# Bayesian Propensity Score Analysis for Observational Data

by

Lawrence Cruikshank McCandless

B.Sc., The University of British Columbia, 2000
M.Sc., The University of British Columbia, 2004

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Statistics)

The University of British Columbia

August 2007

# Abstract

Propensity scores analysis (PSA) involves regression adjustment for the estimated propensity scores, and the method can be used for estimating causal effects from observational data. However, confidence intervals for the treatment effect may be falsely precise because PSA ignores uncertainty in the estimated propensity scores. We propose Bayesian propensity score analysis (BPSA) for observational studies with a binary treatment, binary outcome and measured confounders. The method uses logistic regression models with the propensity score as a latent variable. The first regression models the relationship between the outcome, treatment and propensity score, while the second regression models the relationship between the propensity score and measured confounders. Markov chain Monte Carlo is used to study the posterior distribution of the exposure effect. We demonstrate BPSA in an observational study of the effect of statin therapy on all-cause mortality in patients discharged from Ontario hospitals following acute myocardial infarction. The results illustrate that BPSA and PSA may give different inferences despite the large sample size. We study performance using Monte Carlo simulations. Synthetic datasets are generated using competing models for the outcome variable and various fixed parameter values. The results indicate that if the outcome regression model is correctly specified, in the sense that the outcome risk within treatment groups is a smooth function of the propensity score, then BPSA permits more efficient estimation of the propensity scores compared to PSA. BPSA exploits prior information about the relationship be-

tween the outcome variable and the propensity score. This information is ignored by PSA. Conversely, when the model for the outcome variable is misspecified, then BPSA generally performs worse than PSA.

# Table of Contents

## Appendices

# List of Tables

# List of Figures

# Acknowledgments

I gratefully acknowledge the support of my thesis supervisory committee Paul Gustafson, John Petkau, Rollin Brant and Peter Austin for their time and guidance. I am indebted to Paul Gustafson for his mentorship, insights and wisdom in helping me to direct of my thesis research and pursue an academic career. I especially thank John Petkau for his detailed readings of my thesis drafts, and also for his insight in navigating the job application process. Additionally, I would like to thank Peter Austin for his time and access to the EFFECT data. My thesis research is supported by training awards from the Michael Smith Foundation for Health Research. The EFFECT study is funded by a Canadian Institutes of Health Research Team Grant in Cardiovascular Outcomes Research. Additional funding for this project came from an operating grant (Grant No. NA 5703) from the Heart and Stroke Foundation of Ontario. Lastly, I would like to thank my family, friends and fellow graduate students for their support during my time at UBC.

# Dedication

To Megan

# Chapter 1

# Introduction

The propensity score, defined as the probability of treatment given measured confounders, can be used as a tool to control confounding bias in observational studies [1, 2]. Patients with fixed propensity scores have identical distributions of measured confounders, irrespective of whether they are treated or not [1]. Thus when there are no unmeasured confounders, comparing patient outcomes between treatment groups conditional on the propensity score gives unconfounded estimates of causal treatment effects. Sampling patients with the same propensity score replicates a miniature randomized trial within a subset of the study population.

To adjust for confounding, we can include the propensity score in a regression model for the outcome [3]. Because the functional form for the relationship between the outcome and the propensity score within treatment groups is usually unknown, we have a non-parametric regression problem. A common analytic strategy is to break the study sample into homogeneous groups, and then estimate the treatment effect within each group [2]. The groups are often constructed using quintiles of the empirical distribution of the propensity scores. Other adjustment techniques include treating the propensity score as a continuous covariate or fitting a regression using cubic splines [2]. If matched pairs are available then conditional likelihood estimation is often used [2].

In observational studies, the propensity score is unknown and can be interpreted as missing data. A propensity score analysis (PSA) proceeds using a two step procedure.

First we estimate the propensity score for each patient. Then we fit a regression model on the estimated propensity scores. In most applications, standard errors estimates are calculated from asymptotic approximation to the distribution of the maximum likelihood estimator based on the regression model of the outcome on treatment and estimated propensity scores [4]. This approach to interval estimation ignores uncertainty in the estimated propensity scores. Frequentist inferences are reported under hypothetical repeated sampling of patient outcomes conditional on the observed treatments and estimated propensity scores.

If our objective is to report inferences while acknowledging uncertainty in the estimated propensity scores, then intuitively the conventional standard error calculations will yield interval estimates for the treatment effect which are falsely precise. PSA substitutes point estimates for parameters. We can draw parallels with stepwise model selection techniques used in regression modelling. A preliminary analysis selects one model among many competing models. The usual confidence intervals are calculated by conditioning on the selected model and they will be too narrow because they do not acknowledge model uncertainty [5].

There is little discussion in the literature about methods for incorporating uncertainty from the estimated propensity scores into uncertainty about the treatment effect. Some exceptions include [4, 6]. In contrast, there a large body of work describing the merits of using estimated propensity scores over true propensity scores [7]. Hahn and others [8–11] argue that treatment effect estimates calculated using the estimated propensity scores are generally more efficient than corresponding estimates calculated from true propensity scores. This has been demonstrated both analytically and in simulations in the context of propensity score weighting adjustments for confounding [8–11]. For matched sampling, Rubin and Thomas [12, 13] demonstrate that matching on the estimated propensity scores generally yields greater similarity in empirical distribution of covariates for treated versus control subjects. Similar findings are reported for subclassification on quintiles of the estimated propensity score versus the true propensity score [3].

In this thesis, we develop a Bayesian approach to modelling uncertainty in the estimated propensity scores in observational studies with a dichotomous treatment, dichotomous outcome and a vector of measured confounders. We propose a Bayesian propensity score analysis (BPSA) which models the joint distribution of the data and parameters with the patient propensity scores as latent variables. Markov chain Monte Carlo (MCMC) is used to sample from the posterior distribution of model parameters. This estimation strategy yields interval estimates for the treatment effect which incorporate uncertainty in the estimated propensity scores.

We show that BPSA has certain advantages compared to PSA. The method uses patient outcome data in order to calculate propensity score estimates. Instead of using a two step procedure which first estimates propensity scores and then estimates treatment effects, BPSA estimates both quantities simultaneously. The MCMC algorithm iteratively imputes the propensity scores and estimates the treatment effect. When imputing the propensity scores, posterior information about the outcome distribution flows through the algorithm in order to inform inferences about the propensity scores. Thus BPSA incorporates prior information about the relationship between the outcome and propensity score within treatment groups. This information is ignored by PSA when estimating propensity scores.

With the exception of two articles [14, 15], there appears to be little research combining propensity score methods and Bayesian statistics. One possible reason is because PSA does not use conventional models for the risk of the outcome variable. The usual approach to adjust for confounding is to work with an assumed model for the risk of the outcome given treatment and measured confounders. Instead PSA uses a model for the outcome given treatment and the propensity score. But the propensity score is not a typical covariate. It is a characteristic of the manner in which treatment and covariates were sampled. While the PSA method produces valid treatment effect estimates from a frequentist standpoint, it makes less sense within the Bayesian paradigm. How one chooses to design an experiment (i.e. the true values of propensity scores) should convey no prior information about the risk of the outcome in a given study subject. Robins and Ritov argue the propensity scores should not appear in the likelihood for the distribution of the outcome variable and are irrelevant in a Bayesian analysis [15]. We elaborate on this in more detail in Section 3.2.

This thesis is organized as follows. In Section 1.1, we describe a case-study to motivate the methodology. We describe an observational study of the effectiveness of statin therapy in Ontario residents following myocardial infarction. In Chapter 2, we review the topic of confounding bias in observational studies, and analytic adjustments for confounding bias using propensity scores. Only observational studies with a binary treatment, binary response and a vector of measured covariates are considered. In Chapter 3, we describe the BPSA method which models the propensity score as a latent variable. We demonstrate that BPSA uses patient outcomes to estimate propensity scores by incorporating prior information about the relationship between the outcome and propensity scores. In Section 3.2, we discuss using BPSA as a tool for causal inference. We argue that while the BPSA model is not a realistic representation of the true data generating process, the method can nonetheless be used to construct Bayesian estimators with good frequentist properties. In Chapter 4 we apply our method to the statin data and compare the results to PSA. We illustrate that contrary to intuition, BPSA and PSA may give very different results even in large samples. In Chapter 5 we use simulations to study the frequentist performance of point estimates, interval estimates and prediction error when applied to both the statin data and simulated data. The results demonstrate that if the outcome regression model is correctly specified, in the sense that the outcome risk within treatment groups is a smooth function of the propensity score, than BPSA permits more efficient estimation of the propensity scores compared to PSA. Conversely, when the model for the outcome variable is misspecified, then simulations show that BPSA will generally perform poorly relative to PSA.

## 1.1 A motivating example: Estimating the effectiveness of statins in patients following acute myocardial infarction.

To motivate the propensity score methodology, we begin by considering the example of an observational study estimating the effectiveness of statin therapy in patients discharged alive from hospital following acute myocardial infarction. Statins are a class of lipid lowering medications which are commonly prescribed to persons with multiple risk factors for cardiovascular disease. Several large studies in the United States and elsewhere have demonstrated that statin therapy reduces morbidity and mortality in patients following myocardial infarction [16–18]. Nonetheless, the magnitude of the effectiveness of statins in large populations is less well understood. Randomized trials typically do not provide estimates of population level effects because they exclude vulnerable populations from study, such as the elderly and very sick. Observational studies indicate that statins may be highly effective in these vulnerable groups [19].

Detailed clinical data were obtained from a random sample of 4572 patients discharged alive from Ontario hospitals with myocardial infarction between April 1, 1999 and March 31, 2000. The data were collected in conjunction with the Enhanced Feedback for Effective Cardiac Treatment (EFFECT) study, an ongoing initiative to improve the quality of health services for Ontario residents with cardiovascular disease [2]. For each patient, medical charts were abstracted to obtain information on demographic characteristics, cardiac risk factors, comorbid conditions, vascular history, vital signs at hospital admission, and laboratory tests. Data on prescriptions for statins were also collected. More details about the dataset are given by Austin and Mamdani [2]. A complete list of the measured covariates is given in Table 1.1.

To estimate the effect of statin therapy on mortality, patients were classified as statin users if they were prescribed a statin at hospital discharge, and they were classified as statin non-users otherwise. Death within three years of hospital discharge was established by linking patient records to the Ontario Registered Persons Database.

Before comparing mortality rates among treated and untreated patients, we describe the study sample. Table 1.1 presents summary statistics for the baseline characteristics of treated and untreated patients. We use t-tests and chi-squared tests to compare the covariate distributions among treated and untreated patients for continuous and dichotomous variables respectively. The results indicate that the treatment groups differ systematically with respect to risk factors for mortality. Patients who were prescribed a statin at hospital discharge were typically younger and healthier than patients without a prescription [2]. These results are consistent with previous studies of physician prescribing habits for cardiovascular disease medications [19, 20]. Physicians tend to avoid giving statins to the elderly or to patients with multiple comorbidities, even though they are often indicated in such populations [19].

The results in Table 1.1 indicate that a crude comparison of mortality among treated and untreated patients may be biased due to confounding. Randomized trials demonstrate that statin therapy reduces mortality in patients following acute myocardial infarction [16]. Consequently, we expect the protective effect of statin therapy on mortality to be exaggerated in a crude comparison of mortality rates. Treated patients have lower mortality because of the benefits of statins and because they are more healthy. The crude odds ratio for the association between statin therapy and mortality can be calculated from the 2 × 2 table:

| | Died | Survived | |
|---|---|---|---|
| Treated | 193 | 1161 | 1354 |
| Untreated | 800 | 2418 | 3218 |
| | 993 | 3579 | 4572 |

and is equal 0.50 with 95% confidence interval (0.42, 0.60). This estimate is far lower than previous estimates from randomized controlled trials or observational studies [16–18]. In light of prior information about physician prescribing habits of statins [19, 20], these data provide evidence that association between statin therapy and mortality is confounded. Analytic adjustments are required in order to account for the unequal distribution of mortality risk factors between treatment groups.

Table 1.1: Baseline characteristics of 4572 patients discharged alive from hospital following acute myocardial infarction.

| Characteristic | Statin prescribed (n=1354) | | Statin not prescribed (n=3218) | |
|---|---|---|---|---|
| | Number (%) or Mean ± SD | | | |
| *Demographic characteristics* | | | | |
| Age (mean) | 63 | ±12** | 68 | ±14** |
| Female sex | 398 | (29)* | 1201 | (37)* |
| *Presenting characteristics*$^+$ | | | | |
| Shock | $\leq 5$ | ($\leq 1$) | 24 | (1) |
| *AMI risk factors*$^+$ | | | | |
| Family history of CAD | 525 | (39)** | 973 | (30)** |
| Diabetes | 459 | (26) | 1060 | (26) |
| CVA/TIA | 122 | (9) | 312 | (10) |
| High BP | 548 | (48)* | 1386 | (43)* |
| Current smoker | 459 | (34) | 1060 | (33) |
| Hyperlipdaemia | 794 | (59)** | 604 | (19)** |
| *Comorbidities*$^+$ | | | | |
| Angina | 504 | (37)** | 999 | (31)** |
| Renal disease | 10 | (1) | 13 | ($\leq 1$) |
| *Vital signs on admission*$^\dagger$ | | | | |
| Systolic BP | 149 | ±31 | 148 | ±32 |
| Diastolic BP | 85 | ±18 | 84 | ±18 |
| Heart rate | 81 | ±23* | 84 | ±23* |
| Respiratory rate | 20 | ±5** | 21 | ±6** |
| *Laboratory values*$^\dagger$ | | | | |
| White blood count | 10 | ± 5 | 10 | ± 5 |
| Haemoglobin | 141 | ± 17 ** | 137 | ± 19** |
| Sodium | 139 | ± 3* | 139 | ± 4* |
| Glucose | 9 | ± 6 | 9 | ± 5 |
| Creatinine | 101 | ± 54 | 104 | ± 60 |

* $p < 0.05$, ** $p < 0.001$

† Continous variables, + Dichotomous variables

# Chapter 2

# Background: Control of confounding in observational studies

The concept of confounding is ubiquitous in epidemiology and observational research. See Rothman and Greenland [21], Pearl [22] and Greenland, Rothman and Pearl [23] for reviews. Confounding bias is the problem of mixing of the effects of the putative treatment of interest with that of extraneous outcome risk factors. For example, having yellow fingers does not affect risk of lung cancer. But yellow fingers and lung cancer will tend to be associated because yellow fingers are associated with smoking which also causes lung cancer. Conceptualizing confounding has been controversial. The definition of confounding varies from one reference to another and is an area of ongoing research [24].

In this chapter we review methods for control of confounding bias with emphasis on propensity score methods. This provides the setting for discussing our proposed Bayesian propensity score analysis. Because confounding implies biased estimation, we begin by reviewing the framework for understanding the competing targets of inference in the analysis of observational data. In Section 2.1, we review definitions of causal parameters. In Section 2.2, we discuss estimation of causal parameters in randomized experiments and observational studies. In Section 2.3, we define confounding bias and confounding variables. In Section 2.4, we review stratification on the propensity score for reducing confounding bias. We restrict our discussion to the setting of a binary treatment, binary response and a vector of measured covariates, where the log odds ratio (OR) is the measure of effect.

## 2.1 Defining causal parameters

Suppose that our objective is to estimate the causal effect of applying a dichotomous treatment on the risk of a dichotomous outcome. To define causation, a popular approach is to use potential outcome models [25]. The idea is to model not only the data that is observed in an investigation, but also the data that would be observed under hypothetical treatment interventions that are not observed. To motivate the approach, we paraphrase an example from Pearl [22]. Imagine a barometer which records air pressure measurements dichotomously as $B = High$ or $Low$ each morning on a sequence of days. Additionally, we record the weather each afternoon as either $W = Rain$ or *No Rain*. Suppose that our objective is to measure the causal effect of physically intervening to set the barometer to $B = Low$ each morning on the frequency of rain each afternoon. Furthermore, suppose that this is an observational study in the sense that we cannot actually manipulate the values of the barometer. To define "causal effect", we model the distribution of $W$ given $B$, and also the hypothetical distribution of $W$ given the unobserved value of $B$. The causal effect of $B$ on $W$ is defined based on differences between these two distributions.

To define the causal effect of a treatment on an outcome in human populations, we model the distribution of outcomes among patients who have been sampled from a large hypothetical population in which either all patients received the treatment, or all patients did not receive the treatment. Mathematically, let $Y_{\{1\}}$ denote a dichotomous random variable which models the outcome of a patient had they been sampled from the treated population. The variable $Y_{\{1\}}$ takes the value one if the patient has the outcome and zero otherwise. Similarly, let $Y_{\{0\}}$ denote the outcome for the same patient had they been sampled from the untreated population. The sampled pair of random variables $(Y_{\{1\}}, Y_{\{0\}})$ are *potential outcomes* (sometimes called counterfactuals). Let $X$ model the treatment received by the patient, taking value one if the patient was treated and zero otherwise. Let $C$ denote a $p \times 1$ vector of patient characteristics such as age and gender. Data for each patient is modeled by the collection of random variables $(Y_{\{1\}}, Y_{\{0\}}, X, C)$.

Modelling both $Y_{\{1\}}$ and $Y_{\{0\}}$ may seem strange because we know the patient received treatment $X$ and not $1 - X$. But modelling outcomes under hypothetical treatment interventions allows us to define causation. Defining causal effects based on the distribution of observed data is problematic because it leads to ambiguity about the difference between association and causation. In the barometer example, were we to investigate the effect of $B$ on $W$ in an observational study of the joint distribution of $B$ and $W$, then we might erroneously conclude that changing $B$ does cause a change in the weather because $B$ and $W$ are dependent. But this dependence does not have a causal interpretation. Changes in air pressure affect both $B$ and $W$. Suppose instead that we model the two distributions of $W$ given that we intervene in the experiment by setting $B$ to either *Low* or *High* each day. While we could never observe both distributions at the same time, common sense tells us they are identical. Thus $B$ does not cause $W$. The key to defining causation statistically is to model not only the distribution of the data that we observe, but also the distribution of data that we might have observed under hypothetical interventions. Whether or not we can make inferences about these distributions from the data is a separate issue.

We denote causal parameters from standard definitions [22, 26]. Let

$$f(y_{\{1\}}, y_{\{0\}}, x|c) = f(x|y_{\{1\}}, y_{\{0\}}, c)f(y_{\{1\}}, y_{\{0\}}|c) \tag{2.1}$$

model the joint probability density function of a sampled unit $(Y_{\{1\}}, Y_{\{0\}}, X)$ given $C$. Uppercase and lowercase letters denote random variables and realizations of random variables respectively. Define the *causal log OR conditional on $C$* as

$$\beta_c^* = \log \left[ \frac{P(Y_{\{1\}} = 1|c)/(1 - P(Y_{\{1\}} = 1|c))}{P(Y_{\{0\}} = 1|c)/(1 - P(Y_{\{0\}} = 1|c))} \right].$$

Let $P(Y_{\{x\}} = 1) = \int P(Y_{\{x\}} = 1|c)f(c)dc$ for $x = 0, 1$ where $f(c)$ is the probability density function of $C$. The quantity $P(Y_{\{x\}} = 1)$ can be interpreted as the standardized risk given treatment $X = x$, averaged with respect to the marginal density $P(C)$. Define the *average causal log OR* as

$$\beta_{avg}^* = \log \left[ \frac{P(Y_{\{1\}} = 1)/(1 - P(Y_{\{1\}} = 1))}{P(Y_{\{0\}} = 1)/(1 - P(Y_{\{0\}} = 1))} \right].$$

The parameter $\beta_c^*$ describes the causal effect of $X$ on $Y$ within the subgroup of the population with $C = c$. In contrast, $\beta_{avg}^*$ describes the causal effect assuming that all patients either have or have not been treated.

In general $\beta_{avg}^*$ and $\int \beta_c^* f(c)dc$ are not equal. This is true even when $\beta_c$ does not depend on $C$, meaning that there is no effect modification [27]. This is because the log OR is a non-linear function of risks. The average of the log ORs conditional on $C$ will typically not equal the log OR calculated from average risks. This property of ORs is called non-collapsibility by epidemiologists [23] and is related to the challenges of characterizing odds ratios in marginal and conditional models in longitudinal data analysis. In contrast, the risk difference is always collapsible, while the risk ratio is collapsible when there is no effect modification [23]. We emphasize this point because when using the log OR as a measure of effect, we obtain different parameters depending on whether we condition on $C$. For example, as discussed in Section 2.4, conditioning on the propensity score for reducing confounding bias yields a third causal quantity which may differ from either $\beta_{avg}$ or $\beta_c$. Each parameter is a valid causal quantity, but the fact that they differ from one another could lead one to erroneously conclude that a method is biased. This has been typically overlooked in the literature, particularly with respect to propensity score methods. Recent work on the relationship between causal quantities is investigated by [28–31].

## 2.2 Estimating causal effects in randomized experiments and observational studies

For each sampled unit, only one of $Y_{\{1\}}$ or $Y_{\{0\}}$ is observed while the other is missing. Formalizing, let $Y$ denote the observed outcome for a sampled patient, where

$$Y = \begin{cases} Y_{\{1\}} \text{ if } X = 1 \\ Y_{\{0\}} \text{ if } X = 0. \end{cases}$$

Or more simply, define $Y$ as $Y = Y_{\{X\}}$. Therefore, if $X = 1$ then we observe $Y = Y_{\{1\}}$ and the potential outcome $Y_{\{0\}}$ is missing. If $X = 0$ then $Y = Y_{\{0\}}$ and $Y_{\{1\}}$ is missing.

In observational studies, we observe $(Y, X, C)$ for each subject. Following [26], define the *associational* parameter,

$$\beta_c = \log\left[\frac{P(Y=1|X=1,c)/(1-P(Y=1|X=1,c))}{P(Y=1|X=0,c)/(1-P(Y=1|X=0,c))}\right].$$

Letting $\mu_x = \int P(Y = 1|x, c)f(c)dc$ for $x = 0, 1$, we define

$$\beta_{avg} = \log \left[ \frac{\mu_1/(1 - \mu_1)}{\mu_0/(1 - \mu_0))} \right].$$

The parameter $\beta_c$ is just the usual conditional association between $Y$ and $X$ given $C$, and might be estimated from logistic regression of $Y$ on $X$ and $C$. The parameter $\beta$ is a log OR calculated from standardized risks. The important issue in causal inference is to determine the relationship between $\beta_c$ and $\beta_c^*$, and between $\beta_{avg}$ and $\beta_{avg}^*$. In other words, to determine when association equals causation.

For identification of causal parameters, it is usually assumed that

$$P(X = 1|y_{\{1\}}, y_{\{0\}}, c) = P(X = 1|c)$$

or equivalently, that

$$(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|C,$$

meaning that $(Y_{\{1\}}, Y_{\{0\}})$ are conditionally independent of $X$ given $C$. This assumption appears under different names in the causal inference literature, and we call it the *assumption of no unmeasured confounding.* The assumption means that, within levels of $C$, the treatment variable $X$ is not associated with the observed or unobserved potential outcomes.

