic} \tag{C.14}$$

C.1.3 Firm Size Derivation

For union firms, the first order condition is:

$$\frac{\partial J^u_{icf}(n^u_{icf})}{\partial n} = p_i\frac{\partial y_{icf}(n^u_{icf})}{\partial n} - w^u_{icf} - n^u_{icf}\frac{\partial w^u_{icf}(n^u_{icf})}{\partial n} - \frac{\varkappa\delta}{q_{vc}} = 0 \tag{C.15}$$

Using the quadratic production function and the expression for the union wage, this becomes:

$$p_i\left(\epsilon_{ic} - \sigma_i n^u_{icf}\right) - \frac{\beta p_i(1-\rho(1-\delta))}{1-\beta\rho(1-\delta)}(\epsilon_{icf} - \frac{1}{2}\sigma_i n^u_{icf}) - D^u_{icf} + \frac{\beta p_i(1-\rho(1-\delta))}{1-\beta\rho(1-\delta)}\frac{1}{2}\sigma_i n^u_{icf} - \frac{\varkappa\delta}{q_{vc}} = 0 \tag{C.16}$$

where D^u_{icf} contains the elements of the union wage expression that do not vary with n. Rearranging this expression, we arrive at:

$$n^u_{icf} = \frac{1}{\sigma_i p_i}\Big[p_i\epsilon_{ic} + \psi_{icf} - \frac{1-\beta\rho^2(1-\delta)^2}{1-\beta}\frac{\varkappa}{q_{vc}} - \frac{1-\rho}{1-\rho(1-q_{uc})}b - \frac{(1-\rho)\rho q_{uc}}{1-\rho(1-q_{uc})}E^u_{ic}\Big] \tag{C.17}$$

Similarly for nonunion firm size:

$$\frac{\partial J^n_{icf}(n^n_{icf})}{\partial n} = p_i\frac{\partial y_{icf}(n^n_{icf})}{\partial n} - w^n_{icf} - n^n_{icf}\frac{\partial w^n_{icf}}{\partial n} - \frac{\varkappa\delta}{q_{vc}} = 0 \tag{C.18}$$

Using the production function and the nonunion wage expression, this becomes:

$$\begin{aligned} p_i\left(\epsilon_{ic}-\sigma_i n^n_{icf}\right)-\frac{(1-\rho(1-\delta))}{1-\beta\rho(1-\delta)}\frac{\beta p_i}{1+\beta}(\epsilon_{icf}-\sigma_i n^n_{icf})-D^n_{icf} \\ +\frac{(1-\rho(1-\delta))}{1-\beta\rho(1-\delta)}\frac{\beta p_i}{1+\beta}\sigma_i n^n_{icf}-\frac{\varkappa\delta}{q_{vc}}=0 \end{aligned} \tag{C.19}$$

where D^n_{icf} contains the elements of the nonunion wage expression that do not vary with n. Rearranging this expression, we arrive at:

$$\begin{aligned} n^n_{icf}=\frac{1+\beta}{(1+\beta\rho(1-\delta))}\cdot\frac{1}{\sigma_i p_i}\Bigg[p_i\epsilon_{ic}-\frac{(1-\beta\rho^2(1-\delta)^2)}{1-\beta}\frac{\varkappa}{q_{vc}}-\frac{1-\rho}{1-\rho(1-q_{uc})}b \\ -\frac{(1-\rho)\rho q_{uc}}{1-\rho(1-q_{uc})}E^n_{ic}\Bigg] \end{aligned} \tag{C.20}$$

C.1.4 Wage Equation Linearisation

First note that the contribution of firm size to the nonunion and union wage equations is given by:

$$\tilde{w}^n_{icf}(n)=-\frac{\beta}{1+\beta}\frac{1-\rho(1-\delta)}{1-\beta\rho(1-\delta)}\sigma p_i n \tag{C.21}$$

$$\tilde{w}^u_{icf}(n)=-\frac{\beta}{2}\frac{1-\rho(1-\delta)}{1-\beta\rho(1-\delta)}\sigma p_i n \tag{C.22}$$

where, for the firm size contribution to wages is smaller for union wages as $\beta\in(0,1)$.

Plugging firm size into the respective nonunion and union wage equations gives:

$$\begin{aligned} w^n_{icf}=&\frac{\beta(1-\rho(1-\delta))}{(1-\beta\rho(1-\delta))}\cdot\frac{\beta\rho(1-\delta)}{1+\beta\rho(1-\delta)}p_i\epsilon_{ic} \\ &+\frac{1-\rho}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))}\frac{1-\beta^2\rho(1-\delta)}{1+\beta\rho(1-\delta)}b \\ &+\frac{(1-\rho)\rho q_{uc}}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))}\frac{1-\beta^2\rho(1-\delta)}{1+\beta\rho(1-\delta)}E^n_{ic} \\ &+\frac{\beta(1-\rho(1-\delta))}{(1-\beta\rho(1-\delta))(1-\beta)}\frac{1+\rho(1-\delta)(1-\beta-\beta^2\rho(1-\delta))}{1-\beta\rho(1-\delta)}\frac{\varkappa}{q_{uc}} \end{aligned} \tag{C.23}$$