When there is no unmeasured confounding, the associational parameters are equal to causal parameters, meaning that $\beta_c = \beta_c^*$ and $\beta_{avg} = \beta_{avg}^*$. This is because

$$\begin{aligned}
P(Y=1|X=1,c) &= P(Y_{\{1\}}=1|X=1,c) \\
&= P(Y_{\{1\}}=1|c) \\
P(Y=1|X=0,c) &= P(Y_{\{0\}}=1|X=0,c) \\
&= P(Y_{\{0\}}=1|c).
\end{aligned}$$

In each equation, the first equality follows because $Y = Y_{\{X\}}$, while the second equality follows because $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|C$. Therefore $P(Y=1|x,c) = P(Y_{\{x\}} = 1|c)$ and association equals causation. We have

$$\begin{aligned}
\beta_c &= \log\left[\frac{P(Y=1|X=1,c)/(1-P(Y=1|X=1,c))}{P(Y=1|X=0,c)/(1-P(Y=1|X=0,c))}\right] \\
&= \log\left[\frac{P(Y_{\{1\}}=1|c)/(1-P(Y_{\{1\}}=1|c))}{P(Y_{\{0\}}=1|c)/(1-P(Y_{\{0\}}=1|c))}\right] \\
&= \beta_c^*.
\end{aligned}$$

Similarly $\mu_x = \int P(Y=1|x,c)f(c)dc = \int P(Y_{\{x\}}|c)f(c)dc = P(Y_{\{1\}}=1)$ which means that

$$\begin{aligned}
\beta_{avg} &= \log\left[\frac{\mu_1/(1-\mu_1)}{\mu_0/(1-\mu_0))}\right] \\
&= \log\left[\frac{P(Y_{\{1\}}=1)/(1-P(Y_{\{1\}}=1))}{P(Y_{\{0\}}=1)/(1-P(Y_{\{0\}}=1))}\right] \\
&= \beta_{avg}^*
\end{aligned}$$

To estimate causal effects, it suffices to estimate features of the distribution of $P(Y = 1|x,c)$ using standard statistical techniques, and we can dispense with the potential framework.

The value of potential outcome models is not so much statistical as it is conceptual. In simple settings, the same methods of analysis are used regardless of whether or not we use associational or causal models. But causal modelling gives a clear definition of causal effects, and it characterizes the assumption that is needed to distinguish association from causation. Whether or not the assumption of no unmeasured confounders is valid becomes a separate question. In randomized controlled trials with perfect compliance, blinding and no loss to follow up, this assumption of no unmeasured confounders is valid because the distribution $P(X = 1|c)$ is specified by the investigator. For example, we may choose $P(X = 1|c) = 0.5$ for all subjects. Then we have $P(X = 1|y_{\{1\}}, y_{\{0\}}, c) = P(X = 1|c)$ automatically. In observational studies, there can be no guarantee that the assumption is correct. Prior information is used to select a collection of covariates such that we have approximately $P(X = 1|y_{\{1\}}, y_{\{0\}}, c) \approx P(X = 1|c)$. In other words, causal inference in observational studies necessitates efforts to classify patients into strata such that the treatment is assigned approximately at random.

Causal inference using potential outcome models has also been developed in connection with models for missing data [25]. For each patient, we observe $Y_{\{X\}}$ while $Y_{\{1-X\}}$ is missing. The issue is to characterize the pattern of missing data, or rather, the way in which treatment is assigned to each patient. In the factorization of equation (2.1), the treatment assignment mechanism is modelled by $P(X = 1|y_{\{1\}}, y_{\{0\}}, c)$. If $X = 1$ then $Y_{\{0\}}$ is missing, whereas if $X = 0$ then $Y_{\{1\}}$ is missing. When there is no unmeasured confounding, the missing data mechanism does not depend on the missing potential outcome $Y_{\{1-X\}}$. For Bayesian inference, valid estimation of the causal effects defined in Section 2.1 can proceed by modelling only the observed patient outcomes $Y$ given $X$ and $C$, that is $P(Y = 1|x, c)$, without modelling the treatment assignment mechanism $P(X = 1|c)$.

We elaborate in more detail about Bayesian inference for causal effects because it is relevant to the discussion about Bayesian propensity score analysis in Chapter 3. Following Rubin [25], relabel the potential outcomes as $Y_{\{X\}}$ and $Y_{\{1-X\}}$, where $Y_{\{X\}} = Y$ is the observed outcome while $Y_{\{1-X\}}$ is missing. Suppose we model the joint probability density function of $Y_{\{X\}}$, $Y_{\{1-X\}}$ and $X$ given $C$ parametrically as

$$f(y_{\{x\}}, y_{\{1-x\}}, x|c, \theta, \gamma) = f(x|y_{\{x\}}, y_{\{1-x\}}, c, \gamma) f(y_{\{x\}}, y_{\{1-x\}}|c, \theta).$$

The parameter $\theta$ indexes the parametric model for the potential outcomes and it includes the causal parameter of interest. The quantity $\gamma$ parametrizes the missing data mechanism. Suppose that there are no unmeasured confounders, meaning that $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|C$. Then we have $f(x|y_{\{x\}}, y_{\{1-x\}}, c, \gamma) = f(x|c, \gamma)$. The probability density function of the observed quantities $Y$, $X$ and $C$ is given by

$$\begin{aligned} f(y, x|c, \theta, \gamma) &= f(y_{\{x\}}, x|c, \theta, \gamma) \\ &= \int f(y_{\{x\}}, y_{\{1-x\}}, x|c, \theta, \gamma) dy_{\{1-x\}} \\ &= \int f(x|c, \gamma) P(y_{\{x\}}, y_{\{1-x\}}|c, \theta) dy_{\{1-x\}} \\ &= f(x|c, \gamma) \int f(y_{\{x\}}, y_{\{1-x\}}|c, \theta) dy_{\{1-x\}}. \end{aligned} \quad (2.2)$$

Let $f(\theta, \gamma) = f(\theta) f(\gamma)$ denote a prior density in which $\theta$ and $\gamma$ are independent. Then the missing data mechanism is called *ignorable* [25]. Given $y, x$ and $c$, Bayesian inference for $\theta$ proceeds from the posterior distribution for $\theta$ which obeys the proportionality

$$\begin{aligned} f(\theta|y,x,c) &\propto f(y,x|c,\theta,\gamma)f(\theta)f(\gamma) \\ &\propto \left[f(x|c,\gamma)\int f(y_{\{x\}},y_{\{1-x\}}|c,\theta)dy_{\{1-x\}}\right]f(\theta)f(\gamma) \\ &\propto \int f(y_{\{x\}},y_{\{1-x\}}|c,\theta)dy_{\{1-x\}}f(\theta). \qquad (2.3) \end{aligned}$$

The parameter $\gamma$ does not appear in the posterior distribution. Hence for Bayesian inferences about $\theta$, the treatment assignment mechanism $f(x|c,\gamma)$ conveys no information about $\theta$ and can be ignored when developing a model for the data.

## 2.3 Confounding bias

The literature on confounding bias distinguishes between the notions of *confounding* and a *confounder* [24, 26]. We say that there is confounding if

$$(Y_{\{1\}}, Y_{\{0\}}) \not\perp\!\!\!\perp X|C.$$

When $(Y_{\{1\}}, Y_{\{0\}}) \not\perp\!\!\!\perp X|C$, then this implies that $P(Y_{\{X\}} = 1|c) \neq P(Y_{\{X\}} = 1|x,c) = P(Y = 1|x,c)$ for $X = 0$ or 1. Therefore,

$$\begin{aligned} \beta_c^* &= \log\left[\frac{P(Y_{\{1\}} = 1|c)/(1-P(Y_{\{1\}} = 1|c))}{P(Y_{\{0\}} = 1|c)/(1-P(Y_{\{0\}} = 1|c))}\right] \\ &\neq \log\left[\frac{P(Y = 1|X = 1,c)/(1-P(Y = 1|X = 1,c))}{P(Y = 1|X = 0,c)/(1-P(Y = 1|X = 0,c))}\right] \\ &= \beta_c \end{aligned}$$

Thus $\beta_c^* \neq \beta_c$, and the associational parameter $\beta_c$ does not have a causal interpretation. Similarly, we have $\beta_{avg}^* \neq \beta_{avg}$.

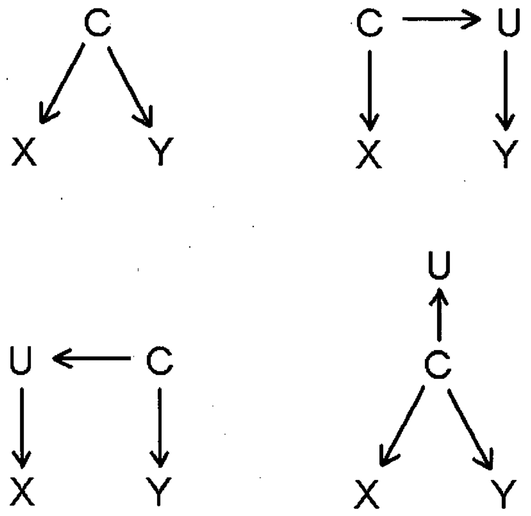
This definition of confounding is appealing because it avoids reference to measures of effect, and it helps distinguish confounding from the unrelated notion of non-collapsibility described in Section 2.1. Lack of a clear separation between confounding and non-collapsibility when inferring causation created some controversy in early attempts to characterized confounding. For example, when analyzing data from randomized trials, conditioning on outcome risk factors in a stratified analysis will drive the log OR parameter away from zero when there is a treatment effect [23, 32]. But this change in the log OR is unrelated to confounding. See [22, 24] for discussion.

Recent definitions of confounders, meaning variables which are responsible for confounding, assume that the causal relationship between variables in a population can be modelled using directed acyclic graphs (DAGs). While we have discussed potential outcome models for causation, DAGs offer a different strategy that can be used for qualitative assessment of bias. For fixed time point treatments, such as in the case of the statin data example of Section 1.2, the approaches can be shown to be equivalent in the sense that both types of models give the same mathematical description of the same quantities [22].

A DAG consists of a set of variables connected by arrows in which no directed paths form loops. An arrow models direct causal effects of the parent variable on the child variable. A scalar variable $C$ is defined as a confounder for the effect of $X$ on $Y$ if $C$ connects to $X$ and $Y$ by forward pointing arrows. For example, Figure 2.1 presents examples of DAGs in which $C$ is a confounding variable. Therefore, $C$ is a confounder if it causes both $X$ and $Y$. If $C$ is smoking, $Y$ is lung cancer, and $X$ is yellow fingers, and their causal relationship can be modelled by any of the diagrams in Figure 2.1, then smoking is a confounder for the effect of yellow fingers on lung cancer.

The definition of a confounder appearing in modern epidemiological textbooks

Figure 2.1: Directed acyclic graphs in which $C$ is a confounder, $X$ is treatment $Y$ is the outcome, and $U$ is an additional measured or unmeasured variable.



says that $C$ is a confounder for the effect of $X$ on $Y$ under two conditions. *Condition 1*: The variable $C$ must be associated with both $X$ and $Y$. More technically, we require that $Y \not\perp\!\!\!\perp C|X$ and $X \not\perp\!\!\!\perp C|Y$. *Condition 2*: The causal interpretation of dependencies between $Y$, $X$ and $C$ is restricted to require that $C$ is not affected by $X$ or $Y$. This definition and the one based on DAGs are usually equivalent [24]. But using DAGs to define a confounder formalizes and generalizes the intent of *Condition 2* to more complex observational data (see [24] for details).

To infer causation from observational data, we must identify a set of covariates $C$ such that we have approximately $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|C$. DAGs can be helpful in this process. Instead of focusing on whether individual covariates meet the traditional definition of being confounders, the investigator can instead attempt to elicit a DAG model for all relevant measured and unmeasured factors. Such a diagram may be elaborate, but simple criteria have been developed based on paths between the treatment and outcome variable that allow the investigator to verify if $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|C$ [22, 26].

## 2.4 Propensity score analysis (PSA) for control of confounding

When there is no unmeasured confounding, the standard approach to estimating causal effects is to estimate $P(Y = 1|x, c)$ over levels of $C$ and then calculate an estimate of $\beta_c$, the log OR conditional on $C$. We call this stratification on measured confounders. When strata are sparse, model-based estimation is often used. We assume a parametric model for $P(Y = 1|x, c)$, such as a logistic regression model of $Y$ on $X$ and $C$, and proceed by maximum likelihood estimation. A difficulty is that this strategy requires accurate specification of $P(Y = 1|x,c)$. This may be difficult if $C$ is high-dimensional with continuous components. In the statin data example of Section 1.1, we have rich patient information on quantities such as vital signs at hospital admission and laboratory values. But the functional form of the dependence between these variables and mortality is poorly understood. If the model for $P(Y = 1|x,c)$ is misspecified, then estimates of $\beta_c$ will be asymptotically biased.

An alternative technique to control confounding is to use propensity scores. The propensity score is defined as the probability that a subject is treated given measured confounders, or mathematically as the quantity $Z = P(X = 1|c)$. The propensity score can be used as a tool to ensure that treatment groups have similar distributions of measured confounders. Rosenbaum and Rubin [1], Theorem 1, showed that patients with fixed propensity score $Z$ have the same distribution for $C$ irrespective of $X$, or more technically that $C \perp\!\!\!\perp X|Z$. This is because $P(X = 1|c,z) = z$, which does not depend on $C$. Further Rosenbaum and Rubin [1], Theorem 3, showed that if there is no unmeasured confounding conditional on $C$, then this implies that there is no unmeasured confounding conditional on $Z$. Or rather, that $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|Z$. The reason is because

$$\begin{aligned}
f(y_{\{1\}}, y_{\{0\}}|x,z) &= \int_{U_z} f(y_{\{1\}}, y_{\{0\}}|x,c,z) f(c|x,z)dc \\
&= \int_{U_z} f(y_{\{1\}}, y_{\{0\}}|c,z) f(c|x,z)dc \\
&= \int_{U_z} f(y_{\{1\}}, y_{\{0\}}|c,z) f(c|z)dc \\
&= f(y_{\{1\}}, y_{\{0\}}|z).
\end{aligned}$$

The integration is over the subsets of the support of $C$ given by the set $U_z = \{c|f(c) = z\}$ for $0 < z < 1$. The set $U_z$ corresponds to the subset of the population who have propensity score $z$. The first line follows because there is no unmeasured confounding given $C$, while the second line follows because $C \perp\!\!\!\perp X|Z$. Thus $P(Y_{\{1\}}, Y_{\{0\}}|x, z) = P(Y_{\{1\}}, Y_{\{0\}}|z)$ and we have $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|Z$.

Rosenbaum and Rubin used the result $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|Z$ to argue that stratifying on the propensity score eliminates confounding bias due to C. The association between $X$ and $Y$ within patients with the same propensity score has a causal interpretation. Since $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|Z$, we have $P(Y = 1|x, z) = P(Y_{\{X\}} = 1|z)$ because

$$\begin{aligned} P(Y = 1|X = 1, z) &= P(Y_{\{1\}} = 1|X = 1, z) = P(Y_{\{1\}} = 1|z) \\ P(Y = 1|X = 0, z) &= P(Y_{\{0\}} = 1|X = 0, z) = P(Y_{\{0\}} = 1|z). \end{aligned}$$

The associational log OR conditional on $Z$, written

$$\beta = \log \left[ \frac{P(Y = 1|X = 1, z)/(1 - P(Y = 1|X = 1, z))}{P(Y = 1|X = 0, z)/(1 - P(Y = 1|X = 0, z))} \right],$$

has a causal interpretation because it is equal to the causal log OR conditional on $Z$

$$\beta^* = \log \left[ \frac{P(Y_1 = 1|z)/(1 - P(Y_1 = 1|z))}{P(Y_0 = 1|z)/(1 - P(Y_0 = 1|z))} \right].$$

Patients with fixed propensity scores have equal probability of treatment, irrespective of their characteristics. Therefore, comparing patient outcomes between treatment groups conditional on the propensity score removes confounding. The parameter $\beta$ will generally differ from either $\beta_{avg}$ or $\beta_c$, the causal parameter defined in Section 2.2, because of non-linearity of the log OR [28, 29]. But $\beta$ can nonetheless be interpreted as a valid causal quantity [30, 31].

### *Point estimation of $\beta$ using propensity score analysis (PSA)*

In observational studies, the propensity score is unknown and can be interpreted as missing data. A propensity score analysis (PSA) proceeds using a two step procedure. First the propensity scores are estimated for each patient. Then we fit a regression model on the estimated propensity scores. In most applications, standard error estimates are calculated from asymptotic approximation to the distribution of the maximum likelihood estimator based on the regression model of the outcome on treatment and estimated propensity scores [4]. Thus interval estimates ignore uncertainty in the estimated propensity scores.

To illustrate in greater detail, suppose that we have an observational study with sample size $n$. To define the data, let $y_i$ and $x_i$ for $i = 1, \ldots, n$, denote the observed values of outcome $Y$ and treatment $X$, and let $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$. Let $c_i$ for $i = 1, \ldots, n$, denote values of the $p \times 1$ vector $C$ of confounding variables, and let $\mathbf{c}$ denote an $n \times p$ design matrix where each row is given by $c_i$. To ease notation in modelling regression intercept terms, we assume that the first component of $C$ is identically equal to one, or rather, that the matrix $\mathbf{c}$ contains a column of ones.

The vector of propensity scores $\mathbf{z} = (z_1, \ldots, z_n)$, where $z_i = P(X = 1|c_i)$, is unknown. To estimate $\beta$, we first estimate $\mathbf{z}$, typically via logistic regression of $\mathbf{x}$ on $\mathbf{c}$ using the model

$$\text{logit}[P(X = 1|c)] = \gamma' c \tag{2.4}$$

The fitted values of the regression provide point estimates of $\mathbf{z}$ denoted $\hat{\mathbf{z}}$. Next, the patients are stratified on the estimated propensity scores, and we estimate characteristics of $P(Y = 1|x, z)$ such as $\beta$. Exact stratification on the propensity score is often not possible because the data are sparse. Instead, we can propose a regression model for $P(Y = 1|x, z)$, such as

$$\text{logit}[P(Y = 1|x, z)] = \beta x + g(z)'\xi, \tag{2.5}$$

and then use maximum likelihood estimation with $\mathbf{z} = \hat{\mathbf{z}}$ to calculate an estimate for $\beta$. The quantity $g(Z)$ is a $k \times 1$ covariate vector which is a known function of the propensity score and models the relationship between the propensity score and $Y$. A popular choice for $g(Z)$ is based on five equal sized quintile groups. Then $g(z)$ is a $5\times1$ vector in which the latter four components are "dummy" variables indicating quintile group membership,

$$g(z)' = \begin{cases} [1, 0, 0, 0, 0] & \text{if } z < c_1 \\ [1, 1, 0, 0, 0] & \text{if } c_1 \leq z < c_2 \\ [1, 0, 1, 0, 0] & \text{if } c_2 \leq z < c_3 \\ [1, 0, 0, 1, 0] & \text{if } c_3 \leq z < c_4 \\ [1, 0, 0, 0, 1] & \text{if } c_4 < z. \end{cases} \tag{2.6}$$

The parameter vector $\xi = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$ governs the risk of $Y$ within each group. The quantities $c_1, c_2, c_3, c_4$ which define the quintile groups are specified in advance based on the empirical distribution of the estimated propensity scores $\hat{z}_1, \ldots, \hat{z}_n$. When subclassifying on propensity score bins, we can also use the Mantel Haenszel method [4]. Alternatively we could model $g(z)'\xi$ using cubic splines

$$g(z)'\xi = \xi_0 + \sum_{j=1}^{3} \xi_j z^j + \sum_{j=1}^{p} \xi_{j+p}(z - c_j)_+^3$$

where $(u)_+^p = u^p I(u \geq 0)$ with knots $c_1, c_2, \ldots c_p$, or as a simple linear predictor $g(z)'\xi = \xi_0 + \xi_1 z$ [2, 33]. If matched pairs are available with matching on the estimated propensity score, then we can use conditional likelihood estimation which does not assume an explicit functional form for $g(z)'\xi$.

PSA is similar to conventional regression of $Y$ on $X$ and $C$, except that it substitutes a high dimensional covariate $C$ with the scalar $Z$. Each of the functional forms for $P(Y = 1|x, z)$ assumes that there is no effect modification due to the propensity score. But this may be unrealistic in some settings and the method can be extended appropriately [31].

### *Interval estimation of $\beta$ using PSA*

Interval estimates for $\beta$ are typically calculated by estimating the asymptotic variance of $\hat{\beta}$, denoted $\hat{Var}\{\hat{\beta}\}$. We report

$$\left[\hat{\beta} - 1.96\sqrt{\hat{Var}\{\hat{\beta}\}}, \hat{\beta} + 1.96\sqrt{\hat{Var}\{\hat{\beta}\}}\right]$$

as a 95% confidence interval for $\beta$. The quantity $\hat{Var}\{\hat{\beta}\}$ is calculated from the observed information matrix based on the parametric model for $P(Y = 1|x, z)$. When $P(Y = 1|x, z)$ follows equation (2.4), this is accomplished by evaluating the matrix

$$\begin{aligned}
& \frac{\partial^2}{\partial(\beta, \xi)^2} \sum_{i=1}^{n} \log\{f(y_i|x_i, z_i)\} \\
= & \frac{\partial^2}{\partial(\beta, \xi)^2} \sum_{i=1}^{n} \log\left\{\frac{\exp(y_i(\beta x_i + g(z_i)'\xi)))}{1 + \exp(\beta x_i + g(z_i, \xi))}\right\} \\
= & \frac{\partial^2}{\partial(\beta, \xi)^2} \sum_{i=1}^{n} y_i(\beta x_i + g(z_i)'\xi) - \log(1 + \exp(\beta x_i + g(z_i)'\xi))
\end{aligned}$$

at the maximum likelihood estimates for $\beta$ and $\xi$. This expression is calculated from $\hat{\mathbf{z}}$ rather than $\mathbf{z}$, and the standard error estimate of $\beta$ does not acknowledge the uncertainty in the estimated propensity scores. Thus we might expect that the interval estimates will be too narrow and not have nominal coverage probability.

There is little discussion in the literature about methods for incorporating uncertainty from the estimated propensity scores into uncertainty about the treatment effect. Some exceptions include [4, 6]. In contrast, there a large body of work describing the merits of using estimated propensity scores over true propensity scores [7]. Hahn and others [8–11] argue that treatment effect estimates calculated from the estimated propensity scores are generally more efficient than corresponding estimates calculated from true propensity scores. This has been demonstrated both analytically and in simulations in the context of propensity score weighting adjustments for confounding [8–11]. An accessible discussion is given by Hirano, Imbens and Ridder [11] for the case of a single dichotomous confounder. For matched sampling, Rubin and Thomas [12, 13] show that matching on the estimated propensity scores yields greater similarity in empirical distributions of covariates for treated versus control subjects relative to matching on true propensity scores. A similar argument is given by Rosenbaum and Rubin [3] based on the analysis of a dataset using PSA stratifying on quintiles of the estimated propensity scores. We return to this topic in more detail in Section 5.4.

Nonetheless, it is unclear whether or not ignoring uncertainty in the estimated propensity scores is harmful to interval estimation for PSA. Many of the arguments in favor of using estimated propensity scores rely on large-sample theory. Interval estimates ignoring the estimated propensity scores may not have nominal coverage in finite samples. The arguments also tend to focus on propensity score weighting methods rather than PSA. The literature investigating interval estimation for PSA is fairly sparse and usually considers continuous outcomes (e.g. [3, 4, 34]). For dichotomous outcomes where the log OR is the measure of effect, the non-linearity of the logit link causes ambiguity about the targets of inference. Thus it becomes difficult to make sense of notions like coverage probability because it depends on what the analyst is trying to estimate. For example, it is sometimes assumed PSA should be used to estimate either $\beta_{avg}$ or $\beta_c$, defined in Section 2.2, even though it is asymptotically biased for both quantities [34, 35]. A detailed discussion of the targets of inference for PSA is given in [28, 29]. We discuss this topic in more detail in Section 5.2.

# Chapter 3

# Bayesian propensity score analysis (BPSA)

In this chapter we present a Bayesian propensity score analysis (BPSA). The method is similar to PSA in the sense that it uses the same models for the data. But inferences are carried out using Bayes theorem rather than maximum likelihood. Because the propensity score for each subject is unknown, it is modelled as a latent covariate which is integrated out of the posterior distribution of model parameters.

Unlike PSA which uses a two step procedure to first estimate the propensity scores and then estimate the treatment effect, BPSA estimates both quantities simultaneously. This can offer unique advantages to BPSA. We observe an "information flow" in which outcome data is used for efficient estimation of the propensity scores. In Section 3.1 we describe the BPSA method including the model, prior distributions and a strategy for posterior simulation. We show that during MCMC, the conditional distribution for the propensity scores depends on the distribution of patient outcomes. The BPSA method incorporates prior information about smoothness assumptions between the response distribution and propensity score, within treatment groups. This information is not used by PSA for estimation of the propensity scores. While this seems like a reasonable estimation strategy, it is a break with tradition [7, 36]. For example, Rubin writes [36]:

"Of substantial importance, the propensity score approach to causal inference ... focuses on the theme that the design of an observational study should parallel the design of a randomized experiment. That is, our propensity score approach is accomplished without any access to outcome data."