$$
\begin{aligned}
w^u_{icf} = & \frac{\beta(1-\rho(1-\delta))}{2(1-\beta\rho(1-\delta))} p_i \epsilon_{ic} - - \frac{2-\beta-\rho(1-\delta)}{2(1-\beta\rho(1-\delta)} \psi_{icf} \\
& + \frac{1-\rho}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))} \frac{2-\beta-\rho(1-\delta)}{2} b \\
& + \frac{(1-\rho)\rho q_{uc}}{(1-\beta\rho(1-\delta))(1-\rho(1-q_{uc}))} \frac{2-\beta-\rho(1-\delta)}{2} E^u_{ic} \\
& + \frac{\beta(1-\rho(1-\delta))}{(1-\beta\rho(1-\delta))(1-\beta)} \frac{1+\rho(1-\delta)(2(1-\beta)-\beta\rho(1-\delta))}{2} \frac{\varkappa}{q_{uc}}
\end{aligned}
\tag{C.24}
$$

We can conclude that for values of $\beta \in (0,1)$, an increase in industrial prices, or a sectoral productivity shock will have larger increase on union wages. Following BGS, the rates of arrival can be expressed as functions of the city employment rate. We re-write the wage equations making explicit that the coefficients on the key variables are nonlinear functions of the employment rate:

$$
w^n_{icf} = \tilde{\beta}^n_1 p_i \epsilon_{ic} + \tilde{\beta}^n_{2c}(ER_c) b + \tilde{\beta}^n_{2c}(ER_c) \tilde{E}^n_{ic} + \tilde{\beta}^n_{3c}(ER_c) \tag{C.25}
$$

$$
w^u_{icf} = \tilde{\beta}^u_1 p_i \epsilon_{ic} + \tilde{\beta}^u_{2c}(ER_c) b + \tilde{\beta}^u_{2c}(ER_c) \tilde{E}^u_{ic} + \tilde{\beta}^u_{3c}(ER_c) + \tilde{\beta}^u_4 \psi_{icf} \tag{C.26}
$$

where $\tilde{E}^n_{ic}$ and $\tilde{E}^u_{ic}$ are outside options, following BGS expressed in terms of weighted averages over wages.

We take a linear approximation of the wage equations above with respect to the vector $[p_i, ER_c, \epsilon_{ic}, \tilde{E}^n_{ic}]$. We expand around the point where cities have a common industrial structure: $[p, ER, \epsilon, 0]$. Our final linearised wage equations are:

$$
w^n_{ic} = \gamma^n_{0i} + \gamma^n_1 \tilde{E}^n_{ic} + \gamma^n_2 ER_c + \gamma^n_4 \epsilon_{ic} \tag{C.27}
$$

and,

$$
w^u_{ic} = \gamma^u_{0i} + \gamma^u_1 \tilde{E}^u_{ic} + \gamma^u_2 ER_c - \gamma^u_3 \psi_{icf} + \gamma^u_4 \epsilon_{ic} \tag{C.28}
$$

where $\frac{\gamma^u_{0i}}{\gamma^n_{0i}} = \frac{\gamma^u_4}{\gamma^n_4} = K \geq 1$ such that $\gamma^u_{0i} = \gamma^n_{0i} K$ and $\gamma^u_4 = K \gamma^n_4$.

C.1.5 The Firm Wage Response

We can characterize the firm decision on whether to pay this wage and, so, stay nonunion by examining the value of the firm if it pays this wage versus if it pays the union wage. An increase in psi increases profits as wages fall. But less so for emulation firms. There is a cutoff where worker is indifferent.

Thus, the value (in steady state) of the firm if it is unionised is:

$$J^u_{icf} = \frac{1}{1-\rho}\left[p_i y_{icf}(n^u_{icf}) - n^u_{icf} w^u_{icf} - \frac{n^u_{icf}\delta\varkappa}{q_{vc}}\right] \tag{C.29}$$

While the value of the firm if it pays the wage to prevent unionisation is:

$$J^*_{icf} = \frac{1}{1-\rho}\left[p_i y_{icf}(n^*_{icf}) - n^*(w^u_{icf} - \lambda_c + \psi_{icf}) - \frac{n^*_{icf}\delta\varkappa}{q_{vc}}\right] \tag{C.30}$$

where, n^*_{icf} is the optimal firm size when a nonunion firm pays a wage of w^*_{icf}.

Now, take the derivative of both value functions with respect to ψ_{icf}:

$$\frac{\partial J^u_{icf}}{\partial\psi} = \frac{1}{1-\rho}\left[\frac{\partial n}{\partial\psi}\left[p_i\frac{\partial y_{icf}}{\partial n} - w^u_{icf} - n^u_{icf}\frac{\partial w^u_{icf}}{\partial n} - \frac{\delta\varkappa}{q_{vc}}\right] - n^u_{icf}\frac{\partial w^u_{icf}}{\partial\psi}\right] \tag{C.31}$$

and,

$$\frac{\partial J^*_{icf}}{\partial\psi} = \frac{1}{1-\rho}\left[\frac{\partial n}{\partial\psi}\left[p_i\frac{\partial y_{icf}}{\partial n} - w^*_{icf} - \frac{\delta\varkappa}{q_{vc}}\right] - n^*_{icf}\frac{\partial w^u_{icf}}{\partial\psi} - n^*_{icf}\right] \tag{C.32}$$

Note that in both cases, the term in brackets multiplying $\frac{\partial n}{\partial\psi}$ is the first order condition associated with choosing n and, so, equals zero. As a result:

$$\frac{\partial J^u_{icf}}{\partial\psi} = \frac{1}{1-\rho}\left[-n^u_{icf}\frac{\partial w^u_{icf}}{\partial\psi}\right] \tag{C.33}$$

and,

$$\frac{\partial J^*_{icf}}{\partial\psi} = \frac{1}{1-\rho}\left[-n^*_{icf}\frac{\partial w^u_{icf}}{\partial\psi} - n^*_{icf}\right] \tag{C.34}$$

$\frac{\partial J^u_{icf}}{\partial\psi}$ is positive since $\frac{\partial w^u_{icf}}{\partial\psi}$ is negative. An increase in ψ reduces the wage that union firms have to pay and their profits increase as a result. But for emulation firms, an increase in ψ requires a one for one increase in the wage they have to paid (partially offset by the fact that the union wage they are trying to emulate has dropped). That is, $\frac{\partial J^*_{icf}}{\partial\psi} < 0$.

Note that at $\psi^u = \lambda_c - (w^u - w^n)$, workers are just indifferent between whether they organize or not. For $\psi_{icf} < \psi^u$, workers will not organize and the firm will be nonunion and will pay the nonunion wage derived earlier, w^n_{icf}. At $\psi_{icf} = \psi^u$, the wage a firm needs to pay to prevent unionisation is just w^n_{icf} and at that wage and associated optimal firm size, the value of the firm is greater than its value unionised. Therefore, at $\psi_{icf} = \psi^u, J^*_{icf} > J^u_{icf}$. With J^u_{icf} rising

and J^*_{icf} declining with increases in ψ, there will be a point, $\tilde{\psi}$, at which $J^*_{icf} = J^u_{icf}$. This arises when $\psi = \lambda_c$ and, therefore, $w^*_{icf} = w^u_{icf}$.

C.1.6 Decomposition of Mean Wage Movement

Consider decomposing the change in the mean log wage for a representative city, c. We can write this as,

$$\Delta w_{ct} = \Delta P^u_{ct} w^u_{ct} + \Delta(1-P^u_{ct}) w^n_{ct} + P^u_{ct-1} \Delta w^u_{ct} + (1-P^u_{ct-1}) \Delta w^n_{ct} \tag{C.35}$$

where, P^u_{ct} is the proportion union in a representative city and the w's correspond to mean log wages. The wages are also for a representative city and, thus, could be captured by the mean wage across all cities.

The first counterfactual is constructed by holding P^u_{ct} constant over time, i.e.,

$$\Delta w^1_{ct} = P^u_{ct-1} \Delta w^u_{ct} + (1-P^u_{ct-1}) \Delta w^n_{ct} \tag{C.36}$$

Then, the difference between Δw_{ct} and Δw^1_{ct} shows the effect of the union composition change - the standard 'between' effect for a union effect decomposition.

We estimate a regression for the nonunion wage change within an industry, i.e.,

$$\Delta w^n_{ict} = \gamma_{it} + \beta_1 \Delta R^n_{ct} + \beta_2 \Delta T_{ict}(R^u_{ct} - R^n_{ct}) + \beta_3 \Delta ER_{ct} \tag{C.37}$$

Relating this to the overall change in the log mean nonunion wage, we can write:

$$\Delta w^n_{ct} = \sum_i \Delta \eta^n_{ict} w^n_{ict} + \sum_i \eta^n_{ict-1} \Delta w^n_{ict} \tag{C.38}$$

where, η^n_{ict} is the proportion of nonunion workers in industry i in city c at time t.

We can also decompose the change in the union expected wage term as (the nonunion expected wage term can be decomposed in an analogous manner):

$$\Delta T_{ict}(R^u_{ct} - R^n_{ct}) = \Delta T_{ict}(R^u_{ct} - R^n_{ct}) + T_{ict-1} \Delta(R^u_{ct} - R^n_{ct}) \tag{C.39}$$

and, finally, we can decompose the rent term as:

$$\Delta R^u_{ct} = \sum_j \Delta \eta^u_{jct} \nu_{jt} + \sum_j \eta^u_{jct-1} \Delta \nu_{jt} \tag{C.40}$$