These issues are explored further in Section 3.2, where we discuss using BPSA as a tool for causal inference.

## 3.1 The method

*Model*

Factorize the density of $Y$, $X$, and $C$ as

$$P(Y = y, X = x|c) = P(Y = y|x, c)P(X = x|c).$$

We use models which are identical to the PSA model given in equations (2.4) and (2.5). We let

$$\begin{aligned} \text{logit}[P(Y = 1|x, c)] &= \beta x + g(z(c, \gamma))'\xi && (3.1) \\ \text{logit}[P(X = 1|c)] &= \gamma' c, && (3.2) \end{aligned}$$

where $z(c, \gamma) = \text{expit}(\gamma' c)$ and $\text{expit}(a) = (1 + \exp(-a))^{-1}$. Here we write $z = z(c, \gamma)$ to acknowledge that the propensity score is a known analytic function of $C$ and $\gamma$.

Equation (3.1) is a logistic regression model for the risk of $Y$ with treatment effect

$\beta$ which does not interact with $C$ or $Z$. The parameter $\beta$ is the primary quantity of interest and models the causal effect of $X$ on $Y$ given $Z$. The quantity $g(Z)'\xi$ is a linear predictor relating the propensity score $Z = z(c, \gamma)$ to the risk of $Y$ via the parameters $\xi$. In this thesis, we restrict our attention the case where $g(Z)$ is a 5×1 vector of indicator variables

$$g(Z)' = \begin{cases} [1,\ 0,\ 0,\ 0,\ 0] & \text{if } Z < c_1 \\ [1,\ 1,\ 0,\ 0,\ 0] & \text{if } c_1 \leq Z < c_2 \\ [1,\ 0,\ 1,\ 0,\ 0] & \text{if } c_2 \leq Z < c_3 \\ [1,\ 0,\ 0,\ 1,\ 0] & \text{if } c_3 \leq Z < c_4 \\ [1,\ 0,\ 0,\ 0,\ 1] & \text{if } c_4 < Z \end{cases} \tag{3.3}$$

The quantities $c_1, c_2, c_3, c_4$ must be specified a priori. They define five separate "bins" in which the risk of $Y$ given $X$ is assumed to be constant. The components of $\xi = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$ model the risk of $Y$ within the bins. To choose $c_1, c_2, c_3, c_4$, we use an approach motivated by PSA. We fit the logistic regression of $X$ on $C$ via maximum likelihood and then use the fitted values to obtain initial estimates of the propensity scores. The values of $c_1, c_2, c_3, c_4$ are selected to define five equal size quintile groups from the empirical distribution of these estimates.

It is important to emphasize that the bins $[0, c_1]$, $[c_2, c_3]$, $[c_3, c_4]$, $[c_4, 1]$ will not define quintile groups when applying BPSA. The parameter $\gamma$ is modelled as a latent variable, and thus the covariate $g(z(c, \gamma))$ is also a latent quantity. We can only classify patients into quintile groups in an average sense, such as based on the posterior expectation of $g(z(c, \gamma))$. Because the quantities $c_1, c_2, c_3, c_4$ are chosen using PSA, we would not necessarily expect 20% of the study observations to fall within each bin.

The choice of model for the outcome variable may not be suitable for the EFFECT data. The idea that the risk of $Y$ is constant within bins is clearly only an approximation, and it it appears that there might be more suitable functional forms for $g(Z)$. Nonetheless, we use the form of $g(Z)$ given above in order to maintain similarity with the published litterature on PSA. In first proposing PSA, Rosenbaum and Rubin [1] advocated using five equal sized bins because previous work by Cochran demonstrated that this would elimianate 90% of the bias induced by a confounder. Furthermore, our method can be readily modified to incorporate more flexible choice of $g(Z)$ through modification of the MCMC algorithm.

Another possible limitation of the model is the choice of using the log OR $\beta$ to model the causal effect of $X$ on $Y$. As discussed in Chapter 2.1, the log OR has well know limitation compared to other effect measures. In this investigation we choose the log OR because of the flexibility of using logistic regression to model binary data. The logit link simplifies modelling when $g(Z)$ takes on other forms such as cubic splines. Nonethless, in priciple, we could define an outcome model using other measures of effect.

### *Prior distributions*

The model parameters $\beta$, $\xi$ and $\gamma$ are all standard regression coefficients, and we use mean-zero diffuse independent normal distributions as prior distributions.

$$\begin{aligned} \beta &\sim N(0, 100), \\ \gamma_1, \ldots, \gamma_k &\sim N(0, 100), \\ \xi_1, \ldots, \xi_5 &\sim N(0, 100). \end{aligned}$$

This should yield similarity between BPSA and PSA, which uses no prior information for model parameters. There are also opportunities for more flexible modelling strategies. Zheng and Little [37] propose a model similar to BPSA for estimating the population mean from a finite population where subjects have non-equal probability of selection. They model the distribution of the data conditional on the selection probabilities, and they use a hierarchical prior distribution for the subgroup means. The resulting inferences lie mid-way between assuming that the means are all identical versus all different. For BPSA, we could use a similar strategy and model the parameters $\xi_1, \xi_2, \ldots \xi_5$ hierarchically and as conditionally independent and identically distributed given an unknown hyperparameter.

### *Posterior simulation*

Following the notation given in Section 2.3, recall that $\mathbf{y}$, $\mathbf{x}$ denote vectors of length $n$ of the observed responses, exposures in $n$ study subjects, and $\mathbf{c}$ denotes an $n \times p$ matrix of measured covariates with a column of ones. Our objective is to study $f(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c})$, the posterior distribution of model parameters given $\mathbf{y}$, $\mathbf{x}$ and $\mathbf{c}$. To accomplish this, we sample from the posterior using MCMC. To illustrate the main ideas, observe that if the propensity scores were known then posterior inferences could be accomplished by fitting the logistic regression models in equation (3.1). Hence posterior simulation may proceed by treating $\gamma$ as latent variable that is integrated out of the joint posterior distribution $f(\beta, \xi, \gamma|\mathbf{y}, \mathbf{x}, \mathbf{c})$. We update successively from $f(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c}, \gamma)$ and $f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi)$.

To update from the conditional densities $f(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c}, \gamma)$ and $f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi)$ we must appropriately condition the joint distribution of the data and parameters. Write this distribution as

$$f(\mathbf{y}, \mathbf{x}, \beta, \xi, \gamma|\mathbf{c}) = f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma) f(\mathbf{x}|\mathbf{c}, \gamma) f(\beta, \xi, \gamma).$$

where $f(\beta, \xi, \gamma)$ is the prior distribution for $\beta, \xi$ and $\gamma$. All of the density functions in the right-hand side of this equation are known from the modelling and prior assumptions. Therefore, $f(\beta, \xi | \mathbf{y}, \mathbf{x}, \mathbf{c}, \gamma)$ obeys the proportionality

$$\begin{aligned} f(\beta, \xi | \mathbf{y}, \mathbf{x}, \mathbf{c}, \gamma) &\propto f(\mathbf{y} | \mathbf{x}, \mathbf{c}, \beta, \xi, \gamma) f(\beta, \xi) \\ &= \prod_{i=1}^{n} \frac{\exp\{y_i(\beta x_i + g(z(c_i, \gamma))'\xi)\}}{1 + \exp\{\beta x_i + g(z(c_i, \gamma)'\xi\}} \times f(\beta, \xi). \end{aligned} \tag{3.4}$$

This is the likelihood for logistic regression of $\mathbf{y}$ on $\mathbf{x}$ and the propensity score $\mathbf{z}(\mathbf{c}, \gamma)$, multiplied by the prior for $\beta$ and $\xi$. Updating from the density $f(\beta, \xi | \mathbf{y}, \mathbf{x}, \mathbf{c}, \gamma)$ involves a single update from Bayesian logistic regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}$.

Bayesian logistic regression can be accomplished using the Metropolis Hastings algorithm [38]. The algorithm is an iterative procedure for sampling from a target density $f(\theta) = k^{-1}\pi(\theta)$, where $k$ is an unknown normalization constant. The implementation proceeds as follows: At iteration $i$, given a current sampled parameter value $\theta^{(i)}$, a candidate value $\theta^*$ is generated from a proposal density $Q(\theta^* | \theta^{(i)})$. We assign $\theta^{(i+1)} \leftarrow \theta^*$ with probability

$$\min \left[ \frac{f(\theta^*) Q(\theta^{(i)} | \theta^{(i)})}{f(\theta^{(i)}) Q(\theta^* | \theta^{(i)})}, 1 \right],$$

or assign $\theta^{(i+1)} \leftarrow \theta^{(i)}$ otherwise. After discarding a suitable number of initial iterations, the series $\theta^{(1)}, \theta^{(2)}, \ldots$ are a dependent sample from the target density $f(\theta)$. The choice of proposal density $Q(\theta^* | \theta^{(i)})$ impacts the performance of the Metropolis Hastings algorithm. For Bayesian logistic regression, a common choice is a multivariate normal density with mean equal to the maximum likelihood estimator and covariance matrix given by the inverse of the observed information. This proposal density approximates the posterior distribution in large samples and yields high acceptance rates for candidate parameter values in posterior updating.

The density of $f(\gamma|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi)$ obeys the proportionality

$$\begin{aligned} f(\gamma|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi) &\propto f(\mathbf{y}|\mathbf{x},\mathbf{c},\beta,\gamma,\xi)f(\mathbf{x}|\mathbf{c},\gamma)f(\gamma) \\ &= \prod_{i=1}^{n}\left[\frac{\exp\{y_i(\beta x_i + g(z(c_i,\gamma)'\xi)\}}{1+\exp\{\beta x_i + g(z(c_i,\gamma)'\xi\}} \times \frac{\exp\{x_i(\gamma' c)\}}{1+\exp\{\gamma' c\}}\right] \\ &\quad \times f(\gamma). \end{aligned} \tag{3.5}$$

This density is not proportional to $f(\mathbf{x}|\mathbf{c},\gamma)f(\gamma)$. Therefore, updating the propensity scores $z(\mathbf{c},\gamma)$ in BPSA does not consist of sampling from the posterior distribution of regression coefficients from logistic regression of $\mathbf{x}$ on $\mathbf{c}$. Instead, information about patient outcomes is also involved in updating information about the propensity scores.

We update from $f(\gamma|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi)$ using a Metropolis-Hastings step. Finding a suitable proposal distribution is challenging because the characteristics of this distribution are not obvious. For example, we found that a proposal based on the approximation

$$f(\gamma|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi) \approx \frac{f(\mathbf{x}|\mathbf{c},\gamma)f(\gamma)}{k}, \tag{3.6}$$

where $k$ is a normalization constant, led to unacceptably high rejection rates. Such a proposal approximates the target density using the asymptotic distribution of the maximum likelihood estimator from fitting the regression in equation (3.2). The fact that the approximation is poor indicates that the patient outcome distribution may supply a lot of information about $\gamma$. Instead, we use a proposal distribution based on a random walk Metropolis Hastings algorithm which updates the $p$ components of $\gamma$, one at a time. This approach samples from $f(\gamma|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi)$ by sampling sequentially from $f(\gamma_1|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi,\gamma_{(-1)}),\ldots,\ f(\gamma_k|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi,\gamma_{(-k)})$, where $\gamma_{(-j)}$ denotes $(\gamma_1,\ldots,\gamma_{j-1},\gamma_{j+1},\ldots,\gamma_k)$. At each step, a proposal for $f(\gamma_j|\mathbf{y},\mathbf{x},\mathbf{c},\beta,\xi,\gamma_{(-j)})$

is given by a random draw from a univariate $t$-distribution with appropriate scale and degrees of freedom. The scale parameter is chosen so as to ensure fast mixing of the MCMC chain through the target distribution. If the scale is too small, the chain will move slowly. A large scale will give high rejection rates. Specifying a small degrees of freedom for the $t$-distribution gives heavy tails to the proposal density, which allows greater flexibility for ensuring rapid movement through the target distribution.

A difficulty with this updating scheme is that the computational cost will be substantial when the number of covariates $p$ is large. For example, in the statin data described in Section 1.1 we have $p = 20$. An alternative strategy is to use a multivariate random walk. We could update $\gamma$ using a proposal equal to $\gamma$ plus a draw from a multivariate $t$-distribution of dimension $p$ with mean zero, small degrees of freedom, and scale matrix equal to the inverse of the observed information from logistic regression of $\mathbf{x}$ on $\mathbf{c}$. This proposal distribution has a similar shape as the density in equation (3.6), but is not constrained to one area of the parameter space.

We assess sampler convergence using the CODA package which is available for the statistical software R [39] designed for output analysis and diagnostics for MCMC. The cumuplot() function plots the evolution of sample quantiles over iterations, and it can be used to identify poor mixing. Diagnostic tools based on the analysis of multiple MCMC chains are also available [40].

To summarize, simulation from the posterior distribution $f(\beta, \xi, \gamma, |\mathbf{y}, \mathbf{x}, \mathbf{c})$ proceeds as follows:

1. Specify the quantities $c_1, c_2, c_3, c_4$ by constructing five quintile groups from the fitted values of a logistic regression of $\mathbf{x}$ on $\mathbf{c}$ fit by maximum likelihood. This defines the "bins" for BPSA.

2. Obtain a starting value for $\gamma^{(0)}$, by sampling once from a normal distribution

with mean and variance given by the maximum likelihood estimator and its asymptotic variance, respectively, calculated form the logistic regression of $\mathbf{x}$ on $\mathbf{c}$ from Step 1. Obtain starting values for $(\beta^{(0)}, \xi^{(0)})$ by sampling from the asymptotic distribution of the maximum likelihood estimators from logistic regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}(\mathbf{c}, \gamma^{(0)})$ using the regression model of equation (3.1).

3. For $t = 1, 2, \ldots$

   (a) Update $\gamma^{(t+1)}$ by updating successively from each of $f(\gamma_1 | \mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi, \gamma_{(-1)}), \ldots,$ $f(\gamma_k | \mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi, \gamma_{(-k)})$ using a random walk Metropolis Hastings step with proposal distribution that is univariate $t$ with a suitable scale parameter and degrees of freedom.

   (b) Update $\beta^{(t+1)}$ and $\xi^{(t+1)}$ using a Metropolis Hastings step with target density $f(\beta, \xi | \mathbf{y}, \mathbf{x}, \mathbf{c}, \gamma^{(t+1)})$ and proposal distribution obtained by logistic regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}(\mathbf{c}, \gamma^{(t+1)})$.

After discarding a suitable number of initial iterations, the sequence $(\beta^{(i)}, \xi^{(i)}, \gamma^{(i)})$ for $i = 1, 2, \ldots$, is a serially dependent sample from the required posterior distribution $f(\beta, \xi, \gamma, | \mathbf{y}, \mathbf{x}, \mathbf{c})$.

## 3.2 Causal inference with BPSA

The BPSA model described in Section 3.1 seems reasonable. BPSA essentially mimics PSA from a Bayesian perspective. We can generate data from the model. Intuition and standard large sample theory suggest that BPSA point and interval estimates will agree with PSA in large samples. This is confirmed in the simulations of Chapter

5. Nonetheless, BPSA is intended to be used as a tool for causal inference, and upon closer inspection we see some technicalities.

Does it make sense to use outcome data in order to estimate propensity scores? In Section 3.1, we showed that when updating the propensity scores, the conditional distribution of $\gamma$, given in equation (3.5) depends on the observed values of $\mathbf{y}$ and the parameters $\xi$ and $\beta$. As we learn about the risk of $Y$ as a function of $g(Z)$ and $X$, this information flows back through the MCMC algorithm to impact estimation of $\gamma$. Unlike PSA, the BPSA method fits both regression models given in equations (3.1) and (3.2) simultaneously rather than one at a time. Thus the two regression models assist each other in order to produce a good fit for the data.

But this estimation strategy is in conflict with the standard approaches to PSA. Rubin [7] emphasizes that when estimating the propensity scores, the investigator should not have access to outcome data. Rosenbaum and Rubin [3] advocate estimating the propensity scores by an iterative process where the investigator first estimates the propensity scores using a model for the distribution of $X$ given $C$, and then checks for balance in the distribution of $C$ within quintile groups. If the covariate distributions differ between treatment groups, this indicates that the model for $P(X = 1|c)$ is incorrect. The investigator can then estimate the propensity scores again using alternate models. PSA is a tool for constructing comparable treatment groups and the method should be blind to the values of the outcome. When there are no unmeasured confounders, PSA replicates randomized groups in a manner similar to actual randomization. Furthermore, there is the issue of the time ordering of when $Y$ and $X$ are measured. If the outcome variable is measured after the start of follow-up, then is it meaningful to use this quantity to make inferences of the probability of treatment?

To reconcile these contradictory viewpoints, we need to establish if the regression model given in equation (3.1) is a sensible model. If we truly believe that the risk of $Y$ given $X$ is determined by the linear predictor $g(Z)$, then it makes sense to use Bayes rule to incorporate patient outcome information when estimating propensity scores. However, in most published articles on propensity score methods, PSA is evaluated assuming that the regression model given in equation (3.1) is *misspecified* (see for example [4, 28, 34, 35]). The ideal analysis estimates the association between $X$ and $Y$ conditional on $Z$ exactly, while stratification of study units into equal sized bins is a only crude approximation of this procedure. Therefore PSA will tend to produce biased estimates because of residual confounding due to incomplete adjustment for $Z$. The justification of this perspective is that while the association between $X$ and $Y$ given $Z$ has a causal interpretation, the interpretation ought not extend to units within the same quintile group. In Section 2.2, we showed that when there is no unmeasured confounding conditional on $C$, then this implies that $(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|Z$. Therefore given $Z$, all patients have equal probability of $X = 1$ irrespective of $C$, and the association between $X$ and $Y$ is unconfounded. But among units with propensity score $Z \in [c_k, c_{k+1})$, for some interval from $c_k$ to $c_{k+1}$, treated and untreated units will not have identical distributions of propensity scores. Consequently, it seems tenuous to argue that the association between $X$ and $Y$ within a fixed quintile has a causal interpretation. If we adhere to this logic, then it implies that equation (3.1) is not a realistic model for causal inference. Thus using Bayes theorem and equation (3.1) to estimate propensity scores BPSA is not a sensible strategy for analyzing observational data.

An alternative perspective is that it may be appropriate to impose additional modelling assumptions which treat the distribution of $Y$ given $X$ as a smooth function in $Z$. For example, if patients A and B have propensity scores equal to 0.75 and 0.80 respectively, then it may be reasonable to assume that they have similar baseline risk of $Y$. In the statin data example, healthy patients are likely to be treated, and consequently, patients with high propensity scores have reduced risk of death. Thus while the distribution of propensity scores may differ among treated and untreated patients within quintile groups, all patients may have roughly equal risk of $Y$. This reasoning is similar to that used to justify standard regression models that categorize confounders. When adjusting for patient age using conventional regression, we may group patients into one year groups rather than one month groups. Inferences are then reported conditional on this smoothness assumption.

If we can assume that

$$(Y_{\{1\}}, Y_{\{0\}}) \perp\!\!\!\perp X|g(Z) \tag{3.7}$$

then this implies that

$$\begin{aligned} P(Y=1|X=1, g(z)) &= P(Y_{\{1\}}=1|X=1, g(z)) = P(Y_{\{1\}}=1|g(z)) \\ P(Y=1|X=0, g(z)) &= P(Y_{\{0\}}=1|X=0, g(z)) = P(Y_{\{0\}}=1|g(z)). \end{aligned}$$

The log ORs calculated from $P(Y=1|X, g(z))$ will have a causal interpretation within subsets of the population with a given g(z). The assumption in equation (3.7) is very similar in spirit to the assumption of no unmeasured confounders from equation (2.2). But it in some sense it is made implicitly in conventional PSA. When we use PSA to calculate model based point and interval estimates for $\beta$, we are assuming that equation (3.1) models the true distribution of the response variable. The choice of linear predictor $g(.)$, whether quintile groups or cubic splines, is presumably guided by prior information about the smoothness of $P(Y=1|x, z)$.

As an aside, we note that the BPSA model, given in equation (3.1), is not compatible with potential outcome models for causal inference. BPSA models the quantities $P(Y = 1|x, c)$ and $P(X = 1|c)$ as dependent *a priori*. We see this because the quantity $P(Y = 1|x, c) = \text{expit}\{\beta x + g(z(c, \gamma))'\xi\}$ is an explicit function of $z(c, \gamma) = P(X = 1|c) = \text{expit}(\gamma' c)$. What this means is that, given $X$ and $C$, we believe that the risk of $Y$ depends on the manner in which treatment is assigned. However, causal inference using potential outcomes models *requires* that $P(X = 1|c)$ and $P(Y = 1|x, c)$ be independent *a priori*. As described in Section 2.2, in order for the treatment assignment mechanism to be ignorable in the sense of Rubin [25], a necessary condition is that the parameters governing the potential outcomes be independent *a priori* from the parameters which model the treatment assignment mechanism.

To elaborate, in the Section 2.2 equation (2.2), we showed that the potential outcome approach to causation models the distribution of $Y$ and $X$ given $C$ as the product of two parts; one which models the potential outcomes and one which models treatment assignment,

$$f(y, x|c, \theta, \gamma) \quad = \quad f(x|c, \gamma) \int f(y_{\{x\}}, y_{\{1-x\}}|c, \theta) dy_{\{1-x\}}.$$

The parameter $\theta$ models the potential outcomes, while $\gamma$ models the treatment assignment. Provided that we assign independent prior distributions to $\theta$ and $\gamma$, denoted $f(\theta, \gamma) = f(\theta) f(\gamma)$, then the missing data mechanism for the potential outcomes is ignorable [25], and Bayesian inference for the parameter $\theta$ proceeds from

$$f(\theta|y, x, c) \quad \propto \quad \int f(y_{\{x\}}, y_{\{1-x\}}|c, \theta) dy_{\{1-x\}} f(\theta). \tag{3.8}$$

According to this model for the observed data, knowledge of $\gamma$ conveys no information about $\theta$.

If $f(y_{\{1\}}, y_{\{x\}}|c)$ and $f(x|c)$ are indexed by different parameters which are *a priori* independent, then this implies that $P(X = 1|c)$ and $P(Y = 1|x, c)$ are also *a priori* independent. This, in turn, invalidates the BPSA model given in equation (3.1). Potential outcome models for causal inference view observational data as arising from two separate processes, one which generates the potential outcomes, and one which assigns treatment and masks potential outcomes. Because the models have distinct parameters, the investigator can ignore the manner in which treatment is assigned. The intuitive explanation for this is that, in a randomized experiment, *how* one chooses to randomize should convey no prior information about causal effects.

Furthermore, the assumption that $f(y_{\{1\}}, y_{\{x\}}|c)$ and $f(x|c)$ are indexed by different parameters which are *a priori* independent, implies that propensity scores are irrelevant to a Bayesian analysis [14]. If we look at the expression for $f(\theta|y, x, c)$ given above, we see that because of the way the model is specified, the parameter $\gamma$ which models the propensity score conveys no information about the parameter $\theta$. This may explain why there is so little published work combining Bayesian statistics with propensity score methods.

In order to make sense of BPSA, we should operate from the premise of using models which assume a relationship between $P(X = 1|c)$ and $P(Y = 1|x, c)$, even if such models seem unrealistic. This reasoning is used by Rubin [14] and Robins and Ritov [15] in the context of propensity score methods. For control of confounding in observational studies, Rubin [14] proposes BPSA when we have prior knowledge of the propensity scores. He argues that because specification of $P(Y = 1|x, z)$ may be easier than specification of $P(Y = 1|x, c)$, BPSA inferences may have good frequentist properties even though the model is not an accurate representation of the true data generating process. Similar logic is argued by Little in the context of survey sampling

[41]:

“...in practice all models are simplifications, and the features of the population that are important to include in the model vary according to the choice of design.... One way of limiting the effects of model misspecification is to restrict attention to models that yield design-consistent estimates.”