Now we are ready to talk about the second counterfactual. In this counterfactual, we hold

the probabilities of nonunion workers in each industry finding jobs constant. So,

$$\Delta w_{ct}^{n2} = \sum_i \Delta \eta_{ict}^n w_{ict}^n + \sum_i \eta_{ict-1}^n \Delta w_{ict}^{n2} \tag{C.41}$$

where,

$$\Delta w_{ict}^{n2} = \gamma_{it} + \beta_1 \Delta R_{ct}^n + \beta_2 T_{ict-1} \Delta (R_{ct}^u - R_{ct}^n) + \beta_3 \Delta ER_{ct} \tag{C.42}$$

The overall counterfactual would be:

$$\Delta w_{ct}^2 = P_{ct-1}^u \Delta w_{ct}^u + (1 - P_{ct-1}^u) \Delta w_{ct}^{n2} \tag{C.43}$$

The difference between this counterfactual and the first gives the effect of changes in the probability that nonunion workers could get union jobs, operating through the outside option effects in bargaining. The difference between counterfactual 2 and the actual change gives the total effect of changes in the union proportions (including the probability of accessing a union job).

Next, we can look at the effect of shifts in the industrial structure - the η's. We can look at this in terms of changes in the industry structure for union workers and the change in the industry structure for nonunion workers.

For the union industry structure, if we were to decompose the change in the union wage into between and within industry effects then we could get a counterfactual change in the union wage holding the industry composition constant. When we hold this constant, we would both get a new, counterfactual version of Δw_{ct}^u and also hold constant the η^u's in the nonunion equation.

The latter would imply a counterfactual nonunion wage:

$$\Delta w_{ct}^{n3} = \sum_i \Delta \eta_{ict}^n w_{ict}^n + \sum_i \eta_{ict-1}^n \Delta w_{ict}^{n3} \tag{C.44}$$

where,

$$\Delta w_{ict}^{n3} = \gamma_{it} + \beta_1 \Delta R_{ct}^n + \beta_2 (T_{ict-1}) (\sum_j \eta_{jct-1}^u \Delta \nu_{jt}^u - \Delta R_{ct}^n) + \beta_3 \Delta ER_{ct} \tag{C.45}$$

The overall counterfactual is then:

$$\Delta w_{ct}^3 = P_{ct-1}^u \sum_k \eta_{kct-1}^u \Delta w_{kct}^u + (1 - P_{ct-1}^u) \Delta w_{ct}^{n3} \tag{C.46}$$

The difference between this and counterfactual 2 is the effect of changes in the union indus-

trial structure holding the nonunion structure constant.

In the fourth step, we hold constant the union wage premium $\nu^u_{it} - \nu^n_{it} = \nu^u_{i1} - \nu^n_{i1} = \tilde{\nu}_{i1}$. From this we derive $\hat{\nu}^u_{it} = \nu^n_{it} + \tilde{\nu}_{i1}$ and form:

$$\Delta w^{n4}_{ict} = \gamma_{it} + \beta_1 \Delta R^n_{ct} + \beta_2 (T_{ict-1})(\sum_j \eta^u_{jct-1} \Delta \hat{\nu}^u_{it} - \Delta R^n_{ct}) + \beta_3 \Delta ER_{ct} \tag{C.47}$$

The overall counterfactual is then:

$$\Delta w^4_{ct} = P^u_{ct-1} \sum_k \eta^u_{kct-1} \Delta w^u_{kct} + (1 - P^u_{ct-1}) \sum_k \eta^n_{kct-1} \Delta w^{n4}_{ct} \tag{C.48}$$

The difference between counterfactual 4 and counterfactual 3 shows the added effect of holding the union wage premium constant. The difference between counterfactual 4 and the raw wage shows the total change in wages due to changes in the union sector. The difference between counterfactual 1 and counterfactual 4 shows the contribution of union outside options to wages.

In the fifth step we hold the nonunion industrial structure. That is,we form:

$$\Delta w^{n5}_{ct} = \sum_i \eta^n_{ict-1} \Delta w^{n4}_{ict} \tag{C.49}$$

where,

$$\Delta w^{n5}_{ict} = \gamma_{it} + \beta_1 \sum_j \eta^n_{jct-1} \Delta \nu^n_{jt} + \beta_2 (T_{ict-1})(\sum_j \eta^u_{jct-1} \Delta \hat{\nu}^u_{it} - \sum_j \eta^n_{jct-1} \Delta \nu^n_{jt}) + \beta_3 \Delta ER_{ct} \tag{C.50}$$

The overall counterfactual is then:

$$\Delta w^5_{ct} = P^u_{ct-1} \sum_k \eta^u_{kct-1} \Delta w^u_{kct} + (1 - P^u_{ct-1}) \sum_k \eta^n_{kct-1} \Delta w^{n4}_{ct} \tag{C.51}$$

The difference between counterfactual 5 and counterfactual 4 shows the added effect of holding the nonunion industrial structure constant.

In the next step, we hold constant changes in the nonunion wage premia, forming:

$$\Delta w^{n6}_{ct} = \sum_i \eta^n_{ict-1} \Delta w^{n6}_{ict} \tag{C.52}$$

where,

$$\Delta w^{n6}_{ict} = \beta_3 \Delta ER_{ct} \tag{C.53}$$

The overall counterfactual is then:

$$\Delta w^{6}_{ct} = P^{u}_{ct-1} \sum_{k} \eta^{u}_{kct-1} \Delta w^{u}_{kct} + (1 - P^{u}_{ct-1}) \sum_{k} \eta^{n}_{kct-1} \Delta w^{n6}_{ct} \tag{C.54}$$

That is, all that is left driving the counterfactual wage change is the effect of changes in the ER and union wages. In step 7 we fix the employment rate, and in step 8 we fix industry-city union wages.

C.2 Data Appendix

Table C.1: SMSA Rankings

1980 Rank	SMSA	1980 Rank	SMSA
1	New York, NY	23	Patterson-Clifton-Passaic, NJ
2	Los Angeles-Long Beach, CA	24	San Diego, CA
3	Chicago, IL	25	Buffalo, NY
4	Philadelphia, PA	26	Miami, FL
5	Detroit, MI	27	Kansas City, MO, KS
6	San Francisco-Oakland, CA	28	Denver, CO
7	Washington, DC, MD, VA	29	San Bernardno-Riverside-Ontario, CA
8	Boston, MA	30	Indianapolis, IN
9	Nassau-Suffolk, NY	31	San Jose, CA
10	Pittsburgh, PA	32	New Orleans, LA
11	St Louis, MO, IL	33	Tampa- St Petersburg, FL
12	Baltimore, MD	34	Portland, OR
13	Cleveland, OH	35	Columbus, OH
14	Houston, TX	36	Rochester, NY
15	Newark, NJ	37	Sacramento, CA
16	Minneapolis-St Paul, MN	38	Birmingham, AL
17	Dallas-Fort Worth, TX	39	Albany-Schenectady-Troy, NY
18	Seattle-Everett, WA	40	Norfolk-Portsmouth, VA
19	Anaheim-Santa Ana-, Garden Grove, CA	41	Akron, OH
		42	Gary-Hammond-East Chicago, IN
20	Milwaukee, WI	43	Greensboro-Winston-Salem- High Point, NC
21	Atlanta, GA		
22	Cincinnati, OH		

Notes: SMSAs consistently available from 1978-2010, ranked by population size in 1980.

Table C.2: Changes to SMSA Definitions 1973-2010

	1973-1980		1981-1989	1993-2003	2004-2010
Chicago	Cook	Lake	Kendall Added		
	Du Page	McHenry	Grundy Added	Dekalb Added	
	Kane	Will			
Philadelphia	Burlington	Chester		Salem Added	
	Camden	Delaware			
	Gloucester	Montgomery			
	Bucks	Philadelphia			
Detroit	Lapeer	Oakland	Monroe Added	Lenawee Added	
	Livingston	St.Clair		Washtenaw Added	
	Macomb	Wayne			
Washington	District of Columbia	Arlington	Calvert Added	Fauquier Added	King George Dropped
	Montgomery	Fairfax	Charles Added	Clarke & Warren Added	Rappahannock Added
	Prince George's	Fairfax city	Frederick Added	Culpeper Added	
	Alexandria	Falls Church	Loudoun Added	King George Added	
			Prince William Added	Spotsylvania Added	
			Masassas Added	Jefferson Added	
			Masassas Park Added	Fredericksburg Added	
			Stafford Added	Berkeley Added	
Boston	Essex	Plymouth	Bristol Added		Bristol Dropped
	MiddleSex	Suffolk			Essex Dropped
	Norfolk		Worchester Added		
Pittsburgh	Allegheny	Washington	Fayette Added	Butler Added	Armstrong Added
	Beaver	Westmoreland			
St Louis	Clinton	Jefferson	Jersey Added	Lincoln Added	Macoupin Added
	Madison	St. Charles		Warren Added	Bond Added
	Monroe	St. Louis			Calhoun Added
	St. Clair	St. Louis city			
	Franklin				
Baltimore	Anne Arundel	Carroll	Queen Anne's Added		
	Baltimore city	Harford			
	Baltimore	Howard			
Cleveland	Cuyahoga	Lake		Added Ashtabula	
	Geauga	Medina		Added Lorain	
Houston	Brazoria	Liberty		Added Chambers	Added Austin
	Fort Bend	Montgomery			Added Galveston
	Harris	Waller			