Hence while the BPSA model may be implausible, it can be viewed as a tool for building good frequentist procedures.

# Chapter 4

# Analysis of the statin data

In Section 1.1, we outlined a case-study of an observational study estimating the effect of statin therapy on mortality in Ontario patients discharged alive from hospital following acute myocardial infarction. We demonstrated that the crude association between statin therapy and mortality was likely biased due to confounding. Statins were generally prescribed to younger and healthier patients. Consequently, the reduction in mortality associated with statin use is partly driven by systematic differences in patient characteristics.

We apply BPSA and PSA to adjust for confounding. For each of the 4572 patients in the sample, we let Y equal one if the patient died within three years of discharge from hospital, and zero otherwise. We let X equal one if the patient was prescribed a statin at hospital discharge, and zero otherwise. We let C equal a 21 × 1 vector of measured covariates listed in Table 1.1, where the first term is equal to one in order to include an intercept term in the regression modelling. In this investigation, we assume that each of the variables in Table 1.1 is a true confounding variable regardless of the observed associations between $C$, $X$ and $Y$.

## 4.1 Conventional regression adjustment for confounding

To reduce confounding due to $C$, we estimate the effect of $X$ on $Y$ using conventional regression on measured confounders. We fit a logistic regression model of $Y$ on $X$ and $C$ with main effects and no interactions. Table 4.1 presents log ORs for the associations between measured covariates and mortality. The adjusted log OR for the association between of $X$ and $Y$ given $C$ is -0.33 with 95% confidence interval (-0.53, -0.13). The odds ratio, exp(-0.33) = 0.72, is closer to one compared to the crude OR, and is more consistent with the results of other population-based observational studies [17, 18]. This suggests that the analysis has reduced some of the confounding.

## 4.2 PSA analysis of the statin data

We apply PSA to the statin data. The method uses a two step procedure where the propensity scores are estimated for each patient and then the estimated propensity scores are included in a regression model for the mortality based on five quintile groups. We estimate the propensity scores using the fitted values from the logistic regression model given in equation (2.4) with main effects and no interactions. The treatment effect is then estimated via the model given in equation (2.5).

Table 4.2 gives estimates for the parameter $\gamma$ which is the log ORs for the association between $X$ and $C$. Before moving on to estimation of $\beta$ and $\xi$, we illustrate some of the properties of propensity scores. In Figure 4.1 we plot kernel density estimates of the density of propensity scores within treatment groups, given by $f(z|x)$ for $x = 0, 1$. The dashed and solid curves refer to the untreated group ($x = 0$) and

Table 4.1: Log ORs for the association between $Y$ and $(X, C)$.

| Characteristic | log OR | (95% CI) |
|---|---|---|
| Statin therapy | -0.33 | (-0.53, -0.13)* |
| *Demographic characteristics* | | |
| Age in years (mean) | 0.07 | (0.06, 0.08)** |
| Female sex | -0.19 | (-0.37, -0.01)* |
| *Presenting characteristics*$^+$ | | |
| Shock | 0.47 | (-0.38, 1.31) |
| *AMI risk factors*$^+$ | | |
| Family history of CAD | -0.19 | (-0.39, 0.01) |
| Diabetes | 0.40 | (0.21, 0.60)** |
| CVA/TIA | 0.36 | (0.12, 0.60)* |
| High BP | -0.08 | (-0.25, 0.10) |
| Current smoker | 0.23 | (0.02, 0.43)* |
| *Comorbidities*$^+$ | | |
| Angina | 0.27 | (0.10, 0.44)* |
| Renal disease | 0.54 | (-0.67, 1.75) |
| *Vital signs on admission*$^\dagger$ | | |
| Systolic BP | -0.01 | (-0.01, 0.00)** |
| Diastolic BP | 0.00 | (-0.01, 0.00) |
| Heart rate | 0.01 | (0.01, 0.02)** |
| Respiratory rate | 0.03 | (0.02, 0.05)** |
| *Laboratory values*$^\dagger$ | | |
| White blood count | 0.02 | (0.00, 0.03)* |
| Haemoglobin | -0.02 | (-0.02, -0.01)** |
| Sodium | -0.02 | (-0.04, 0.00) |
| Glucose | 0.02 | (0.00, 0.04)* |
| Creatinine | 0.00 | (0.00, 0.01)** |

* $p < 0.05$, ** $p < 0.001$

† Continous variables, + Dichotomous variables

Table 4.2: Point estimates for $\gamma$, the log ORs for the association between $X$ and $C$.

| Characteristic | log OR | (95% CI) |
|---|---|---|
| *Demographic characteristics* | | |
| Age (mean) | -0.03 | (-0.04, -0.02)$^{**}$ |
| Female sex | -0.17 | (-0.33, -0.01)$^{*}$ |
| *Presenting characteristics*$^{+}$ | | |
| Shock | -0.54 | (-1.54, 0.45) |
| *AMI risk factors*$^{+}$ | | |
| Family history of CAD | 0.16 | (0.02, 0.31)$^{*}$ |
| Diabetes | -0.05 | (-0.22, 0.12) |
| CVA/TIA | 0.17 | (-0.07, 0.40) |
| High BP | 0.31 | (0.17, 0.44)$^{**}$ |
| Current smoker | -0.27 | (-0.42, -0.12)$^{**}$ |
| *Comorbidities*$^{+}$ | | |
| Angina | 0.37 | (0.23, 0.51)$^{**}$ |
| Renal disease | 1.06 | (0.01, 2.11)$^{*}$ |
| *Vital signs on admission*$^{\dagger}$ | | |
| Systolic BP | 0.00 | (0.00, 0.00) |
| Diastolic BP | 0.00 | (-0.01, 0.00) |
| Heart rate | 0.00 | (-0.01, 0.00) |
| Respiratory rate | -0.01 | (-0.02, 0.00) |
| *Laboratory values*$^{\dagger}$ | | |
| White blood count | 0.00 | (-0.01, 0.01) |
| Haemoglobin | 0.00 | (0.00, 0.01) |
| Sodium | 0.01 | (0.01, 0.03) |
| Glucose | 0.01 | (0.00, 0.02) |
| Creatinine | 0.00 | (0.00, 0.00) |

* $p < 0.05$, ** $p < 0.001$

† Continous variables, + Dichotomous variables

treated group ($x = 1$) respectively. The curves were generated via Gaussian kernel density estimation using the R function density() with default settings. In Figure 4.2, we plot the estimated log odds of death, as a function of $x$ and $z$, given by logit$[P(Y = 1|x, z)]$. This quantity is estimated by second-order local polynomial regression of $Y$ on $Z$ for fixed $X$ using the R function loess() [39] with default settings. In Figure 4.2, the four vertical bars indicate the values of $c_1, c_2, c_3, c_4$ which define the five equal-sized quintile groups. Each of the intervals [0, 0.21), [0.21, 0.26), [0.26, 0.31), [0.31,0.38), [0.38, 1] contains an equal amount of the data. Consequently, estimates of logit$[P(Y = 1|x, z)]$ in the outermost quintiles are more imprecise. In Figure 4.2, we see that the dashed line is systematically higher than the solid line within quintile groups #1, #2, #3, #4. This suggests that statin therapy reduces mortality, in patients with a given propensity score. The curves are roughly parallel, indicating that the effect of $X$ on $Y$ is not modified by $Z$. Figure 4.2 reveals that high propensity scores are associated with a reduced risk of death. This is consistent with the literature on statin prescribing in Ontario residents [19] and the results of Table 1.1. Healthy patients are more likely to be treated with statins. In Figure 4.1, we see that untreated patients have lower propensity scores than treated patients. This is expected because $P(X = 1|z) = z$ implying that $X$ and $Z$ are dependent. Because $Z$ is simultaneously associated with $Y$ given $X$, and is also associated with $Y$ given $X$, this implies that $Z$ acts like a confounder, and Figures 4.1 and 4.2 allow us to visualize the confounding action of $C$.

To estimate the effect of $X$ on $Y$ while controlling for $C$, we apply PSA and stratify on quintiles of the propensity score. The results are given in Table 4.3. We see that risk of death is greatest in the 1st quintile and decreases in quintiles 2 though 5, as is illustrated in Figure 4.2. PSA assumes that there is no effect modification between $X$ and $Z$ and the log odds ratio of the effect of $X$ on $Y$ is -0.36 (-0.54, -0.18).

Figure 4.1: The distribution of propensity scores among the treated and the untreated, denoted $f(z|x)$ for $x = 0, 1$. The solid curve refers to the treated patients, while the dashed curve refers to the untreated patients.

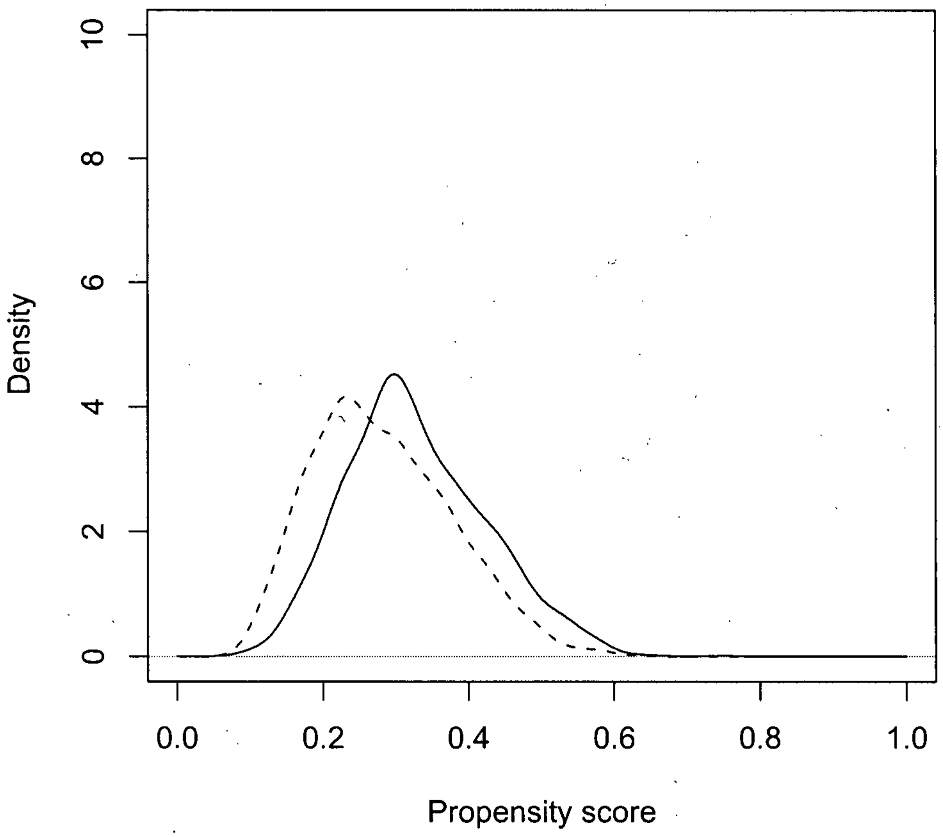


Figure 4.2: Log odds of death as a function of $x$ and $z$, denoted logit$[P(Y = 1|x, z)]$. The solid curve refers to the treated patients, while the dashed curve refers to the untreated patients.

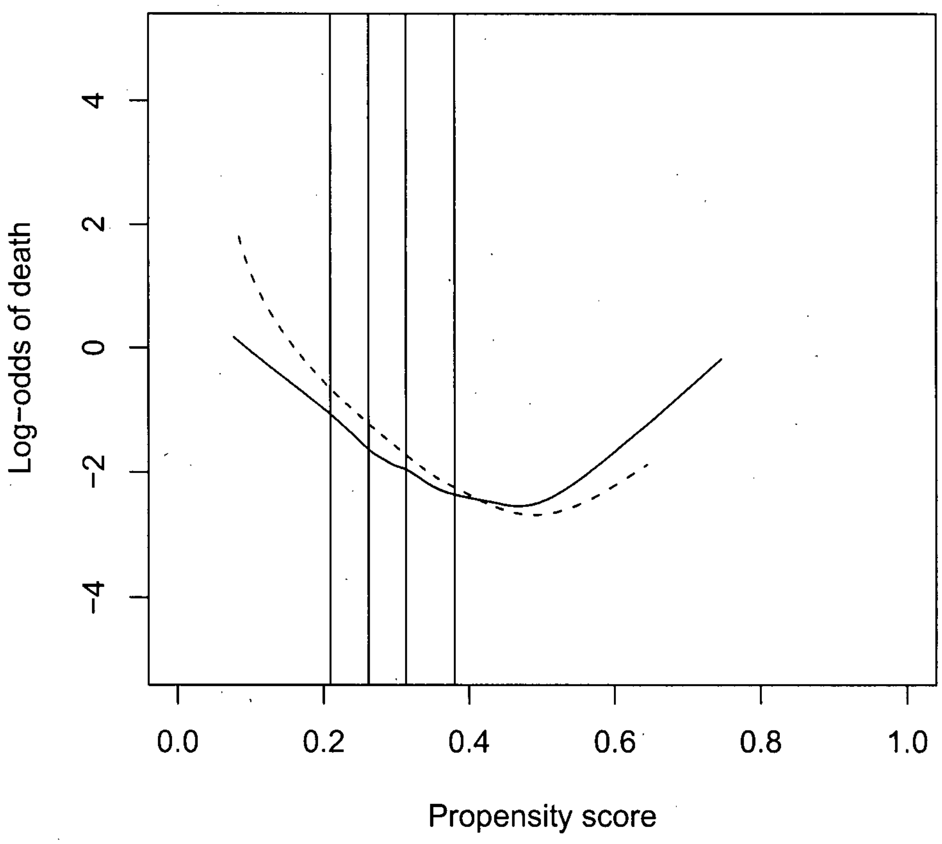


Table 4.3: Parameter estimates for the treatment effect $\beta$, and the baseline risk of $Y$ within quintile groups $(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$, calculated from PSA

| Parameter | log OR | (95% CI) |
|---|---|---|
| $\beta$ | -0.36 | (-0.54, -0.18) |
| $\xi_1$ | -0.13 | (-0.27, 0.00) |
| $\xi_2$ | -0.78 | (-0.98, -0.59) |
| $\xi_3$ | -1.44 | (-1.66, -1.22) |
| $\xi_4$ | -1.69 | (-1.93, -1.45) |
| $\xi_5$ | -2.18 | (-2.45, -1.90) |

### *Assessing the covariate balance produced by PSA*

To assess whether PSA succeeds in reducing confounding, we investigate the balancing properties of the propensity score. We examine the distribution of measured confounders among treated and untreated patients within quintile groups. The results are given in Table 4.4. The distributions of measured confounders within quintile groups are compared using two methods: by comparing sample means and proportions using t-tests, and by comparing the standardized difference between the distributions, calculated as the mean difference divided by the pooled standard deviation of the distributions [2].

Table 4.4 illustrates the ability of PSA to reduce confounding bias. Consider patient age which is a strong risk factor for mortality. In Table 1.1 we see that younger patients are more likely to be prescribed a statin. But within quintile groups, much of the systematic differences in age between treated and untreated patients is removed. Hence PSA reduces confounding bias due to age. The same effect is observed for other covariates which are strongly associated with treatment, such as angina.

Table 4.4: Means of measured covariates among treatment groups, within quintiles of the propensity score, calculated from PSA.

| | Quintile 1 | | Quintile 2 | | Quintile 3 | | Quintile 4 | | Quintile 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n = 914$ | | $n = 914$ | | $n = 914$ | | $n = 914$ | | $n = 916$ | |
| | Statin | | Statin | | Statin | | Statin | | Statin | |
| | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No |
| *Demographic characteristics* | | | | | | | | | | |
| Age | $77^{**\dagger}$ | $80^{**\dagger}$ | $72^{**\dagger}$ | $74^{**\dagger}$ | 67 | 67 | $62^{\dagger}$ | $60^{\dagger}$ | 53 | 52 |
| Female sex | $64\%^{\dagger}$ | $59\%^{\dagger}$ | 48% | 46% | 33% | 33% | 21% | 23% | $11\%^{\dagger}$ | $14\%^{\dagger}$ |
| *AMI risk factors* | | | | | | | | | | |
| Family history of CAD | 9.0% | 8.6% | 22% | 21% | 32% | 31% | 43% | 41% | 61% | 61% |
| Diabetes | $32\%^{\dagger}$ | $26\%^{\dagger}$ | 28% | 29% | 26% | 28% | 24% | 26% | 24% | 23% |
| CVA/TIA | 11% | 11% | 10% | 13% | 8.6% | 9.7% | 8.7% | 8.5% | $8.4\%^{\dagger}$ | $5.3\%^{\dagger}$ |
| High BP | 29% | 29% | 41% | 42% | 42% | 44% | 51% | 48% | 61% | 59% |
| Current smoker | $37\%^{*\dagger}$ | $27\%^{*\dagger}$ | $35\%^{\dagger}$ | $29\%^{\dagger}$ | 34% | 34% | 36% | 40% | $31\%^{*\dagger}$ | $37\%^{*\dagger}$ |
| *Comorbidities* | | | | | | | | | | |
| Angina | $7.6\%^{*\dagger}$ | $13\%^{*\dagger}$ | $22\%^{*\dagger}$ | $29\%^{*\dagger}$ | 34% | 35% | 39% | 39% | $57\%^{*\dagger}$ | $45\%^{**\dagger}$ |
| Renal disease | 0% | 0% | 0% | 0% | 0.3% | 0.5% | 0.3% | 0.3% | 2.1% | 1.5% |
| *Vital signs on admission* | | | | | | | | | | |
| Shock | 3% | 2% | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| Systolic BP | $148^{\dagger}$ | $144^{\dagger}$ | 149 | 149 | $147^{\dagger}$ | $151^{\dagger}$ | 151 | 151 | 149 | 149 |
| Diastolic BP | $86^{*\dagger}$ | $79^{*\dagger}$ | 82 | 82 | $83^{\dagger}$ | $86^{\dagger}$ | $85^{\dagger}$ | $86^{\dagger}$ | 87 | 87 |
| Heart rate | 94 | 92 | 85 | 83 | 82 | 83 | $78^{*\dagger}$ | $81^{*\dagger}$ | $78^{\dagger}$ | $76^{\dagger}$ |
| Respiratory rate | 24 | 24 | 21 | 21 | 20 | 20 | 20 | 20 | 19 | 19 |
| *Laboratory values* | | | | | | | | | | |
| White blood count | 11 | 11 | 10 | 10 | 10 | 10 | 9.7 | 10 | 9.9 | 10 |
| Haemoglobin | $132^{*\dagger}$ | $127^{*\dagger}$ | 134 | 135 | 140 | 139 | 142 | 143 | 146 | 147 |
| Sodium | 138 | 138 | 139 | 139 | 139 | 139 | 139 | 139 | $139^{\dagger}$ | $140^{\dagger}$ |
| Glucose | $10^{\dagger}$ | $9.8^{\dagger}$ | 9.5 | 9.4 | 9.2 | 9.5 | $8^{\dagger}$ | $9.2^{\dagger}$ | 10 | 9.4 |
| Creatinine | $110^{*\dagger}$ | $121^{*\dagger}$ | 103 | 103 | 97 | 101 | $101^{*\dagger}$ | $93^{*\dagger}$ | 99 | 96 |

* $p < 0.05$, ** $p < 0.001$, † Standardized difference $\geq 10\%$

In Table 4.4 we see that there are a number of statistically significant imbalances within quintile groups, particularly within the quintiles #1 and #5. But if the PSA model for $P(Y = 1|x, c)$ given in equation (3.1) is correct, then these imbalances do not necessarily indicate that the associations between $X$ and $Y$ within quintile groups are confounded. The difficulty with using Table 4.4 to identify residual confounding is that the table assumes nothing about the relationship between the propensity score and the outcome. Such an estimation strategy is valid in the sense that if we see balance in the distribution of confounders, then it suggests that there is no confounding. But the approach may be overly pessimistic. It focuses entirely on efforts to create comparable treatment and control groups, while ignoring prior information about the relationship between the outcome and the propensity score.

If the risk of $Y$ given $X$ is a smooth function of $Z$ within each interval $[0, c_1]$, $[c_1, c_2]$, $[c_2, c_3]$, $[c_3, c_4]$, $[c_4, 1]$, then systematic imbalances in the distribution of propensity scores within quintile groups do not necessarily indicate that there is confounding. To put this another way, covariate imbalances in Table 4.4 may cancel themselves out in such a way that all individuals within each interval have the same risk of $Y$. Furthermore, Figure 4.2 illustrates that smoothness assumptions may be reasonable. Modest change in the propensity score are associated with modest change in risk of mortality.

## 4.3 BPSA analysis of the statin data

Before applying BPSA to the statin data, we first set the bins $[0, c_1]$, $[c_1, c_2]$, $[c_2, c_3]$, $[c_3, c_4]$, $[c_4, 1]$ which define intervals of homogeneous risk of $Y$ given $X$, using the values $c_1 = 0.21$, $c_2 = 0.26$, $c_3 = 0.31$, $c_4 = 0.38$ from Section 4.2. We then apply BPSA to the statin data. We run a single MCMC chain of length 100 000 after

Table 4.5: Parameter estimates for the treatment effect $\beta$, and the baseline risk of $Y$ within each of the five bins $(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$, calculated from BPSA.

| Characteristic | log odds ratio (95% CI) |
|---|---|
| $\beta$ | -0.30 (-0.50, -0.10) |
| $\xi_1$ | 0.62 (0.38, 0.87) |
| $\xi_2$ | -1.04 (-1.39, -0.67) |
| $\xi_3$ | -1.93 (-2.27, -1.60) |
| $\xi_4$ | -3.04 (-3.45, -2.60) |
| $\xi_5$ | -3.98 (-4.43, -3.55) |

discarding the initial 10 000 iterations, and we then thin the chain by retaining every tenth iteration to obtain a sample of size 10 000. Thinning the chain is advantageous for computer storage of the analysis results. Sampler convergence for $\gamma$ was worse than for $\beta$ or $\xi$. Figures 4.3 and 4.4 contains mixing plots for $\beta$, $\xi$ and the first six components of $\gamma$ based on the last one thousand iterations of the thinned chain. We can see that samples for $\xi$ and $\beta$ move more rapidly through the target distribution than for $\gamma$. To assess the effect of sampler convergence, we obtained three additional MCMC chains of length 100 000 iterations with overdispersed starting values, and we inspected the marginal distributions of the components of $\beta$, $\xi$ and $\gamma$. The variation of sample means between MCMC chains was found to be small in relation to the variation within individual chains. Thus while mixing is not ideal, it does not appear to greatly affect estimation.

Table 4.5 contains point and interval estimates for $\beta$ and $\xi$ from BPSA. The log odds ratio for the effect of $X$ on $Y$ is similar to that of PSA and is given by -0.30 with 95% credible interval (-0.50, -0.10). But there are large differences in the estimates for $\xi$ which underscore the differences between BPSA and PSA. To illustrate why BPSA and PSA give different results, consider Table 4.6 which presents the distribution of

Figure 4.3: BPSA sampler convergence for $\beta$ and $\xi$.