Newark	Essex	Sussex			Union Dropped
	Morris	Union			
Minneapolis-	Anoka	Ramsey	Isanti Added	Sherburne Added	
St Paul	Carver	Scott			
	Chisago	Washington			
	Dakota	Wright			
	Hennepin				
Dallas-	Collin	Wise	Wise Dropped	Henderson Added	Wise Added
Fort Worth	Dallas	Hood	Hood Dropped	Hunt Added	Somerwell Added
	Denton	Johnson		Hood Added	
	Ellis	Tarrant			
	Kaufman	Parker			
	Rockwall				
Seattle-Everett	King	Snohomish		Island Added	Pike Added
Atlanta	Cherokee	Gwinnett	Barrow Added	Butts dropped	Butts Added
	Clayton	Henry	Coweta Added	Carroll Added	Dawson Added
	Cobb	Newton	Spalding Added	Bartow Added	Haralson Added
	De Kalb	Paulding			Heard Added
	Douglas	Rockdale			Jasper Added
	Fayette	Walton			Lamar Added
	Forsyth	Butts			Meriwether Added
	Fulton				Morgan Added
Cincinnati	Dearborn	Clermont		Ohio Added	Union Added
	Boone	Hamilton		Gallatin Added	Bracken Added
	Campbell	Warren		Grant & Brown Added	Butler Added
	Kenton			Pendelton Added	
Kansas City	Johnson	Jackson	Lafayette Added	Clinton Added	Linn Added
	Wyandotte	Platte	Leavenworth Added		Bates Added
	Cass	Ray	Miami Added		Caldwell Added
	Clay				
Denver	Adams	Denver			Adams Dropped
	Arapahoe	Douglas			Broomfield Added
	Boulder	Jefferson			Clear Creek Added
					Elbert & Park Added
					Gilpin Added
Indianapolis	Boone	Johnson			Brown Added
	Hamilton	Marion			Putnam Added
	Hancock	Morgan			

	Hendricks	Shelby			
New Orleans	Jefferson	St. Bernard	St Charles Added	St James Added	
	Orleans	St.Tammany	St John the Bap. Added	Plaquemines Added	
Tampa-	Hillsborough	Pinellas	Hernando Added		
St Petersburg	Pasco				
Portland	Clackamas	Washington		Clark Added	
	Multnomah	Yamhill		Columbia Added	
Columbus	Delaware	Madison	Licking Added	Licking Dropped	Licking Added
	Fairfield	Pickaway	Union Added		Hocking Added
	Franklin				Morrow Added
Rochester	Livingston	Orleans		Genesee Added	
	Monroe	Wayne			
	Ontario				
Sacramento	Placer	Yolo		El Dorado Added	
	Sacramento				
Birmingham	Jefferson	Walker	Blount Added	Walker Dropped	Walker Added
	Shelby	St. Clair			Bibb & Chilton Added
Albany-	Albany	Schenectady	Greene Added	Greene Dropped	
Schenectady-Troy	Rensselaer	Montgomery		Schoharie Added	
	Saratoga				
Norfolk-	Currituck	Portsmouth	Currituck Dropped	Currituck Added	
Portsmouth	Chesapeake	Virginia Beach	Gloucester Added	Isle of Wight Added	
	Norfolk		Hampton & Suffolk Added	Mathews Added	Gloucester Added
			James & York Added		
			Newport News Added		
			Poquoson Added		
			Williamsburg Added		
Greensboro-	Forsyth	Yadkin	Davie Added	Alamance Added	Alamance Dropped
Winston-Salem-	Guilford	Stokes			
High point	Randolph	Davidson			
Gary-Hammond	Lake	Porter			Jasper Added
East Chicago					Newton Added
Portland	Clackamas	Multnomah		Columbia Added	
	Washington	Yamhill			

Notes: Changes to the counties/cities/parishes, included in the SMSA definitions over the sample period. There are no county changes for New York, Patterson, Nassau-Suffolk, Los Angeles, San Francisco, Anaheim, Milwaukee, San Diego, Buffalo, Miami, San-Bernadino, San Jose, Akron.

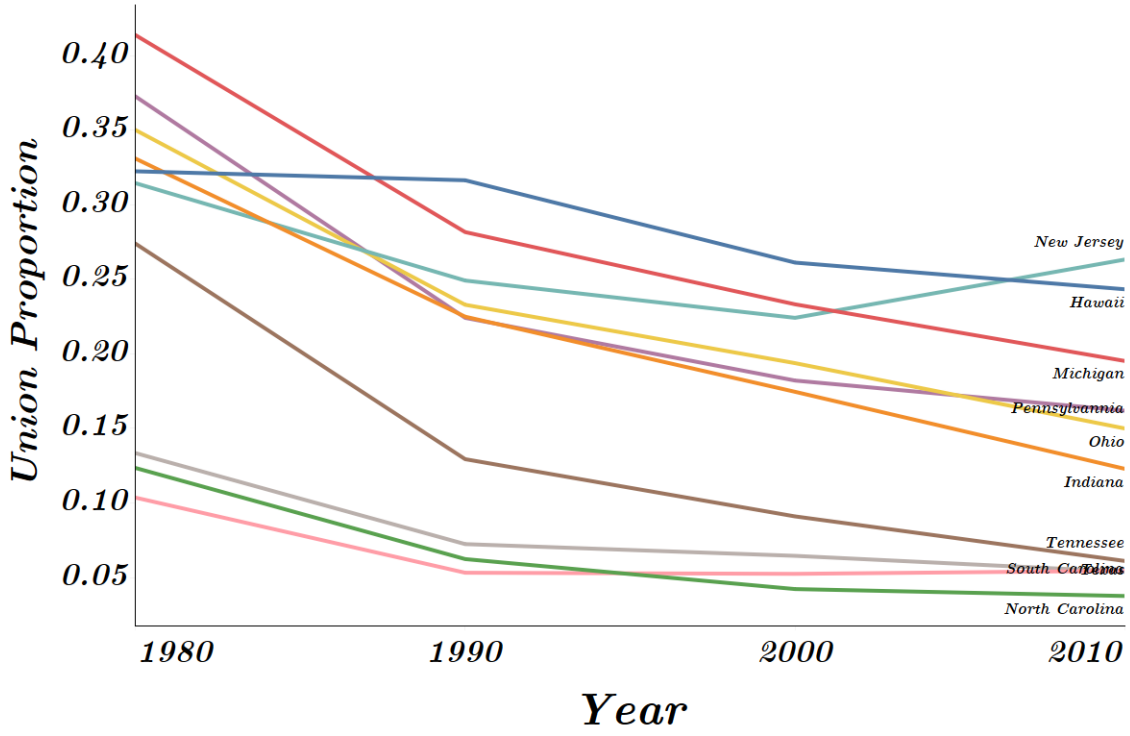
Table C.3: Aggregated Industry Definitions

Category	Code	1990 Industry Codes
Agriculture Service	1	12, 20, 21 , 30
Other Agriculture	2	10 - 11
Mining	3	40 - 50
Construction	4	60
Lumber and Wood Products, except Furniture	5	230 - 241
Furniture and Fixtures	6	242
Stone Clay, Glass, and Concrete Product	7	250 - 262
Primary Metals	8	270 - 280
Fabricated Metal	9	281 - 300
Not Specified Metal Industries	10	301
Machinery, except Electrical	11	310 - 332
Electrical Machinery, Equipment, and Supplies	12	340 - 350
Motor Vehicles and Equipment	13	351
Aircraft and Parts	14	352
Other Transportation Equipment	15	360 - 370
Professional and Photographic Equipment, and Watches	16	371 - 382
Toys, Amusements, and Sporting Goods	17	390
Miscellaneous and Not Specified Manufacturing Industries	18	391 - 392
Food and Kindred Products	19	100 - 122
Tobacco Manufactures	20	130
Textile Mill Products	21	132 - 150
Apparel and Other Finished Textile Products	22	151 - 152
Paper and Allied Products	23	160 - 162
Printing, Publishing and Allied Industries	24	171 - 172
Chemicals and Allied Products	25	180 - 192
Petroleum and Coal Products	26	200 - 201
Rubber and Miscellaneous Plastics Products	27	210 - 212
Leather and Leather Products	28	220 - 222
Transportation	29	400 - 432
Communications	30	440 - 442
Utilities and Sanitary Services	31	450 - 452, 460 - 472
Wholesale Trade	32	500 - 571
Retail Trade	33	580 - 691
Banking and Other Finance	34	700 - 710
Insurance and Real Estate	35	711 - 712
Private Household Services	36	761
Business Services	37	721, 722, 731 - 750, 892
Repair Services	38	751 - 760
Personal Services, except Private Household	39	762 - 791
Entertainment and Recreation Services	40	800 - 802, 810
Hospitals	41	831
Health Services, except Hospitals	42	812 - 830, 832 - 840
Educational Services	43	842 - 860
Social Services	44	861 - 871
Other Professional Services	45	730, 841, 872 - 891, 893
Forestry and Fisheries	46	31 - 32
Justice, Public Order and Safety	47	910
Administration Of Human Resource Programs	48	922
National Security and Internal Affairs	49	932
Other Public Administration	50	900, 901, 921, 930, 931

Notes: List of aggregated industries and corresponding 1990 codes used by the US Census Bureau.