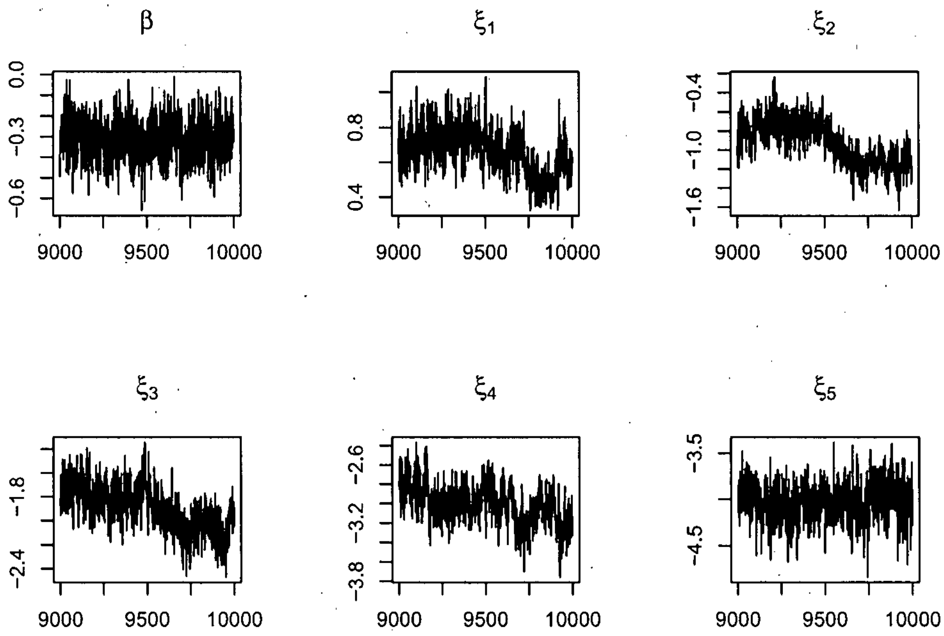
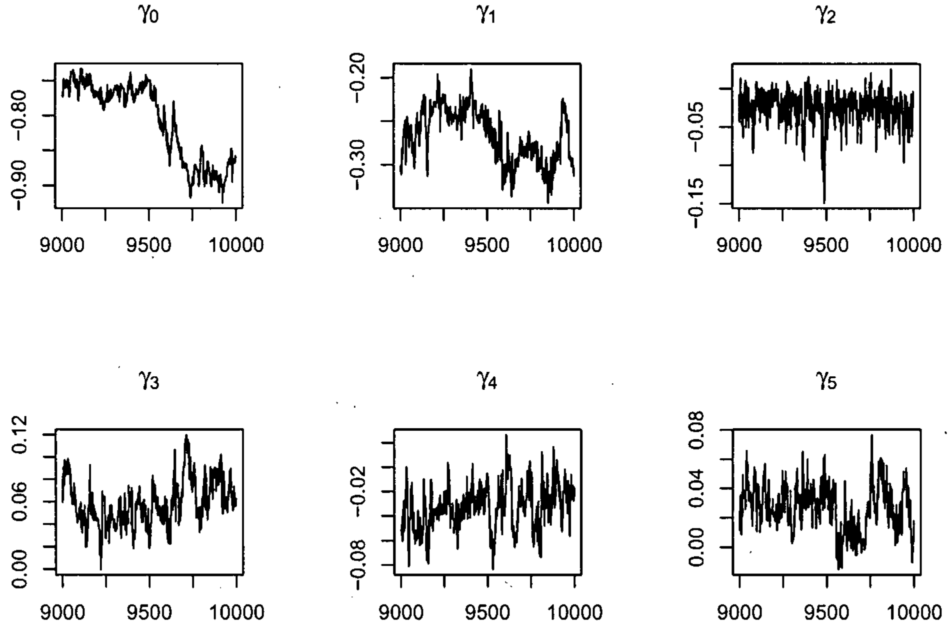


Figure 4.4: BPSA sampler convergence for $\gamma$.



measured confounders among treated and untreated patients within propensity score bins. In the table, each study unit is assigned to a bin based on the posterior mean of the propensity score. This means that for a unit with covariate C, we assign a level of $g(.)$ equal to $g(E\{z(C,\gamma)|\mathbf{y},\ \mathbf{x},\ \mathbf{c}\})$ where

$$E\{z(C,\gamma)|\mathbf{y},\ \mathbf{x},\ \mathbf{c}\} = \int \text{expit}(\gamma' C) f(\gamma|\mathbf{y},\ \mathbf{x},\ \mathbf{c}) d\gamma.$$

A comparison of Table 4.4 and Table 4.6 reveals striking differences. While PSA assigns an equal number of patients into each bin (roughly 4572/5=914), this is not the case for BPSA. From Table 4.6, we see that bins #1, #2, #3, #4, #5, contain 591, 781, 945, 1180, 1075 patients respectively. PSA classifies patients into bins using only information about the relationship between $C$ and $X$. If a confounder is strongly predictive of treatment, then this association is largely reduced after applying PSA because the variation in $C$ is re-distributed across propensity score bins.

In contrast, BPSA estimates propensity scores by incorporating modelling information about the relationship between $Y$ and $g(Z)$. The consequence is that BPSA assigns patients to bins based on how sick they are. For example, consider the indicator variable for diabetes. Diabetes is a strong risk factor for death with odds ratio 1.5 (1.2, 1.8) based on the conventional regression of Section 4.1 (see Table 4.1). For PSA, the prevalence of diabetes is fairly well balanced within each of the five bins because diabetes is not strongly associated with statin prescribing. In contrast, for BPSA, we see that the prevalence of diabetes is high in bins #1 and #2, while lower in bins #3, #4, and #5. In Figure 4.2, we see that patients with low propensity scores have the greatest risk of death. Consequently, when estimating the propensity scores, BPSA assigns sicker patients to bins #1 and #2. For every single covariate that is a strong risk factor for $Y$, we see a tendency for BPSA to assign patients with these risk factors to lower bins. Another example is renal disease. For PSA all patients with renal disease are in bins #3, #4 and #5, whereas for BPSA these same patients are assigned to bin #1.

What can we conclude about the differences between the BPSA and PSA analyses? Which analysis is more valid? Intuitively one might think that modelling uncertainty in the estimated propensity scores will only negligibly impact on inferences when the sample size is large. But this analysis indicates that the reverse may be true. As sample size increases, we get better estimates for $\xi$, the baseline risk of $Y$ within bins. This information flows back to affect propensity score estimation.

Furthermore, in Table 4.6 we see that BPSA yields fairly severe imbalances in the distribution of measured covariates between treated and untreated compared to PSA. The method does not appear to be producing homogeneous subgroups. What are the implications for the ability of BPSA to reduce confounding? We investigate these questions in Chapter 5. We study estimator performance in synthetic data under competing models for the outcome variable and various parameter values. We also consider the balance in $C$ induced by using BPSA versus PSA.

Table 4.6: Means of measured covariates among treatment groups, within bins of the propensity score, calculated from BPSA.

| | Bin 1 | | Bin 2 | | Bin 3 | | Bin 4 | | Bin 5 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $n = 591$ | | $n = 781$ | | $n = 945$ | | $n = 1180$ | | $n = 1075$ | |
| | Statin | | Statin | | Statin | | Statin | | Statin | |
| | Yes | No | Yes | No | Yes | No | Yes | No | Yes | No |
| *Demographics* | | | | | | | | | | |
| Age | 77**† | 81**† | 75**† | 78**† | 71*† | 72*† | 63 | 63 | 50 | 51 |
| Female sex | 48% | 51% | 47% | 49% | 38% | 42% | 26%† | 31%† | 15%*† | 20%*† |
| *AMI risk factors* | | | | | | | | | | |
| Family history of CAD | 12%† | 8%† | 19%† | 14%† | 29% | 25% | 39% | 38% | 59% | 57% |
| Diabetes | 63%**† | 41%**† | 46%*† | 34%*† | 33% | 29% | 21% | 22% | 10% | 11% |
| CVA/TIA | 31%† | 26%† | 18% | 15% | 13%† | 9%† | 5%*† | 3%*† | 1% | 2% |
| High BP | 64%*† | 46%*† | 55%† | 48%† | 53%*† | 44%*† | 43% | 41% | 43% | 38% |
| Current smoker | 24% | 20% | 20% | 19% | 23%*† | 29%*† | 36% | 39% | 48% | 52% |
| *Comorbidities* | | | | | | | | | | |
| Angina | 40% | 38% | 50%*† | 40%*† | 44%*† | 35%*† | 35%**† | 26%**† | 28%**† | 19%**† |
| Renal disease | 10%*† | 2%*† | 0% | 0% | 0% | 0% | 0% | 0% | 0% | 0% |
| *Vital signs on admission* | | | | | | | | | | |
| Shock | 3% | 3% | 1% | 1% | 0% | 0% | 0% | 0% | 0% | 0% |
| Systolic BP | 140 | 139 | 147 | 145 | 148 | 150 | 149† | 152† | 152 | 152 |
| Diastolic BP | 80† | 78† | 81 | 80 | 82 | 82 | 84*† | 86*† | 90 | 90 |
| Heart rate | 113**† | 100**† | 93† | 90† | 83 | 81 | 76*† | 79*† | 74 | 74 |
| Respiratory rate | 26 | 26 | 22 | 22 | 21 | 20 | 19 | 20 | 19 | 19 |
| *Laboratory values* | | | | | | | | | | |
| White blood count | 14† | 13† | 11† | 10† | 10 | 10 | 9.6 | 9.7 | 9.6 | 9.7 |
| Haemoglobin | 124† | 119† | 129 | 131 | 137 | 137 | 143 | 144 | 149 | 150 |
| Sodium | 137 | 137 | 139 | 138 | 139 | 139 | 139 | 139 | 140*† | 140*† |
| Glucose | 14*† | 12*† | 11† | 10† | 10.0 | 9.4 | 8.8 | 8.9 | 8.0 | 7.7 |
| Creatinine | 192† | 165† | 111† | 107† | 99† | 95† | 92**† | 87**† | 86 | 85 |

* $p < 0.05$, ** $p < 0.001$, † Standardized difference $\geq 10\%$

## 4.4 Decomposition of the posterior variance for $\beta$ and $\xi$

One strategy for investigating the difference between BPSA and PSA is to study the effect of admitting uncertainty in the estimated propensity scores on the posterior variance of $\beta$ and $\xi$. Gustafson and Clarke discuss factorizing the posterior variance of model parameters in the context of hierarchical models [42]. Using the relation $Var[A] = E\{Var[A|B]\} + Var\{E[A|B]\}$, they note that for hierarchical models with data $D$, parameter $\theta$ and hyperparameter $\phi$, we can write the posterior variance for $\theta$ as

$$Var[\theta|D] = E\{Var[\theta|D,\phi]|D\} + Var\{E[\theta|D,\phi]|D\}. \tag{4.1}$$

When the parameter $\phi$ is known a priori, the model is not hierarchical because there is a single prior distribution for $\theta$. Then the quantity $Var\{E[\theta|D,\phi]|D\}$ is equal to zero and the posterior variance for $\theta$ is given by $E\{Var[\theta|D,\phi]|D\}$. Consequently, we can conceptualize the term $Var\{E[\theta|D,\phi]|D\}$ as modelling the extent to which admitting uncertainty in the hyperparameter $\phi$ increases the posterior variance. This gives an "ANOVA" type representation for decomposing posterior variance in latent variable models. The first term on the right hand side of the equation models variation "within" specific priors, while the second term models variation in the posterior arising "between" different choices of priors.

The BPSA method can be investigated within this framework. When estimating $\beta$ and $\xi$, PSA uses a degenerate prior distribution for $\gamma$ which is equal to the maximum likelihood estimator from logistic regression of $X$ on $C$. In contrast, BPSA estimates $\beta$ and $\xi$ by admitting uncertainty in $\gamma$ and modelling $\gamma$ as a latent quantity. We can write

$$Var[\beta, \xi|D] = E\{Var[\beta, \xi|D, \gamma]|D\} + Var\{E[\beta, \xi|D, \gamma]|D\}, \tag{4.2}$$

where $D = (\mathbf{y}, \mathbf{x}, \mathbf{c})$. Letting $\hat{\gamma}$ denote the maximum likelihood estimator for $\gamma$ from PSA, the expression $E\{Var[\beta, \xi|D, \hat{\gamma}]|D\} = Var[\beta, \xi|D, \hat{\gamma}]$ equals the posterior variance for PSA. For BPSA, the quantity $E\{Var[\beta, \xi|D, \gamma]|D\}$ models the within variation of $\beta$ and $\xi$ given $\gamma$, while the quantity $Var\{E[\beta, \xi|D, \gamma]|D\}$ models the between variation over levels of $\gamma$.

Thus to characterize the different inferences for BPSA and PSA we can factorize the posterior variance for the statin data and calculate both $Var[\beta, \xi|D, \hat{\gamma}]$ and the quantities in equation (4.2). We can calculate

$$\frac{E\{Var[\beta, \xi|D, \gamma]|D\}}{E\{Var[\beta, \xi|D, \gamma]|D\} + Var\{E[\beta, \xi|D, \gamma]|D\}}$$

which is the variation in $\beta, \xi$ within levels of $\gamma$, divided by the posterior variance. If this quantity is close to one, then admitting uncertainty in $\gamma$ has little effect on posterior variance.

A simple estimate of this ratio for BPSA is provided from the MCMC algorithm described in Section 4.1. When updating $\beta$ and $\xi$, given some current value $\gamma^*$, the proposal distribution is given by the normal approximation to the asymptotic distribution of the maximum likelihood estimator from logistic regression of $Y$ on $X$ and the quintile group $G$. The mean and variance of this distribution are estimates of $E[\beta, \xi|D, \gamma^*]$ and $Var[\beta, \xi|D, \gamma^*]$. Thus if we collect and store these quantities during simulation along with the sample model parameters, we can take the empirical average across sampled values of $\gamma$ to estimate $E\{Var[\beta, \xi|D, \gamma]|D\}$ and $Var\{E[\beta, \xi|D, \gamma]|D\}$.

Table 4.7: Posterior variances for $\beta$ and $\xi$ from the analysis of the statin data using BPSA and PSA

| | Posterior variance from PSA | Posterior variance from BPSA | Decomposition of BPSA posterior variance | | |
|---|---|---|---|---|---|
| | | | Within | Between | Within/Total |
| $\beta$ | 0.009 | 0.010 | 0.010 | 0.001 | 0.948 |
| $\xi_1$ | 0.005 | 0.016 | 0.008 | 0.010 | 0.445 |
| $\xi_2$ | 0.010 | 0.034 | 0.013 | 0.019 | 0.412 |
| $\xi_3$ | 0.013 | 0.030 | 0.015 | 0.015 | 0.498 |
| $\xi_4$ | 0.015 | 0.049 | 0.021 | 0.030 | 0.407 |
| $\xi_5$ | 0.020 | 0.050 | 0.043 | 0.016 | 0.729 |

Table 4.7 presents a decomposition of the posterior variances of $\beta$ and $\xi$ from BPSA and PSA. The first two columns are posterior variance for BPSA versus PSA. The final three columns give the decomposition of the posterior variance for BPSA. The main result is that the ratio of the within variation divided by total variation is estimated as 0.948 for the parameter $\beta$ and is much smaller for each component of $\xi$. In other words, admitting uncertainty in the estimated propensity score greatly increases posterior uncertainty $\xi$, but only marginally increases uncertainty in $\beta$. This makes sense when we compare the interval estimates from BPSA and PSA in Tables 4.3 and 4.5. Interval estimates for $\xi$ from BPSA are roughly twice as wide compared to PSA, whereas for the parameter $\beta$, BPSA only modestly increases posterior uncertainty compared to PSA.

# Chapter 5

# Simulation studies of the performance of BPSA and PSA

In Chapter 4, we applied BPSA to the statin data and contrasted the results with those from PSA. Because of the large sample size, intuition suggests that there should be little uncertainty in the estimated propensity score, and we would expect both methods to give similar inferences. But the statin data example shows that this may not be the case. The estimates for $\beta$ are comparable, but we see large differences in the point estimates for $\xi$. BPSA groups patients into propensity score bins based on their health status, and the result is that the estimated components of $\xi$ are spread apart compared to PSA. Sick patients are grouped into bin #1 and this drove the estimate of $\xi_1$ upwards to reflect the fact that this group had greater risk of death. Similarly, healthy patients were grouped into bin #5. A plot of the risk of death as a function of the propensity score, given in Figure 4.2, shows that this approach seems reasonable because patients with high propensity scores have lower mortality. But the question remains whether or not BPSA inferences are more valid. Can we expect that the frequentist inferences from BPSA will be superior in some settings? If so, then what are these settings and how do they compare to those for the statin data?

In this chapter we investigate the frequentist performance of BPSA estimates using simulations and further analysis of the statin data. In Section 5.1, we use simulations to study the bias, relative efficiency and relative mean squared error (MSE) of point estimators, and the coverage probability and length of interval estimates. BPSA has the potential drawback that it relies more heavily on modelling assumptions for the outcome variable than PSA. The method uses a model for $P(Y = 1|x, c)$ for propensity score estimation, while PSA does not. Consequently, we might expect that PSA inferences are more robust to model misspecification. To investigate further, Section 5.2 investigates the performance of BPSA and PSA when applied to synthetic data generated under competing models for the outcome. In Section 5.3, we consider prediction error. We use cross-validation techniques to study the ability of BPSA and PSA to accurately forecast the outcome variable when applied to real and synthetic data. Finally, Section 5.4 revisits the idea of covariate balance induced by BPSA versus PSA. Chapter 4 illustrated that BPSA appears to produce treatment and control groups which less similar with respect to measured confounders, compared to PSA. Thus the method does not appear to be effectively reducing confounding. We explore covariate balance using simulations.

## 5.1 Simulation study when the distribution of the outcome follows the propensity score model

We use simulations to investigate the frequentist performance of point and interval estimators for $\beta, \xi, \gamma$ when the data are generated *according to the BPSA model* given in equations (3.1) and (3.2), for fixed parameter values of $\beta, \xi$ and $\gamma$.

*Simulation design*

We consider the case where $C$ has four continuous components (meaning that $C$

is a $5 \times 1$ vector with first component equal to one), and we simulate datasets for four different choices of model parameters,

| Design | $\beta$ | $\xi$ | $\gamma$ |
|---|---|---|---|
| #1 | -1/2 | (1/2, -1/2, 1/2, -1/2, 1/2) | (1/2, -1/2, 1/2, -1/2, 1/2) |
| #2 | -1/2 | (2, -2, 2, -2, 2) | (1/2, -1/2, 1/2, -1/2, 1/2) |
| #3 | -1/2 | (1/2, -1/2, 1/2, -1/2, 1/2) | (2, -2, 2, -2, 2) |
| #4 | -1/2 | (2, -2, 2, -2, 2) | (2, -2, 2, -2, 2) |

and sample size $n = 1000$.

Design #1 models the setting where there are strong associations between between $Y$, $X$ and bin membership, with odds ratios of either $\exp(-1/2) = 0.61$ or $\exp(1/2) = 1.64$. Designs #2, #3 and #4 are similar, but use more extreme odds ratios of either $\exp(-2) = 0.13$ or $\exp(2) = 7.4$. While these designs are less realistic, they can indicate settings in which BPSA or PSA break down. Design #2 is of particular interest. The components of $\xi$ are heterogeneous while the components of $\gamma$ are quite similar. We expect that PSA will misclassify patients into the wrong propensity score bins and this will adversely affect estimation of $\xi$ and $\beta$.

For each design and fixed sample size, we generate and analyze 400 simulated datasets, to yield a sample of 400 point estimates and 80% interval estimates for $\beta$, $\xi$ and $\gamma$, using the following algorithm:

1. Generate the $n \times 5$ design matrix $\mathbf{c}$. The first column is a column of ones. The latter four columns are the sampled covariates for the dataset of size $n$. Each element of each column is simulated as an independent draw from a N(0,1) random variable.

2. Generate the $n \times 1$ vector $\mathbf{x}$ using the logistic regression model of equation (3.2), given by

$$\text{logit}[P(X = 1|c)] = \gamma' c$$

where $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ is a $5 \times 1$ vector.

3. Because of the way the simulation is designed, we have $\gamma' C \sim N(\gamma_0, \sum_{i=1}^{4} \gamma_i^2)$ for fixed $\gamma$. Thus the values $c_1, c_2, c_3, c_4$ defining the true quintiles of the propensity score are given exactly by $c_k = \text{expit}\{\gamma_0 + (\sum_{i=1}^{4} \gamma_i^2) q_k\}$ for $k = 1, 2, 3, 4$ and $q_k = \Phi^{-1}(0.2k)$, where $\Phi^{-1}(.)$ is the quantile function of a N(0,1) random variable. Generate the $n \times 1$ vector $\mathbf{y}$ using the logistic regression model of (3.1) given by

$$\text{logit}[P(Y = 1|x, z)] = \beta x + g(z(c, \gamma))' \xi$$

where

$$g(z) = \begin{cases} [1, 0, 0, 0, 0] & \text{if } z(c, \gamma) < c_1 \\ [1, 1, 0, 0, 0] & \text{if } c_1 \leq z(c, \gamma) < c_2 \\ [1, 0, 1, 0, 0] & \text{if } c_2 \leq z(c, \gamma) < c_3 \\ [1, 0, 0, 1, 0] & \text{if } c_3 \leq z(c, \gamma) < c_4 \\ [1, 0, 0, 0, 1] & \text{if } c_4 < z(c, \gamma) \end{cases}$$

and $z(c, \gamma) = \text{expit}(\gamma' c)$.

4. Analyze the datasets using PSA and BPSA with an MCMC chain of length 10 000 after discarding 500 initial iterations. Obtain point and 80% interval estimates for $\beta, \xi$ and $\gamma$ from each method.

Careful tuning of the MCMC sampler is needed for each simulation design, and this is accomplished using separate trial simulation runs.

*Results*

Table 5.1 summarizes the performance of point and interval estimators for $\beta$, $\xi$ and $\gamma$ from BPSA and PSA, in the case where datasets are simulated according to Design #1. The first two columns indicate of the magnitude of bias of each method. Each cell contains the sample mean of the collection of 400 sampled estimators, as well as the z-score for the sample mean relative to the true underlying parameter value. The z-score was calculated as

$$\text{z-score for mean of simulated point estimators} = \frac{\text{sample mean - true parameter value}}{\text{sample standard deviation}}.$$

A large z-score indicates that the point estimator is biased. The third and fourth columns in the table contain the estimated relative efficiency and the relative MSE of BPSA point estimators compared to PSA. These quantities are calculated as

$$\text{Estimated relative efficiency} = \frac{\text{Sample variance of BPSA point estimates}}{\text{Sample variance of PSA point estimates}}$$

and

$$\text{Estimated relative MSE} = \frac{\text{Sample average squared error of BPSA point estimates}}{\text{Sample average squared error of PSA point estimates}}.$$

To aid with interpretation of results, we calculated *simulation standard errors* for the relative efficiency and relative MSE estimates via the bootstrap. In Table 5.1, estimates denoted with a "*" imply that a 90% bootstrapped confidence interval for the parameter excludes 1. If these estimates are significantly less than one, it indicates that BPSA point estimates have smaller variance or smaller MSE compared to PSA. The final four columns contain estimates of the coverage probability and average length of interval estimates. The symbol † indicates that a 90% confidence intervals for the coverage probability excludes the nominal level of 80%. To clarify, we estimate the coverage probabilities which are nominally equal to 80%, and we construct 90% confidence interval for these quantities. We use an $\alpha$-level of 0.1 for the analysis of simulation results in order to increase power at the expense of the Type I error rate. Tables 5.2, 5.3 and 5.4 are identical to Table 5.1, but correspond to data simulated under Designs #2-4.

The results of the simulation study indicate that BPSA and PSA perform very comparably in terms of estimation of $\beta$. The quality of inferences for $\beta$ from BPSA and PSA are so similar that any systematic differences in performance is swamped by variation from the Monte Carlo simulation. However, a qualitative assessment of the results across the four simulation designs suggests that PSA point estimators are more efficient than for BPSA. PSA interval estimators also appear to have lower coverage probability.

Inferences for $\xi$ calculated from BPSA are generally superior to those from PSA. BPSA point estimates of $\xi$ have similar efficiency to those from PSA, but they have smaller bias and this reduces overall MSE. In Tables 5.1 through 5.4, the estimated relative efficiencies of point estimators for $\xi$ are generally not systematically different from one across the simulation designs. But for the parameters $\xi$, the z-scores calculated from BPSA are much smaller than for PSA. For example, under Design #2, PSA point estimators are biased with z-scores of greater than 20. This reduction in bias from using BPSA for estimating $\xi$ reduces overall MSE compared to PSA. In all four simulation designs, the estimated relative MSEs tend to be significantly less than one.

The improvements of BPSA compared to PSA also applies to interval estimation of $\xi$. For each simulation design, BPSA credible intervals have greater average length and estimated coverage probabilities which do not differ significantly from 80%. In contrast, PSA interval estimates always have lower coverage probability, and in Design #2, the estimated coverage probabilities are never greater than 50%. The increase in average interval length for BPSA is usually modest, but is occasionally quite substantial. Nonetheless, the increased length appears to be justifiable because the resulting interval estimates have proper coverage probability.

BPSA point and interval estimates of $\gamma$ also appear to have better performance. Thus BPSA appears to do a better job of estimating the propensity scores. In terms of bias, BPSA and PSA are similar, but BPSA point estimators are much more efficient. Under each of the four simulation designs, the estimated relative efficiencies for BPSA compared to PSA are significantly less than one. Under Designs #2 and #4, the relative efficiencies were typically less than 0.2, meaning that BPSA point estimators of $\gamma$ are perhaps five times more efficient. Consequently, in the four simulation studies, the BPSA estimators of $\gamma$ have smaller MSE compared to PSA.

The performance of interval estimates for $\gamma$ are also improved for BPSA. Under Design #4, BPSA interval estimates have length which is roughly one fifth of that of PSA, and yet they maintain roughly nominal coverage probability. BPSA interval estimates of $\gamma$ tend to have lower coverage probability compared to PSA. This is particularly true under Design #2. In this case, the empirical variance of the BPSA point estimates of $\gamma$ is far greater than the posterior variances. Thus BPSA tends to modestly under report uncertainty in $\gamma$. One possible explanation for this under-coverage is poor MCMC mixing. The sampler may not fully explore the posterior distribution.

Table 5.1: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #1.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean | PSA Sample mean | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | (z-score) | (z-score) | | | Coverage | Length | Coverage | Length |
| $\beta = -0.5$ | -0.51 (-1.7) | -0.49 (0.9) | 1.05* | 1.06* | 0.84 | 0.38 | 0.84 | 0.37 |
| $\xi_1 = 0.5$ | 0.49 (-1.0) | 0.45 (-7.4) | 1.07 | 0.94 | 0.84 | 0.40 | 0.80 | 0.38 |
| $\xi_2 = -0.5$ | -0.46 (3.6) | -0.29 (18.4) | 1.04 | 0.58* | 0.78 | 0.58 | 0.60† | 0.53 |
| $\xi_3 = 0.5$ | 0.52 (1.5) | 0.33 (-15.8) | 1.20* | 0.74* | 0.78 | 0.59 | 0.64† | 0.54 |
| $\xi_4 = -0.5$ | -0.48 (1.6) | -0.26 (21.7) | 0.92 | 0.43* | 0.81 | 0.57 | 0.57† | 0.55 |
| $\xi_5 = 0.5$ | 0.53 (2.4) | 0.47 (-2.1) | 0.96 | 0.96 | 0.81 | 0.59 | 0.78 | 0.57 |
| $\gamma_0 = 0.5$ | 0.49 (-3.3) | 0.50 (-1.0) | 0.16* | 0.16* | 0.79 | 0.07 | 0.78 | 0.19 |
| $\gamma_1 = -0.5$ | -0.51 (-5.8) | -0.51 (-2.2) | 0.22* | 0.24* | 0.77 | 0.09 | 0.79 | 0.19 |
| $\gamma_2 = 0.5$ | 0.51 (4.5) | 0.51 (1.8) | 0.23* | 0.24* | 0.77 | 0.09 | 0.78 | 0.19 |
| $\gamma_3 = -0.5$ | -0.51 (-4.7) | -0.50 (-0.6) | 0.32* | 0.34* | 0.76† | 0.09 | 0.81 | 0.19 |
| $\gamma_4 = 0.5$ | 0.51 (3.9) | 0.51 (1.9) | 0.26* | 0.26* | 0.76† | 0.09 | 0.80 | 0.19 |

* Quantity differs from 1, $p < 0.1$, † Coverage probability is less than 80%, $p < 0.1$

Table 5.2: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #2.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean | PSA Sample mean | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | (z-score) | (z-score) | | | Coverage | Length | Coverage | Length |
| $\beta = -0.5$ | -0.50 (0.4) | -0.42 (9.7) | 1.24* | 1.00 | 0.81 | 0.48 | 0.75$^\dagger$ | 0.44 |
| $\xi_1 = 2$ | 2.00 (0.0) | 1.71 (-20.3) | 0.73* | 0.36* | 0.78 | 0.56 | 0.43$^\dagger$ | 0.51 |
| $\xi_2 = -2$ | -1.96 (1.6) | -1.39 (30.6) | 1.55 | 0.47* | 0.82 | 0.66 | 0.24$^\dagger$ | 0.62 |
| $\xi_3 = 2$ | 2.19 (5.5) | 0.47 (-56.9) | 1.66* | 0.20* | 0.82 | 1.43 | 0.03$^\dagger$ | 0.77 |
| $\xi_4 = -2$ | -1.97 (1.1) | -1.34 (37.4) | 1.72 | 0.38* | 0.80 | 0.68 | 0.20$^\dagger$ | 0.64 |
| $\xi_5 = 2$ | 2.16 (5.1) | 1.30 (-19.0) | 0.71* | 0.40* | 0.81 | 1.37 | 0.35$^\dagger$ | 0.97 |
| $\gamma_0 = 0.5$ | 0.50 (-1.5) | 0.50 (-0.3) | 0.59* | 0.59* | 0.68$^\dagger$ | 0.01 | 0.79 | 0.19 |
| $\gamma_1 = -0.5$ | -0.50 (-0.9) | -0.50 (-1.2) | 0.14* | 0.14* | 0.70$^\dagger$ | 0.01 | 0.80 | 0.19 |
| $\gamma_2 = 0.5$ | 0.50 (0.8) | 0.51 (1.6) | 0.14* | 0.14* | 0.74$^\dagger$ | 0.01 | 0.80 | 0.19 |
| $\gamma_3 = -0.5$ | -0.50 (-0.8) | -0.51 (-2.3) | 0.12* | 0.12* | 0.70$^\dagger$ | 0.01 | 0.77 | 0.19 |
| $\gamma_4 = 0.5$ | 0.50 (1.1) | 0.51 (1.8) | 0.13* | 0.13* | 0.68$^\dagger$ | 0.01 | 0.78 | 0.19 |

* Quantity differs from 1, $p < 0.1$, $^\dagger$ Coverage probability is less than 80%, $p < 0.1$

Table 5.3: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #3.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean | PSA Sample mean | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | (z-score) | (z-score) | | | Coverage | Length | Coverage | Length |
| $\beta = -0.5$ | -0.51 (-0.5) | -0.49 (0.5) | 1.03 | 1.03 | 0.80 | 0.58 | 0.81 | 0.57 |
| $\xi_1 = 0.5$ | 0.49 (-0.9) | 0.48 (-3.4) | 1.01 | 0.98 | 0.81 | 0.38 | 0.79 | 0.37 |
| $\xi_2 = -0.5$ | -0.47 (2.6) | -0.41 (7.6) | 1.02 | 0.91* | 0.77 | 0.60 | 0.74† | 0.57 |
| $\xi_3 = 0.5$ | 0.52 (1.1) | 0.44 (-4.4) | 1.01 | 0.97 | 0.80 | 0.73 | 0.76† | 0.71 |
| $\xi_4 = -0.5$ | -0.49 (1.0) | -0.38 (8.2) | 1.01 | 0.87* | 0.82 | 0.76 | 0.78 | 0.74 |
| $\xi_5 = 0.5$ | 0.53 (1.7) | 0.48 (-1.6) | 0.95* | 0.95* | 0.80 | 0.77 | 0.79 | 0.75 |
| $\gamma_0 = 2$ | 2.00 (-0.2) | 2.03 (3.2) | 0.25* | 0.24* | 0.78 | 0.19 | 0.80 | 0.42 |
| $\gamma_1 = -2$ | -2.02 (-3.2) | -2.03 (-3.1) | 0.34* | 0.34* | 0.78 | 0.25 | 0.81 | 0.44 |
| $\gamma_2 = 2$ | 2.01 (2.2) | 2.03 (3.4) | 0.36* | 0.35* | 0.82 | 0.25 | 0.82 | 0.44 |
| $\gamma_3 = -2$ | -2.02 (-4.1) | -2.04 (-4.1) | 0.41* | 0.41* | 0.75† | 0.26 | 0.80 | 0.44 |
| $\gamma_4 = 2$ | 2.01 (1.9) | 2.03 (3.7) | 0.37* | 0.36* | 0.79 | 0.25 | 0.81 | 0.44 |

* Quantity differs from 1, $p < 0.1$, † Coverage probability is less than 80%, $p < 0.1$

Table 5.4: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #4.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean | PSA Sample mean | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | (z-score) | (z-score) | | | Coverage | Length | Coverage | Length |
| $\beta = -0.5$ | -0.51 (-0.9) | -0.48 (1.6) | 1.05 | 1.04 | 0.78 | 0.66 | 0.77 | 0.65 |
| $\xi_1 = 2$ | 2.03 (2.5) | 1.89 (-9.8) | 0.95 | $0.78^{*}$ | 0.77 | 0.56 | $0.65^{\dagger}$ | 0.54 |
| $\xi_2 = -2$ | -2.01 (-0.9) | -1.75 (16.2) | 0.92 | $0.56^{*}$ | 0.82 | 0.71 | $0.60^{\dagger}$ | 0.70 |
| $\xi_3 = 2$ | 2.15 (4.3) | 1.09 (-30.3) | $1.32^{*}$ | $0.42^{*}$ | 0.80 | 1.46 | $0.26^{\dagger}$ | 1.06 |
| $\xi_4 = -2$ | -2.00 (-0.2) | -1.64 (19.5) | 1.22 | $0.63^{*}$ | 0.81 | 0.90 | $0.55^{\dagger}$ | 0.87 |
| $\xi_5 = 2$ | 2.14 (2.9) | 1.52 (-7) | 0.52 | 0.48 | 0.81 | 1.46 | $0.46^{\dagger}$ | 1.19 |
| $\gamma_0 = 2$ | 2.00 (-0.9) | 2.04 (4.6) | $0.13^{*}$ | $0.13^{*}$ | 0.80 | 0.04 | 0.82 | 0.42 |
| $\gamma_1 = -2$ | -2.00 (-0.1) | -2.03 (-4.3) | $0.09^{*}$ | $0.08^{*}$ | $0.73^{\dagger}$ | 0.05 | 0.83 | 0.44 |
| $\gamma_2 = 2$ | 2.00 (-1.0) | 2.04 (4.9) | $0.08^{*}$ | $0.08^{*}$ | 0.78 | 0.05 | 0.81 | 0.44 |
| $\gamma_3 = -2$ | -2.00 (1.1) | -2.04 (-4.4) | $0.07^{*}$ | $0.06^{*}$ | 0.79 | 0.05 | 0.79 | 0.44 |
| $\gamma_4 = 2$ | 2.00 (-1.0) | 2.03 (4.0) | $0.06^{*}$ | $0.06^{*}$ | 0.79 | 0.05 | 0.79 | 0.44 |

$^{*}$ Quantity differs from 1, $p < 0.1$, $^{\dagger}$ Coverage probability is less than 80%, $p < 0.1$

*Discussion*

The results of the simulation study indicate that BPSA does a better job of estimating $\xi$ and $\gamma$. For estimation of $\beta$, the methods are very comparable, with BPSA point estimators appearing to perform slightly worse under the "realistic" Design #1. Contrary to intuition, differences in inferences between BPSA and PSA may be substantial. This is particularly true when the true data generating mechanism is given by Designs #2 or #4. To explain this behavior, we observe that Design #2 involves large values of $\xi$ and small values of $\gamma$. Because the components of $\gamma$ are similar, correct classification of patients to bins using PSA is error prone. Because the components of $\xi$ are large and heterogeneous, bin misclassification adversely affects estimation of $\xi$ since study subjects with very different outcome risks are being grouped together. Under Design #2, patient outcomes contribute a great deal of information about the propensity scores. BPSA has an advantage because it uses this information while PSA ignores it.

While the simulations demonstrate that BPSA can perform well compared to PSA, the findings may not generally apply to the analysis of typical epidemiologic data. In the statin data example of Section 4, the risk of $Y$ as a function of the propensity score is not particularly heterogeneous. In Figure 4.2 we see that this risk function is a fairly smooth function of $Z$. Future simulations studies could try to more carefully mimic real observational studies.

It is interesting that PSA point estimates of $\beta$ are more efficient than for BPSA. Note that $\beta$ is the primary parameter of interest because it models the treatment effect. The parameters $\xi$ and $\gamma$ are essentially nuisance parameters. Because BPSA does a better job of estimating the propensity scores, we would expect that this should improve control of confounding and estimation of $\beta$. BPSA and PSA both use estimated propensity scores, but BPSA incorporates modelling information about the relationship between $Y$ and $Z$ given $X$. While the simulations indicate that this improves estimation of $\gamma$ and $\xi$, the BPSA approach may be harmful to the efficiency of point estimation of $\beta$. For interval estimation of $\beta$, BPSA seems to perform favorably. For example, under Design #2, BPSA interval estimates of $\beta$ have nominal coverage probability while those calculated from PSA do not.

A feature of our simulation study is that we have evaluated BPSA and PSA assuming that the quantities $c_1, c_2, c_3, c_4$ which define propensity score bins are known. Because of the manner in which the data are simulated, we can determine the quintiles of $Z = P(X = 1|C)$ exactly. In contrast, in the practical application of BPSA, $c_1, c_2, c_3, c_4$ are unknown. We recommend estimating the propensity scores for each subject using PSA, and then calculating the quintiles of the empirical distribution. Thus a perceived limitation of our simulation is that we are not evaluating BPSA and PSA as they would be applied in practice. However, another perspective is that $c_1, c_2, c_3, c_4$ form part of the specification for the model for $P(Y = 1|x, c)$. If we were to estimate $c_1, c_2, c_3, c_4$ when applying BPSA and PSA to the synthetic datasets, then the interpretation of the parameters $(\xi_1, \xi_2, \xi_3, \xi_4, \xi_5)$ would change from one dataset to the next. Averaging point estimates over repeated simulations would not be meaningful. Instead, we study BPSA and PSA under repeated application to data where $c_1, c_2, c_3, c_4$ take on specific values.

## 5.2 Simulation study when the distribution of the outcome follows a conventional regression model for $Y$ on $X$ and $C$

A limitation of the BPSA method is that it relies on more modelling assumptions than PSA, in some sense. Both methods use the same models for the data, but they handle information in different ways. PSA estimates the propensity scores using the marginal model for $P(X = 1|c)$, whereas BPSA estimates propensity scores using models for both $P(Y = 1|x, c)$ and $P(X = 1|c)$. Consequently, we might expect that the PSA methodology is more robust. If the model for $P(Y = 1|x, c)$ does not follow equation (3.1), then this would adversely affect BPSA because the method would be classifying study units into propensity score bins based on an incorrect model. The improved performance of BPSA that is observed in simulations may be sensitive to modelling assumptions. To investigate this further, we repeat the simulations of Section 5.1 by generating synthetic datasets using a more conventional regression model for $Y$ on $X$ and $C$.

*Simulation design*

We consider the case where $C$ has four continuous components (thus a $5 \times 1$ vector with first component is equal to one for the y-intercept), and we simulate datasets where the outcome variable $Y$ follows the regression model

$$\text{logit}\{P(Y = 1|x, c)\} = \tau x + c'\rho, \tag{5.1}$$

rather than the propensity score model given in equation (3.1). Equation (5.1) is a logistic regression model of $Y$ on $X$ and $C$ with treatment effect $\tau$ and covariate effects $\rho$.

We consider four different simulation designs with model parameters given by

| Design | $\tau$ | $\rho$ | $\gamma$ |
|---|---|---|---|
| #1 | -1/2 | (1/2, -1/2, 1/2, -1/2, 1/2) | (1/2, -1/2, 1/2, -1/2, 1/2) |
| #2 | -1/2 | (1, -1, 1, -1, 1) | (1/2, -1/2, 1/2, -1/2, 1/2) |
| #3 | -1/2 | (1/2, -1/2, 1/2, -1/2, 1/2) | (1, -1, 1, -1, 1) |
| #4 | -1/2 | (1, -1, 1, -1, 1) | (1, -1, 1, -1, 1) |

and datasets of fixed sample size $n = 1000$. The simulation designs parallel those of Section 5.1. Designs #1 models the case where the components of $C$ are modestly associated with $X$ and $Y$, while Designs #2, #3 and #4 model stronger associations. Thus we consider instances where $C$ are strong or weak confounders for the effect of $X$ on $Y$.

For each design, we generate and analyze 400 synthetic datasets using the following algorithm:

1. Generate the $n \times 5$ design matrix $\mathbf{c}$. The first column is a column of ones. The latter four columns are the sampled covariates for the dataset of size $n$. Each element of each column is simulated as an independent draw from a N(0,1) random variable.

2. Generate the $n \times 1$ vector $\mathbf{x}$ using the logistic regression model of equation (3.2), where $\gamma = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$ is a $5 \times 1$ vector.

3. Generate the $n \times 1$ vector $\mathbf{y}$ using the logistic regression model given in equation (5.1).

4. Because of the way the simulation is designed, we have $\gamma' C \sim N(\gamma_0, \sum_{i=1}^4 \gamma_i^2)$ for fixed $\gamma$. Thus the values $c_1, c_2, c_3, c_4$ defining the true quintiles of the propensity score are given exactly by $c_k = \text{expit}\{\gamma_0 + (\sum_{i=1}^4 \gamma_i^2) q_k\}$ for $k = 1, 2, 3, 4$ and $q_k = \Phi^{-1}(0.2k)$, where $\Phi^{-1}(.)$ is the quantile function of a N(0,1) random variable. Given $c_1, c_2, c_3, c_4$, analyze the datasets using BPSA and PSA to obtain point and 80% interval estimates for $\beta, \xi$ and $\gamma$ from each method.

A feature of this simulation is that we are generating $Y$, $X$ and $C$ using one model but then analyzing the data using a different model. BPSA and PSA give limiting estimates which will differ from $\tau$, $\rho$ and $\gamma$ used to generate the data. In large samples, PSA yields estimates of the quantity $\gamma^*$ which solves

$$E\left\{\frac{\partial}{\partial\gamma} \log p(X_i|C_i)\right\} = 0, \tag{5.2}$$

where $p(X_i|C_i) = \exp\{X_i(\gamma^T C_i)\}/(1 + \exp\{\gamma^T C_i)\}$ from equation (3.2), and the expectation is with respect to the distribution of $X_i$ given $C_i$ used to generate the data. PSA also estimates $(\beta^*, \xi^*)$ which solves

$$E\left\{\frac{\partial}{\partial(\beta, \xi)} \log p(Y_i|X_i, C_i)\Big|_{\gamma=\gamma^*}\right\} = 0,$$

where $p(Y_i|X_i, C_i) = \exp\{Y_i(\beta X_i + \xi^T g(C_i, \gamma^*))\}/(1 + \exp\{\beta X_i + \xi^T g(C_i, \gamma)\})$ from equation (3.1), with $\gamma^*$ as the solution to equation (5.2), and where the expectation is with respect to the distribution of $Y_i$ given $X_i$ and $C_i$ in equation (5.1). The quantity $\gamma^*$ is identical to the true value of $\gamma$ used to generated the synthetic data because PSA estimates the propensity scores using the correct model specification for $p(X_i|C_i)$. But $\beta^*$ may differ from $\tau$ in equation (5.1) because the quantities parametrize different regression models for the outcome.

In contrast, BPSA estimates $(\beta^*, \xi^*, \gamma^*)$ which solves

$$E\left[\frac{\partial}{\partial(\beta, \xi, \gamma)}\Big\{ \log p(Y_i|X_i, C_i) + \log p(X_i|C_i)\Big\}\right] = 0.$$

The quantities $(\beta^*, \xi^*, \gamma^*)$ give the best overall fit for the BPSA model, and they need not equal the limiting estimates obtained from PSA. In particular, $\gamma^*$ from BPSA need not equal $\gamma$ used to generate the synthetic data. Thus in this particular simulation, when the outcome variable follows equation (5.1), BPSA may give propensity score estimates which are asymptotically biased.

This raises questions about identifying suitable targets of inference in simulation. We focus on the quantities estimated from PSA. The reason is because these parameters model the best fit for the model given in equation (3.1) in the case where the true propensity scores are know. Thus, for example, $\xi^*$ model the average outcome risks within the propensity score bins. To compute these quantities, we write the estimating equation for PSA as

$$\Omega(\theta^*) = E\big\{R[Y - \text{expit}(R'\theta^*)]\big\} = \mathbf{0}, \tag{5.3}$$

where $R$ is a $6 \times 1$ vector with first component equal to $X$ and remaining five components equal to the $5 \times 1$ vector $g(z(\gamma, C))$. Additionally, $\theta^* = [\beta^*, \xi_1^*, \xi_2^*, \xi_3^*, \xi_4^*, \xi_5^*]$. The expectation is with respect to the true joint probability density function for $Y, X$ and $C$. The solution to equation (5.3) yields the density $P(Y = 1|x, c) = \text{expit}\{\beta^* x + g(z(\gamma, C)'\xi^*)\}$ with the smallest Kullback-Liebler distance from the density $\text{expit}\{\tau x + c'\rho)\}$, where $\tau$ and $\rho$ are equal to the values from the specific simulation design [38].

Because $R$ is categorical, we can calculate the solution to this estimating equation numerically using Monte Carlo integration and the Newton Raphson method. Equation (5.3) is the expectation of the score function for a single observation of logistic regression. It can be written as a finite sum of quantities which can be estimated by Monte Carlo, and which depend on $\theta^*$. If we take the derivative of this sum with respect to $\theta^*$, then we can use Newton Raphson to solve $\Omega(\theta^*) = 0$. Complete details are given in the Appendix.

As discussed in Section 2.4, simulation studies of PSA in which a dichotomous outcome variable $Y$ is generated using models like in equation (5.1) are common in the literature (see for example [28, 29, 34, 35, 43]). But investigators often treat the quantity $\hat{\beta}^*$, calculated from PSA, as a point estimator for $\tau$ rather than $\beta^*$ [34, 35]. But Austin and others [28–30] show that in general we have $\tau \neq \beta^*$ because of non-collapsibility of the odds ratio. Thus $\hat{\beta}^*$ will be asymptotically biased for $\tau$. The alternative approach that we adopt here to study the operating characteristics of $\hat{\beta}^*$ and $\hat{\xi}^*$ as estimates of $\beta^*$ and $\xi^*$ rather than $\tau$ and $\rho$.

### *Results*

Table 5.5 summarizes the performance of point and interval estimates for $\beta^*$, $\xi^*$ and $\gamma$ from PSA, in the case where datasets are simulated according to Design #1 with sample size $n = 1000$. For PSA, we have $\gamma^* = \gamma$. The left most column gives the true values of $\beta^*$ and $\xi^*$ which are estimated using the Monte Carlo algorithm given in the Appendix. The uncertainty in these quantities is negligible with standard errors less than $10^{-5}$. As we discuss in more detail below, the PSA point estimates computed from analyzing the synthetic data appear to be essentially unbiased. This gives us some reassurance that the Monte Carlo algorithm has been implemented correctly. As in Section 5.1, the second and third columns in Table 5.5 provide descriptive information about the magnitude of bias for each method, while the fourth and fifth columns in the table contain the estimated relative efficiencies and relative MSE of BPSA point estimators compared to PSA. Tables 5.6, 5.7 and 5.8 are identical to Table 5.5, but correspond to data simulated under Designs #2, #3 and #4.

With respect to the treatment effect parameter $\beta^*$, both BPSA and PSA perform comparably well. BPSA point estimates appear to be slightly less efficient with larger overall MSE, although the difference in the estimated variances is not statistically significant. BPSA interval estimates are generally longer on average, but this does not appear to greatly impact coverage.