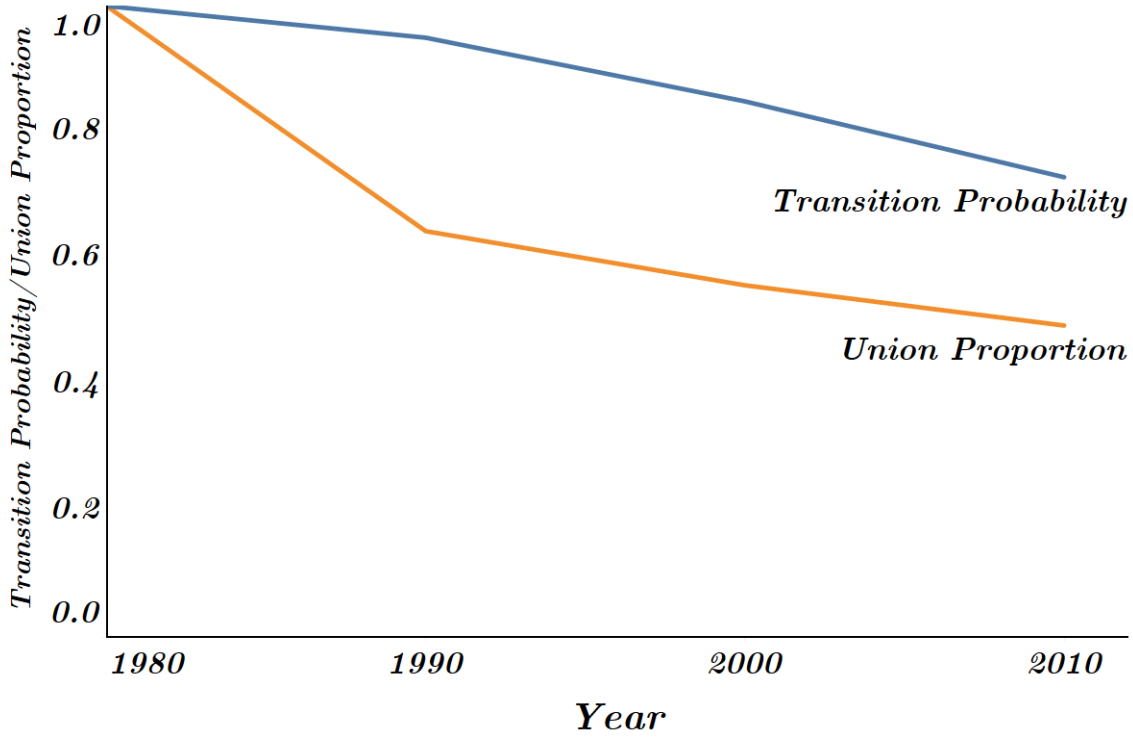
C.3 Tables and Figures

Figure C.1: Declining Unionisation Across Selected States



Notes: The proportion of unionised workers is plotted from 1980-2010 for selected states. States with the largest, and smallest decline over 1980-2010 are presented.

Figure C.2: Percentage Decline in Unionisation and Transitions into Union Jobs



Notes: Transition Probability Calculated at the national level for all workers using linked CPS records. Union proportion calculated at the national level using CPS data.

Table C.4: Alternative Specifications: Alternative Transition Measure

	OLS (1)	OLS (2)	IV (3)
ΔP_{ic}	-0.003		
	(0.018)		
ΔR_c^n	2.31***	2.31***	1.79***
	(0.30)	(0.30)	(0.50)
ΔE_c^{3n}	2.99***	2.99***	3.29***
	(0.44)	(0.44)	(1.01)
ΔER	0.88***	0.88***	0.95***
	(0.18)	(0.18)	(0.21)
Observations	6927	6927	6927
R^2	0.50	0.50	0.50
Year × Ind.	Yes	Yes	Yes
ERIV			No
IVs			all
F-Stats:			
ΔR_c^n			36.22
ΔE_c^{3n}			22.18
p-val:			
ΔR_c^n			0.00
ΔE_c^{3n}			0.00
	Test $\gamma_{11}^n = \gamma_{12}^n$		
p-val:	.11	.11	.20
F-Stat.	2.62	2.64	1.65

Notes Standard errors in parentheses clustered at the city-year level. * denotes significance at the 10% level, ** denotes significance at the 5% level, *** denotes significance at the 1% level. The dependent variable is the change in the regression adjusted average hourly wage of nonunion workers in an industry-city cell. Estimates obtained using decadal changes over 1980-2010 across 50 industries and 93 cities.

Table C.5: Alternative Specifications - Public Sector

	Modification 1		Modification 2	
	OLS	IV	OLS	IV
ΔR_c^n	2.04***	1.66**	2.18***	1.81***
	(0.30)	(0.55)	(0.27)	(0.41)
ΔE_{ic}^n	3.69***	2.64***	3.50***	3.03***
	(0.64)	(1.13)	(0.65)	(1.02)
ΔER	1.08***	1.18***	1.05***	1.13***
	(0.18)	(0.21)	(0.18)	(0.20)
Observations	6349	6349	6349	6349
R^2	0.50	0.50	0.50	0.49
Year × Ind.	Yes	Yes	Yes	Yes
ERIV		No		No
IVs		all		all
F-Stats:				
ΔR_c^n		38.65		40.64
ΔE_{ic}^n		81.66		98.88
p-val:				
ΔR_c^n		0.00		0.00
ΔE_{ic}^n		0.00		0.00
		Test $\gamma_{11}^n = \gamma_{12}^n$		
p-val.	.01	.49	.04	.29
F-Stat.	6.46	.49	4.34	1.14

Notes Standard errors in parentheses clustered at the city-year level. * denotes significance at the 10% level, ** denotes significance at the 5% level, *** denotes significance at the 1% level. The dependent variable is the change in the regression adjusted average hourly wage of nonunion workers in an industry-city cell. Estimates obtained using decadal changes over 1980-2010 across 50 industries and 93 cities. In 'Modification 1' we drop $i - c$ cells in the public sector, but we construct rents using public sector wage premia. In 'Modification 2' we omit the public sector in the construction of rents.