The results of the simulation study indicate that point estimates for $\xi^*$ calculated from BPSA have inferior performance compared to PSA. They are less efficient and badly biased. In each of the four tables we see that the relative efficiencies are significantly greater than one, indicating that the variance of BPSA estimates of $\xi^*$ are greater than for PSA. The BPSA point estimates are biased with z-score values in the range of 5 to 10 or more. This increases mean squared error. BPSA interval estimates of $\xi^*$ also perform worse compared to PSA. In each of the tables, they have lower coverage probability and greater average length.

Table 5.5: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #1.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean<br>(z-score) | PSA Sample mean<br>(z-score) | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | | | | | Coverage | Length | Coverage | Length |
| $\beta^* = -0.44$ | -0.46 (-3.1) | -0.44 (0.2) | 1.02 | 1.04$^*$ | 0.78 | 0.40 | 0.78 | 0.40 |
| $\xi_1^* = -0.87$ | -0.97 (-8.3) | -0.87 (0.7) | 1.33$^*$ | 1.56$^*$ | 0.76$^\dagger$ | 0.50 | 0.76$^\dagger$ | 0.42 |
| $\xi_2^* = 0.81$ | 0.92 (8.8) | 0.83 (1.1) | 1.15$^*$ | 1.37$^*$ | 0.77 | 0.66 | 0.80 | 0.56 |
| $\xi_3^* = 1.33$ | 1.43 (7.3) | 1.33 (-0.7) | 1.19$^*$ | 1.35$^*$ | 0.78 | 0.66 | 0.78 | 0.58 |
| $\xi_4^* = 1.87$ | 1.98 (7.3) | 1.83 (-2.8) | 1.27$^*$ | 1.41$^*$ | 0.79 | 0.73 | 0.77 | 0.60 |
| $\xi_5^* = 2.66$ | 2.93 (14.4) | 2.65 (-0.7) | 1.76$^*$ | 2.67$^*$ | 0.72$^\dagger$ | 0.85 | 0.78 | 0.67 |
| $\gamma_0 = 0.5$ | 0.50 (0.1) | 0.50 (0.2) | 1.37$^*$ | 1.37$^*$ | 0.60$^\dagger$ | 0.16 | 0.77 | 0.19 |
| $\gamma_1 = -0.5$ | -0.49 (3.9) | -0.51 (-3.1) | 0.85$^*$ | 0.86$^*$ | 0.67$^\dagger$ | 0.15 | 0.82 | 0.19 |
| $\gamma_2 = 0.5$ | 0.49 (-4.0) | 0.50 (1.2) | 0.83$^*$ | 0.86$^*$ | 0.66$^\dagger$ | 0.15 | 0.80 | 0.19 |
| $\gamma_3 = -0.5$ | -0.49 (4.5) | -0.51 (-2.0) | 0.75$^*$ | 0.78$^*$ | 0.66$^\dagger$ | 0.14 | 0.80 | 0.19 |
| $\gamma_4 = 0.5$ | 0.48 (-4.9) | 0.51 (2.1) | 0.74$^*$ | 0.77$^*$ | 0.68$^\dagger$ | 0.14 | 0.79 | 0.19 |

$^*$ Quantity differs from 1, $p < 0.1$, $^\dagger$ Coverage probability is less than 80%, $p < 0.1$

Table 5.6: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #2.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean | PSA Sample mean | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | (z-score) | (z-score) | | | Coverage | Length | Coverage | Length |
| $\beta^* = -0.40$ | -0.43 (-3.0) | -0.39 (1.1) | 1.04* | 1.06* | 0.84 | 0.46 | 0.82 | 0.45 |
| $\xi_1^* = -1.60$ | -1.78 (-12.8) | -1.58 (1.3) | 1.33* | 1.87* | $0.74^\dagger$ | 0.69 | $0.70^\dagger$ | 0.52 |
| $\xi_2^* = 1.50$ | 1.61 (9.1) | 1.48 (-1.6) | 1.12 | 1.35* | 0.82 | 0.77 | 0.80 | 0.64 |
| $\xi_3^* = 2.52$ | 2.68 (10.3) | 2.48 (-3.1) | 1.21* | 1.50* | 0.78 | 0.82 | $0.76^\dagger$ | 0.67 |
| $\xi_4^* = 3.57$ | 3.80 (11.5) | 3.51 (-3.1) | 1.26* | 1.63* | $0.74^\dagger$ | 1.00 | $0.68^\dagger$ | 0.76 |
| $\xi_5^* = 5.03$ | 5.59 (16.9) | 5.08 (1.1) | 0.49 | 0.85 | $0.70^\dagger$ | 1.51 | $0.74^\dagger$ | 1.09 |
| $\gamma_0 = 0.5$ | 0.53 (5.8) | 0.51 (2.0) | 1.64* | 1.76* | $0.44^\dagger$ | 0.13 | 0.80 | 0.19 |
| $\gamma_1 = -0.5$ | -0.48 (5.3) | -0.50 (-0.5) | 0.65* | 0.70* | $0.56^\dagger$ | 0.11 | 0.82 | 0.19 |
| $\gamma_2 = 0.5$ | 0.49 (-3.8) | 0.50 (-0.6) | 0.78* | 0.80* | $0.57^\dagger$ | 0.11 | 0.82 | 0.19 |
| $\gamma_3 = -0.5$ | -0.49 (4.7) | -0.51 (-1.5) | 0.69* | 0.72* | $0.59^\dagger$ | 0.11 | 0.82 | 0.19 |
| $\gamma_4 = 0.5$ | 0.48 (-5) | 0.50 (-0.2) | 0.63* | 0.67* | $0.55^\dagger$ | 0.11 | 0.81 | 0.19 |

* Quantity differs from 1, $p < 0.1$, $^\dagger$ Coverage probability is less than 80%, $p < 0.1$

Table 5.7: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #3.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean | PSA Sample mean | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | (z-score) | (z-score) | | | Coverage | Length | Coverage | Length |
| $\beta^* = -0.41$ | -0.43 (-1.8) | -0.40 (1.1) | 0.99 | 1.00 | 0.80 | 0.46 | 0.79 | 0.46 |
| $\xi_1^* = -0.87$ | -0.93 (-6.3) | -0.88 (-0.3) | $1.18^*$ | $1.30^*$ | 0.80 | 0.46 | 0.78 | 0.41 |
| $\xi_2^* = 0.80$ | 0.85 (4.9) | 0.79 (-0.7) | $1.09^*$ | $1.16^*$ | 0.85 | 0.63 | 0.80 | 0.57 |
| $\xi_3^* = 1.31$ | 1.38 (5.5) | 1.30 (-0.7) | 1.04 | $1.11^*$ | 0.80 | 0.67 | 0.77 | 0.61 |
| $\xi_4^* = 1.85$ | 1.90 (3.7) | 1.81 (-2.8) | 1.03 | 1.04 | 0.82 | 0.74 | 0.78 | 0.66 |
| $\xi_5^* = 2.63$ | 2.81 (10.7) | 2.61 (-0.8) | $1.22^*$ | $1.56^*$ | $0.74^\dagger$ | 0.83 | 0.78 | 0.72 |
| $\gamma_0 = 1$ | 1.00 (-0.1) | 1.01 (3.0) | $1.32^*$ | $1.29^*$ | $0.69^\dagger$ | 0.23 | 0.78 | 0.24 |
| $\gamma_1 = -1$ | -0.98 (4.7) | -1.01 (-2.5) | $1.19^*$ | $1.24^*$ | $0.65^\dagger$ | 0.23 | 0.82 | 0.26 |
| $\gamma_2 = 1$ | 0.98 (-4.3) | 1.01 (1.8) | $1.28^*$ | $1.33^*$ | $0.69^\dagger$ | 0.23 | 0.86 | 0.26 |
| $\gamma_3 = -1$ | -0.98 (3.3) | -1.01 (-1.6) | $1.16^*$ | $1.18^*$ | $0.67^\dagger$ | 0.23 | 0.79 | 0.25 |
| $\gamma_4 = 1$ | 0.99 (-2.2) | 1.01 (2.7) | 1.00 | 0.99 | $0.70^\dagger$ | 0.23 | 0.77 | 0.26 |

$^*$ Quantity differs from 1, $p < 0.1$, $^\dagger$ Coverage probability is less than 80%, $p < 0.1$

Table 5.8: Performance of point and interval estimators from BPSA and PSA, when data are simulated under Design #4.

| Parameter | Point Estimation | | | | Interval Estimation | | | |
|---|---|---|---|---|---|---|---|---|
| | BPSA Sample mean | PSA Sample mean | Rel. efficiency | Rel. MSE | BPSA | | PSA | |
| | (z-score) | (z-score) | | | Coverage | Length | Coverage | Length |
| $\beta^* = -0.35$ | -0.38 (-3.5) | -0.35 (-0.3) | 1.00 | 1.03 | 0.80 | 0.51 | $0.76^{\dagger}$ | 0.50 |
| $\xi_1^* = -1.60$ | -1.70 (-8.6) | -1.61 (-0.4) | 1.09 | $1.29^*$ | 0.78 | 0.58 | $0.73^{\dagger}$ | 0.50 |
| $\xi_2^* = 1.47$ | 1.56 (6.9) | 1.49 (1.7) | $1.09^*$ | $1.21^*$ | 0.79 | 0.72 | 0.78 | 0.65 |
| $\xi_3^* = 2.48$ | 2.59 (7.1) | 2.48 (0.3) | $1.11^*$ | $1.24^*$ | $0.76^{\dagger}$ | 0.79 | 0.77 | 0.70 |
| $\xi_4^* = 3.52$ | 3.65 (6.6) | 3.49 (-2.1) | 1.09 | $1.19^*$ | 0.78 | 0.94 | $0.76^{\dagger}$ | 0.80 |
| $\xi_5^* = 4.97$ | 5.33 (12.7) | 5.00 (1.3) | $1.16^*$ | $1.63^*$ | $0.73^{\dagger}$ | 1.35 | $0.70^{\dagger}$ | 1.11 |
| $\gamma_0 = 1$ | 1.03 (4.9) | 1.01 (2.5) | $2.10^*$ | $2.20^*$ | $0.47^{\dagger}$ | 0.20 | 0.80 | 0.24 |
| $\gamma_1 = -1$ | -1.00 (0.1) | -1.01 (-3.0) | $1.15^*$ | $1.13^*$ | $0.59^{\dagger}$ | 0.19 | 0.80 | 0.26 |
| $\gamma_2 = 1$ | 1.00 (-0.8) | 1.01 (2.0) | 1.01 | 1.00 | $0.60^{\dagger}$ | 0.19 | 0.78 | 0.26 |
| $\gamma_3 = -1$ | -0.99 (1.9) | -1.01 (-1.0) | $1.15^*$ | $1.16^*$ | $0.58^{\dagger}$ | 0.19 | 0.81 | 0.25 |
| $\gamma_4 = 1$ | 0.99 (-1.2) | 1.00 (1.2) | 0.98 | 0.98 | $0.57^{\dagger}$ | 0.19 | $0.76^{\dagger}$ | 0.26 |

$^*$ Quantity differs from 1, $p < 0.1$, $^{\dagger}$ Coverage probability is less than 80%, $p < 0.1$

The simulations indicate that BPSA point and interval estimates of $\gamma$ occasionally have better performance compared to PSA. When the true values of the components of $\gamma$ are small in magnitude (Designs #1 and #2), we see an improvement in estimation of $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ due to better efficiency. The relative efficiencies are significantly less than one. Whereas when the components of $\gamma$ are large, the efficiency of BPSA and PSA point estimates for $\gamma$ are essentially the same. Across all the simulation designs, BPSA point estimation of the quantity $\gamma_0$ is always poor. BPSA interval estimates for $\gamma$ perform much worse than PSA. They have smaller average length across all four simulation designs, but this is not accompanied by proper coverage probability. Thus BPSA appears to produce interval estimates which are falsely precise. In contrast, PSA interval estimates always have correct coverage levels.

### *Discussion*

The simulations indicate that when the model for $Y$ obeys equation (5.1), many of the advantages of BPSA over PSA disappear. Section 5.1 indicated that BPSA may yield improved point and interval estimation of $\xi^*$ and $\gamma$. Across each of the simulation designs in Tables 5.1 through 5.4, BPSA point estimates of $\gamma$ were more efficient compared to PSA and had smaller MSE. This led to improved classification of study units into propensity score bins and better estimation of $\xi^*$. When the model for $P(Y = 1|x, c)$ follows the model of equation (5.1), we see nearly the reverse results. The estimates of the components of $\gamma$ are sometimes more efficient compared to PSA, but are generally highly biased. BPSA does not appear to give good estimates of the propensity scores, consequently, we also see poor estimation of $\xi^*$.

Inspection of Tables 5.5 through 5.8 reveals that PSA interval estimates of $\xi^*$ do not have nominal coverage probability. This would be expected even if $\gamma$ were known because the model for $P(Y = 1|x,c)$ is incorrect. Standard practice for maximum likelihood estimation under misspecified models uses a robust "sandwich" estimate of the estimator variance [38]. PSA uses no such approach to interval estimation, and confidence intervals should not be expected to have the correct coverage levels. This reasoning does not apply to PSA interval estimation of $\gamma$ because the marginal model for $P(X = 1|C)$ is correctly specified. Comparing the results from Section 5.1 and 5.2, we see that when the model for $P(Y = 1|x,c)$ is correctly specified, PSA yields interval estimates for $\xi^*$ which are too narrow, presumably because PSA ignores uncertainty in $\gamma$. In this case, Section 5.1 shows that using BPSA to model uncertainty in $\gamma$ yields interval estimates of $\xi^*$ with correct coverage levels. In contrast, if the model for $P(Y = 1|x,c)$ is incorrect, then neither method can be expected to give interval estimates with proper coverage.

One feature of Tables 5.5 through 5.8 is that BPSA estimates of the quantity $\xi_5^*$ seems to be particularly bad. Under each of the four simulation designs, the z-score for the BPSA sample mean is between 15 and 20. This is a likely consequence of the simulation design involving sparse data within the fifth propensity score bin. In this case, PSA seems to produce point and interval estimates with better performance.

As was the case in Section 5.1, BPSA and PSA perform comparably well in point and interval estimation of the treatment effect $\beta^*$. PSA point estimates are slightly more efficient across the four simulation designs. For interval estimation, we see a small increase in the average length for BPSA, and this is accompanied by an increase in coverage probability. But the differences between the two methods are sufficiently modest to be swamped by simulation error.

## 5.3 Predictive performance in real and simulated data

Simulations studies have the disadvantage that they only describe estimator performance where the data are generated under certain specific circumstances. These may not be representative of real epidemiologic investigations such as the statin data example. Consequently, our ability to generalize about the performance of BPSA is somewhat limited. One strategy for characterizing performance for the statin data is to investigate prediction error using cross validation. The original dataset involved 4572 patients discharged from hospital between 1999-2001. However, data from an additional 4599 patients discharged the following year are also available for a total of 9171 observations. If we randomly split the the entire collection of data in half, we can use the first half (called the *build data*) to construct a predictive model to estimate the probability of death for a future patient. We can then study prediction error when applied to the other half of the data (called the *test data*). To average over variability in the choice of build data, we can replicate the random splitting of the data.

Denote a random sample of data of size $n = 4500$ as $(\mathbf{y}, \mathbf{x}, \mathbf{c})$. Let $(Y^*, X^*, C^*)$ denote the data for a future patient from the same population for whom only $X^*$ and $C^*$ are observed. Let $\hat{\beta}, \hat{\xi}$ and $\hat{\gamma}$ denote the point estimates for model parameters from PSA applied to $(\mathbf{y}, \mathbf{x}, \mathbf{c})$, and define

$$\hat{Y}_{PSA} = \text{expit}\{\hat{\beta}X^* + g(z(C^*, \hat{\gamma}))'\hat{\xi}\}$$

as the predictive model from PSA which estimates $P(Y^* = 1|X^*, C^*)$. The posterior distribution for $(\beta, \xi, \gamma)$ from BPSA of (**y**, **x**, **c**) is $P(\beta, \xi, \gamma | \mathbf{y}, \mathbf{x}, \mathbf{c})$. Define

$$\hat{Y}_{BPSA} = \iiint \text{expit}\{\beta X^* + g(z(C^*, \gamma))'\xi\} P(\beta, \xi, \gamma | \mathbf{y}, \mathbf{x}, \mathbf{c}) d\beta d\gamma d\xi.$$

as the predictive model from BPSA for estimation of $P(Y^* = 1 | X^*, C^*)$. The quantity $\hat{Y}_{PSA}$ is a prediction based on substitution of $(\hat{\beta}, \hat{\xi}, \hat{\gamma})$ into the propensity score model for the outcome, while $\hat{Y}_{BPSA}$ is the posterior predictive distribution for $Y^*$. The estimate $\hat{Y}_{BPSA}$ acknowledges uncertainty in the bin group membership of the future patient, while $\hat{Y}_{PSA}$ does not.

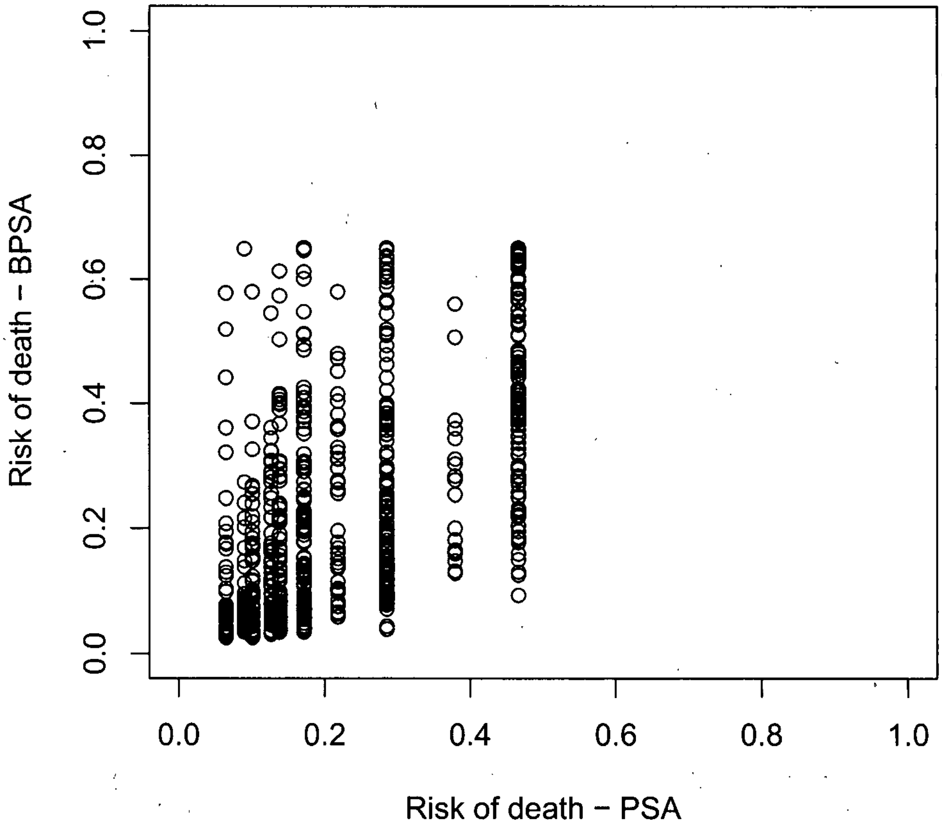
To give a sense of the extent that predictions from BPSA and PSA may differ, Figure 5.1 plots $\hat{Y}_{PSA}$ versus $\hat{Y}_{BPSA}$ for the 4572 patients from the original dataset of Chapter 4. In other words, we analyze the data using BPSA and PSA, and we plot $\hat{Y}_{PSA}$ versus $\hat{Y}_{BPSA}$ based on predictions within the same data. In Figure 5.1, we see that there is disagreement between predictions. The estimates $\hat{Y}_{PSA}$ take on at most ten different values because patients can be classified into only one of ten different treatment-bin combinations. BPSA averages over uncertainty in $\gamma$, or rather, uncertainty in bin membership. We see that the distribution of $\hat{Y}_{BPSA}$ given $\hat{Y}_{PSA}$ is highly variable.

We quantify prediction error using the loss function

$$L(\hat{Y}, Y^*) = -[Y^* \log(\hat{Y}) + (1 - Y^*) \log(1 - \hat{Y})],$$

where $\hat{Y}$ is a prediction. This is called the predictive log score [5, 44] because, for a sample of data $Y_1, \ldots, Y_n$ and predictions $\hat{Y}_1, \ldots, \hat{Y}_n$, the quantity $-\sum_{i=1}^{n} L(\hat{Y}_i, Y_i^*)$ is the log-likelihood for the data calculated from the predictive model. We define prediction error as the expected loss $E[L(\hat{Y}, Y^*)]$, where the expectation is with re-

Figure 5.1: Estimated risks of death in statin data for BPSA compared to PSA. ($\hat{Y}_{BPSA}$ versus $\hat{Y}_{PSA}$)



spect to variability in both in $\hat{Y}$ and $Y^*$. Randomness in $\hat{Y}$ arises from the random sampling of both the build data ($\mathbf{y}$, $\mathbf{x}$, $\mathbf{c}$) and $X^*, C^*$.

Because of the way the loss function is defined, small prediction errors are desirable. Predictions with small error have the appealing property that they assign high probability to the observed data, irrespective of how those data are generated. When comparing predictive models, the model with the smaller prediction error has the smaller Kullback-Leibler distance from the true distribution which generated the data. For a dichotomous variable $Y$ with density $f(y)$ and a predictive model $\hat{f}(y)$, this distance is given by

$$E\left[\log\left\{\frac{f(Y)}{\hat{f}(Y)}\right\}\right] = E[\log(f(Y))] - E[\log(\hat{f}(Y))].$$

The first term does not depend on the choice of predictive model, while the second term is equal to the prediction error. This is because

$$\log\{\hat{f}(Y)\} = Y\log\{\hat{f}(1)\} + (1-Y)\log\{1-\hat{f}(0)\}.$$

Hence minimizing prediction error gives a predictive model which approximates the true distribution of the data.

We can estimate both $E[L(\hat{Y}_{BPSA}, Y^*)]$ and $E[L(\hat{Y}_{PSA}, Y^*)]$ for the statin data using a variation of 5-fold cross-validation. From the total sample of 9171 patients, we randomly select $n = 4500$ patients without replacement. We analyze the dataset to obtain predictive models, and we then evaluate these predictions using the remaining

4671 patients by calculating

$$\begin{aligned}\hat{\phi}_{BPSA} &= \frac{1}{4671}\sum_{i=1}^{4671} L(\hat{Y}_{BPSA,\,i}, Y_i^*) \\ &= -\frac{1}{4671}\sum_{i=1}^{4671}[Y_i^* \log(\hat{Y}_{BPSA,\,i}) + (1-Y_i^*)\log(1-\hat{Y}_{BPSA,\,i})]\end{aligned}$$

and

$$\begin{aligned}\hat{\phi}_{PSA} &= \frac{1}{4671}\sum_{i=1}^{4671} L(\hat{Y}_{PSA,\,i}, Y_i^*) \\ &= -\frac{1}{4671}\sum_{i=1}^{4671}[Y_i^* \log(\hat{Y}_{PSA,\,i}) + (1-Y_i^*)\log(1-\hat{Y}_{PSA,\,i})],\end{aligned}$$

where the index $i$ is over observations in the test data. The quantities $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ are unbiased estimates of $E[L(\hat{Y}_{BPSA}, Y^*)|\mathbf{y},\ \mathbf{x},\ \mathbf{c}]$ and $E[L(\hat{Y}_{PSA}, Y^*)|\mathbf{y},\ \mathbf{x},\ \mathbf{c})]$ respectively, where the conditioning is with respect to the build data. Thus $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ provide "half the story" of predictive performance in the sense that they quantify prediction error for specific build data. To fully characterize $E[L(\hat{Y}_{BPSA}, Y^*)]$ we can repeatedly split the data and examine the sequence of $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ over the random splittings.

We do not consider traditional 5-fold cross-validation where the data are split into five parts. In the analysis of the statin data, we found that the computational cost of analyzing four fifths of the 9171 observations for the statin data was high for BPSA. The large sample size combined with a fairly inefficient MCMC algorithm, which requires individual updating of the components of $\gamma$ one at a time over many measured covariates, reduced the efficiency of the algorithm. Instead our approach was to assess predictive performance using the "random splitting" approach described here.

Table 5.9: The quantities $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ from five different random splittings of the statin data into build data ($n = 4500$) and test data ($n = 4671$).

| | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ |
|---|---|---|
| | 0.40 | 0.47 |
| | 0.38 | 0.45 |
| | 0.39 | 0.46 |
| | 0.39 | 0.45 |
| | 0.39 | 0.46 |
| Mean | 0.39 | 0.46 |

Table 5.9 presents $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ for five random splittings of the data. Each row in the table corresponds to $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ for one random splitting. Standard errors for the estimates are calculated as the sample standard deviation of the replicates of $L(\hat{Y}, Y^*)$ over subjects in the test data, and they are less than 0.008. For each row in the table, $\hat{\phi}_{BPSA}$ is significantly smaller than $\hat{\phi}_{PSA}$. Thus the prediction estimates $\hat{Y}_{BPSA}$ calculated by applying BPSA to the build data have smaller prediction error compared to the corresponding estimates calculated from using PSA. The improvement in prediction persists across random splittings of the data.

To shed some insight into the results, we can attempt to confirm these findings using simulations. Because of the flexibility of simulation studies, we need not worry about correlation in estimated prediction error because of repeated re-analysis of the same data. Instead we can simulate a sequence of build datasets, analyze them to obtain predictive models, and then estimate prediction error using massive simulated test datasets. Furthermore, in simulations the quantity $\hat{Y}_{TRUE} = P(Y^* = 1|X^*, C^*) = \text{expit}(\beta X^* + g(z(C^*, \gamma)'\xi))$ is known exactly because we know $\beta$, $\xi$ and $\gamma$. We can calculate the quantity

$$\begin{aligned}\hat{\phi}_{TRUE} &= \frac{1}{N}\sum_{i=1}^{N} L(\hat{Y}_{TRUE,\,i}, Y_i^*) \\ &= -\frac{1}{N}\sum_{i=1}^{N}[Y_i^* \log(\hat{Y}_{TRUE,\,i}) + (1 - Y_i^*)\log(1 - \hat{Y}_{TRUE,\,i})],\end{aligned}$$

where $N$ is the size of a simulated test dataset. The quantity $\hat{\phi}_{TRUE}$ estimates $E[L(P(Y^* = 1), Y^*)|\mathbf{y}, \mathbf{x}, \mathbf{c})]$ which is the smallest possible prediction error because the true model has Kullback-Leibler distance of zero from itself. Thus calculating $\hat{\phi}_{TRUE}$ for each build dataset gives a benchmark of the best predictions and quantifies the extent that BPSA improves upon PSA.

Accordingly, Table 5.10 presents estimates of $\hat{\phi}_{TRUE}$, $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ for simulated build datasets of sample size $n = 1000$ under the four simulation designs described in Section 5.1. In other words, we generate ten synthetic datasets using the propensity score model given in equations (3.1) and (3.2). For each design, a row in the table corresponds to one simulated build dataset. We analyze the simulated build dataset using PSA and BPSA, and then study the predictive performances by calculating the quantities $\hat{\phi}_{BPSA}$ and $\hat{\phi}_{PSA}$ using large test datasets ($n = 40000$). Standard errors for the quantities are less than 0.001 for Designs #1 and #3, and 0.003 for Designs #2 and #4..

In Table 5.10, $\hat{\phi}_{BPSA}$ is always less than or equal to $\hat{\phi}_{PSA}$, irrespective of the simulation design or build dataset used to obtain the predictive models. As expected, $\hat{\phi}_{TRUE}$ is less than $\hat{\phi}_{BPSA}$ because BPSA cannot have smaller prediction error than the true model. To give an indication of performance across the ten simulated build datasets, the bottom row of each table reports the averages across the ten build datasets.

Table 5.10: The quantities $\hat{\phi}_{BPSA}$, $\hat{\phi}_{TRUE}$ and $\hat{\phi}_{PSA}$ from ten simulated datasets generated under Designs #1, #2, #3 or #4.

| | Design #1 | | | Design #3 | | |
|---|---|---|---|---|---|---|
| | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ |
| | 0.66 | 0.67 | 0.68 | 0.67 | 0.67 | 0.67 |
| | 0.66 | 0.67 | 0.68 | 0.67 | 0.67 | 0.68 |
| | 0.66 | 0.67 | 0.68 | 0.66 | 0.67 | 0.68 |
| | 0.66 | 0.67 | 0.68 | 0.66 | 0.67 | 0.68 |
| | 0.66 | 0.67 | 0.68 | 0.67 | 0.67 | 0.68 |
| | 0.66 | 0.67 | 0.68 | 0.67 | 0.67 | 0.68 |
| | 0.66 | 0.67 | 0.68 | 0.67 | 0.67 | 0.67 |
| | 0.66 | 0.67 | 0.67 | 0.66 | 0.68 | 0.68 |
| | 0.66 | 0.67 | 0.67 | 0.67 | 0.67 | 0.68 |
| | 0.66 | 0.67 | 0.68 | 0.66 | 0.67 | 0.68 |
| Mean | 0.66 | 0.67 | 0.68 | 0.67 | 0.67 | 0.68 |

| | Design #2 | | | Design #4 | | |
|---|---|---|---|---|---|---|
| | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ |
| | 0.40 | 0.40 | 0.50 | 0.40 | 0.40 | 0.42 |
| | 0.40 | 0.42 | 0.47 | 0.40 | 0.41 | 0.47 |
| | 0.40 | 0.40 | 0.47 | 0.39 | 0.40 | 0.47 |
| | 0.40 | 0.41 | 0.51 | 0.40 | 0.41 | 0.49 |
| | 0.40 | 0.41 | 0.49 | 0.39 | 0.40 | 0.43 |
| | 0.41 | 0.41 | 0.52 | 0.40 | 0.40 | 0.43 |
| | 0.40 | 0.40 | 0.49 | 0.40 | 0.40 | 0.46 |
| | 0.40 | 0.40 | 0.52 | 0.40 | 0.40 | 0.43 |
| | 0.39 | 0.41 | 0.45 | 0.40 | 0.40 | 0.45 |
| | 0.39 | 0.40 | 0.52 | 0.40 | 0.40 | 0.45 |
| Mean | 0.40 | 0.41 | 0.49 | 0.40 | 0.40 | 0.46 |

From Table 5.10 we see that BPSA predictions strongly outperform PSA under Designs #2 and #4. For Designs #1 and #3 we see more modest improvements. This fits well with the findings of Section 5.1, where we studied the performance of point and interval estimation using the same simulation designs. Design #2 corresponds to the case where the components of $\xi$ are large in magnitude and heterogeneous while $\gamma$ is small. Thus in model fitting the outcome variable supplies a lot of information about the propensity score. PSA is disadvantaged because it does not use this information. In Table 5.2, PSA point estimators of $\xi$ are badly biased because error in propensity score estimation causes study units to be misclassified into propensity score bins. Poor estimation of $\xi$ and $\beta$ should be expected to yield unreliable predictions for $Y$ given $X$ and $C$. For Designs #1 and #3 we see smaller improvement in the quality of predictions for BPSA versus PSA. For these designs, BPSA point estimators for $\xi$ have only modestly improved mean squared error compared to PSA, so we would not expect large improvements in predictive performance.

To study the sensitivity of these findings across different specifications of the outcome model, we study predictive performance in synthetic data generated according to the regression model of Section 5.2. Table 5.11 presents the quantities $\hat{\phi}_{BPSA}$, $\hat{\phi}_{TRUE}$ and $\hat{\phi}_{PSA}$ from datasets generated according to Designs #1 through #4 from Section 5.3. The layout of the table is identical to that of Table 5.10. Standard errors are less than 0.002 for Designs #1 and #3, and 0.003 for Designs #2 and #4. For each of the simulated datasets, we see a tendency for $\hat{\phi}_{TRUE} < \hat{\phi}_{BPSA} < \hat{\phi}_{PSA}$, indicating that the model from BPSA yields better predictions than PSA. The differences $\hat{\phi}_{BPSA} - \hat{\phi}_{PSA}$ are fairly modest compared to the case where the model for the outcome is correctly specified.

The simulations show that the improvement in predictive performance for BPSA relative to PSA is sensitive to correct model specification for the outcome. Comparing

Table 5.11: The quantities $\hat{\phi}_{BPSA}$, $\hat{\phi}_{TRUE}$ and $\hat{\phi}_{PSA}$ from ten simulated datasets generated under Designs #1, #2, #3 or #4 using the conventional regression model of equation (5.1).

| | Design #1 | | | Design #3 | | |
|---|---|---|---|---|---|---|
| | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ |
| | 0.61 | 0.62 | 0.62 | 0.61 | 0.62 | 0.63 |
| | 0.61 | 0.62 | 0.62 | 0.62 | 0.63 | 0.63 |
| | 0.60 | 0.61 | 0.62 | 0.61 | 0.62 | 0.62 |
| | 0.61 | 0.62 | 0.63 | 0.62 | 0.62 | 0.63 |
| | 0.61 | 0.62 | 0.62 | 0.62 | 0.63 | 0.63 |
| | 0.61 | 0.61 | 0.62 | 0.62 | 0.62 | 0.63 |
| | 0.61 | 0.61 | 0.62 | 0.62 | 0.62 | 0.63 |
| | 0.61 | 0.62 | 0.62 | 0.61 | 0.63 | 0.63 |
| | 0.61 | 0.62 | 0.62 | 0.62 | 0.63 | 0.63 |
| | 0.60 | 0.61 | 0.62 | 0.61 | 0.62 | 0.63 |
| Mean | 0.61 | 0.62 | 0.62 | 0.62 | 0.62 | 0.63 |

| | Design #2 | | | Design #4 | | |
|---|---|---|---|---|---|---|
| | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ | $\hat{\phi}_{TRUE}$ | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ |
| | 0.46 | 0.47 | 0.48 | 0.47 | 0.48 | 0.48 |
| | 0.46 | 0.47 | 0.49 | 0.47 | 0.48 | 0.49 |
| | 0.46 | 0.48 | 0.48 | 0.47 | 0.48 | 0.49 |
| | 0.46 | 0.47 | 0.48 | 0.47 | 0.48 | 0.49 |
| | 0.46 | 0.47 | 0.48 | 0.47 | 0.48 | 0.48 |
| | 0.46 | 0.48 | 0.48 | 0.46 | 0.48 | 0.48 |
| | 0.46 | 0.47 | 0.48 | 0.47 | 0.48 | 0.49 |
| | 0.46 | 0.48 | 0.49 | 0.47 | 0.49 | 0.49 |
| | 0.46 | 0.48 | 0.48 | 0.47 | 0.48 | 0.49 |
| | 0.46 | 0.47 | 0.48 | 0.47 | 0.49 | 0.50 |
| Mean | 0.46 | 0.47 | 0.48 | 0.47 | 0.48 | 0.49 |

Design #2 in Tables 5.10 and 5.11, we see sensitivity in predictive performance. In the bottom rows of the tables, we report the mean predictive performance when averaged across the build datasets. For Table 5.10, comparing BPSA to PSA, we have a mean difference of 0.49 - 0.41 = 0.08, whereas in Table 5.11 we have a mean difference of only 0.48 - 0.47 = 0.01. Thus correct model specification is important. In fact, when we look at prediction in the statin data, the improvement in performance for BPSA is better than would be expected from the simulation findings.

## 5.4 An investigation of covariate balance produced by BPSA versus PSA

In Chapter 4, we applied BPSA to the statin data and compared the results to those obtained with PSA. One of the findings was that PSA produced treatment and control groups with similar distributions of measured confounders, whereas BPSA produced treatment and control groups that are not particularly similar. PSA appeared to do a better job of reducing confounding bias.

To illustrate, recall Tables 4.4 and 4.6. Such tables are commonly used to assess the adequacy of models for the propensity score. They give summary statistics for the distribution of the components of C, among treatment and control groups, within bins of the estimated propensity scores. As discussed in Section 2.3, stratifying on the true propensity score causes $C$ to be identically distributed across treatment groups. The reason is because the distribution of $X$ given the propensity score does not depend on $C$.

Table 4.4 was generated using propensity scores estimates from PSA, and indicates that the method effectively reduces much of the confounding due to $C$. Within each of the bins, the components of $C$ are roughly evenly distributed in treatment and control. For example, while age is a strong confounding variable (see Table 1.1), the strength of the association between age and statin treatment is largely reduced within each of the bins. Table 4.6 presents identical summary statistics, based on propensity scores estimated from BPSA. Within most of the propensity score bins, there are numerous covariates with distributions that differ in treatment versus control. Thus BPSA produces worse covariate balance compared to PSA, and does not appear to effectively reduce confounding.

These findings are surprising in light of the results of Sections 5.1 and 5.2. In simulations, point and interval estimates for the propensity scores calculated from BPSA perform very favorably compared to PSA when the model for $P(Y = 1|x, c)$ is correctly specified. The point estimates have smaller variance and MSE. Interval estimates have shorter length while retaining roughly nominal coverage probability. Based on these results, we might expect that stratifying on propensity scores estimated from BPSA should reduce covariate imbalances in treatment versus control and do a better job of reducing confounding compared to PSA.

In this chapter we explore these apparently contradictory findings by examining the lack of covariate balance that arises in using BPSA for control of confounding. We analyze synthetic data which closely approximate the statin data, with the following two objectives:

1. To replicate the covariate imbalance that is observed when applying BPSA to the statin data.

2. To determine if poor covariate balance from BPSA may occur simultaneously with better estimation of the propensity scores.

This investigation should shed insight into whether or not it is possible for improved estimation of the propensity scores to result in treatment and control groups which are less similar in terms of the distribution of measured confounders.

### *Simulation design*

Rather than sampling numerous datasets and studying the operating characteristics of our methods across datasets, we focus on the analysis of a pair of datasets which closely approximate the statin data. This should facilitate drawing a connection between previous simulations and the data analysis of Chapter 4. Moreover, Section 5.1 demonstrates that the improvement of point and interval estimation of $\gamma$ from using BPSA is sufficiently large that it should be detectable without repeated sampling of datasets.

The two datasets of sample size $n = 5000$, which we denote as **Dataset A** and **Dataset B**, are generated using the following algorithm:

1. Generate $\mathbf{c}$ as a $5000 \times 21$ design matrix consisting of a column of ones and twenty columns of covariates drawn independently from $N(0,1)$.

2. Generate $\mathbf{y}$ and $\mathbf{x}$ as $5000 \times 1$ response and treatment vectors for the $n$ subjects, generated from either:

   - **Dataset A:** The propensity score model given in equations (3.1) and (3.2), with parameters

$$\begin{aligned} \beta &= -0.3, \\ \xi' &= (1, 0, -1, -2, -3), \\ \gamma' &= (-1, -0.1, 0.1, -0.1, 0.1, \ldots, -0.1, 0.1). \end{aligned}$$

- **Dataset B:** The regression models given in equations (5.1) and (3.2), with parameters

$$\begin{aligned}\tau &= -0.3;\\ \rho &= (-2, -0.2, 0.2, \ldots, -0.2, 0.2),\\ \gamma' &= (-1, -0.1, 0.1, -0.1, 0.1, \ldots, -0.1, 0.1).\end{aligned}$$

The choice of design including sample size, number of covariates and parameter values is guided by the analysis results for the statin data. The values of $\beta$ and $\xi$ are guided by Tables 4.3 and 4.5. The choice of $\tau$ and $\rho$ is guided by Table 4.1, and the choice of $\gamma$ is guided by Table 4.2. We generate the outcome under different models in order reflect the fact that the true model for $P(Y = 1|x, c)$ in the statin dataset is unknown.

*Results*

To illustrate rough similarity between Datasets A and B and the statin data, Figure 5.2 gives plots of the density $f(z|X = x)$ for $x = 0, 1$ and logit$[P(Y = 1|x, z)]$ for $x = 0, 1$ and $z \in [0, 1]$ for the pair of datasets. The plots are exactly comparable to Figures 4.1 and 4.2, and are produced in the same manner using the true propensity scores. Dashed and solid lines correspond to the untreated group and treated group, respectively. The plots in Figure 5.2 are fairly similar to those in Figures 4.1 and 4.2. The risk of $Y$ within treatment groups is roughly decreasing in $Z$. This mimics the notion that healthy subjects are the most likely to receive treatment. Further, the untreated group has a lower distribution of propensity scores than the treated group. Thus Datasets A and B roughly approximate the statin data in the sense that they have similar sample size, number of covariates, and dependencies between $X$, $Y$ and $C$.

We apply BPSA and PSA to each of the datasets. To investigate the distribution of $C$ in treatment versus control when stratifying on competing propensity score estimates, we use a descriptive technique given by Imai and Van Dyk [45]. As was discussed in Section 2.3, Rosenbaum and Rubin [1] showed that $X \perp\!\!\!\perp C|Z$ because the distribution of $X$ given the propensity score is equal to $P(X = 1|C, z) = z$ which does not depend on $C$. This implies that $X \perp\!\!\!\perp C_j|Z$ for $j = 1, \ldots, 20$. Thus the following models are correct:

$$\text{logit}[P(X = 1|C_j, Z)] = \phi_j + \theta_j C_j + \omega_j \text{logit}[Z] \text{ for } j = 1, \ldots, 20, \tag{5.4}$$

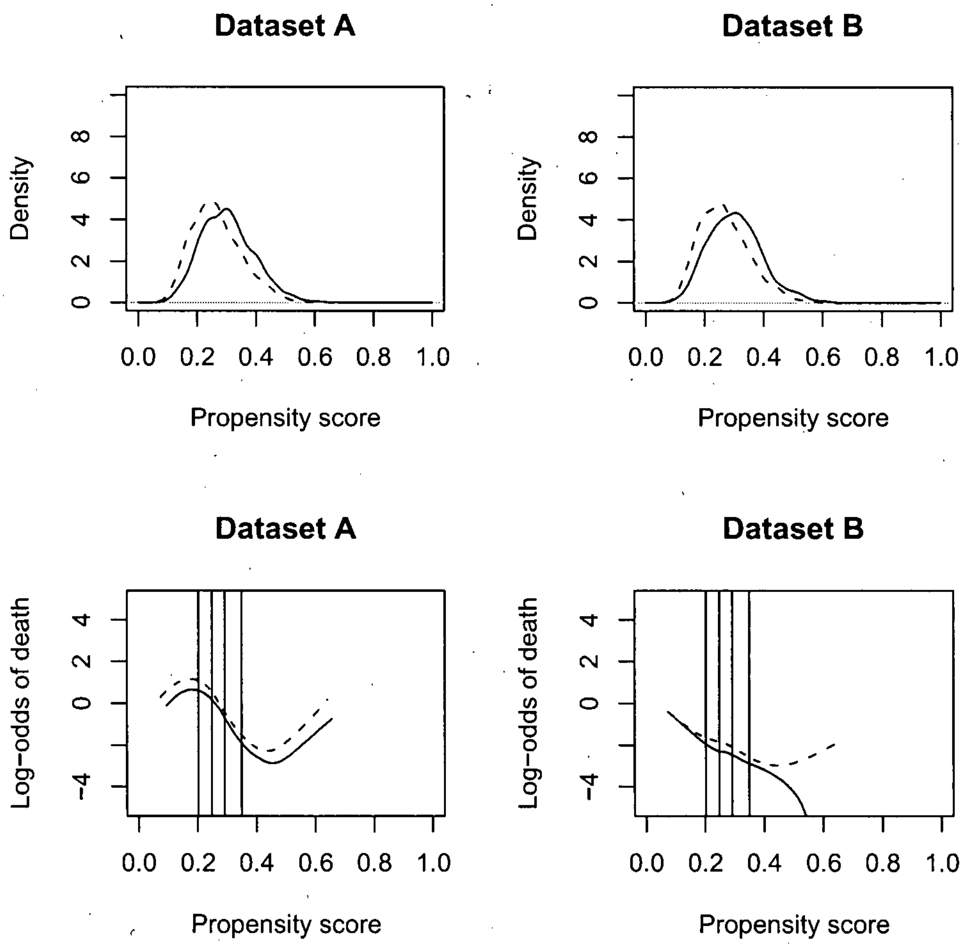
where $\phi_j = \theta_j = 0$ and $\omega_j = 1$ for $j = 1, \ldots, 20$.

We can fit these regression models by substituting the propensity scores estimates $\hat{Z}_{BPSA}$ or $\hat{Z}_{PSA}$ for $Z$, where

$$\begin{aligned} \hat{Z}_{PSA} &= \text{expit}(\hat{\gamma}'C) \\ \hat{Z}_{BPSA} &= \int \text{expit}(\gamma'C) f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c}) d\gamma, \end{aligned}$$

and $f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c})$ is the posterior distribution for $\gamma$. The point estimates for $\theta_1, \theta_2, \ldots, \theta_{20}$, can be used as diagnostic tools to assess the similarity of the distribution of $C$ in treatment versus control. If we fit the models using the true propensity scores $Z = \text{expit}(\gamma'C)$, then Imai and Van Dyk [45] point out that the resulting z-statistics for point estimates of $\theta_1, \theta_2, \ldots, \theta_{20}$ will be normally distributed with mean zero and variance equal to one. This is because the true values of $\theta_1, \theta_2, \ldots, \theta_{20}$ are equal to zero since $X \perp\!\!\!\perp C_j|Z$. To elaborate, if we calculate the maximum likeli-

Figure 5.2: The density $f(z|X = x)$ for $x = 0, 1$ and $\text{logit}[P(Y = 1|x, z)]$ for $x = 0, 1$ and $z \in [0, 1]$ for datasets A and B. Solid curves correspond to treated patients and dashed curves correspond to untreated patients.



hood estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_{20}$, and we divide them by their estimated standard errors $\hat{SE}_1, \hat{SE}_2, \ldots, \hat{SE}_{20}$, then asymptotically we will have

$$\frac{\hat{\theta}_1}{\hat{SE}_1}, \frac{\hat{\theta}_2}{\hat{SE}_2}, \ldots, \frac{\hat{\theta}_{20}}{\hat{SE}_{20}} \sim N(0,1).$$

Conditional on $Z$, any association between $X$ and $C_j$ is due to chance. We can study the distribution of the z-statistics calculated from competing estimates of the propensity scores, while using the $N(0,1)$ density as a "benchmark" for comparison.

Accordingly, we fit the regression models given in equation (5.4) for Datasets A or B, and competing propensity scores estimates $Z$, $\hat{Z}_{PSA}$ and $\hat{Z}_{BPSA}$. In other words, we fit a total of $2 \times 20 \times 3 = 120$ regression models, where each regression has a specific dataset (Dataset A or B), covariate $C_j$ $(j = 1, 2, 3 \ldots 20)$, and vector of propensity scores ($Z$, $\hat{Z}_{PSA}$ or $\hat{Z}_{BPSA}$). For each regression, we compute the z-statistic $\frac{\hat{\theta}_j}{\hat{SE}_j}$. We then produce normal quantile plots of $\frac{\hat{\theta}_1}{\hat{SE}_1}, \frac{\hat{\theta}_2}{\hat{SE}_2}, \ldots, \frac{\hat{\theta}_{20}}{\hat{SE}_{20}}$ for Dataset A or B and for $Z$ or $\hat{Z}_{PSA}$ or $\hat{Z}_{BPSA}$.

The results are given in Figure 5.3. The ($\circ$) symbols and ($\triangle$) symbols correspond to z-statistics computed from regressing on $\hat{Z}_{PSA}$ or $\hat{Z}_{BPSA}$, respectively. The (+) symbols correspond to z-statistics computed from regressing on the true propensity score Z.

Figure 5.3: Normal quantile plots for the z-statistics $\frac{\hat{\theta}_1}{\hat{SE}_1}, \frac{\hat{\theta}_2}{\hat{SE}_2}, \ldots, \frac{\hat{\theta}_{20}}{\hat{SE}_{20}}$ in either Dataset A or B. The symbols ○, △ and + correspond to regressing on $\hat{Z}_{PSA}, \hat{Z}_{BPSA}$ or $Z$ respectively.

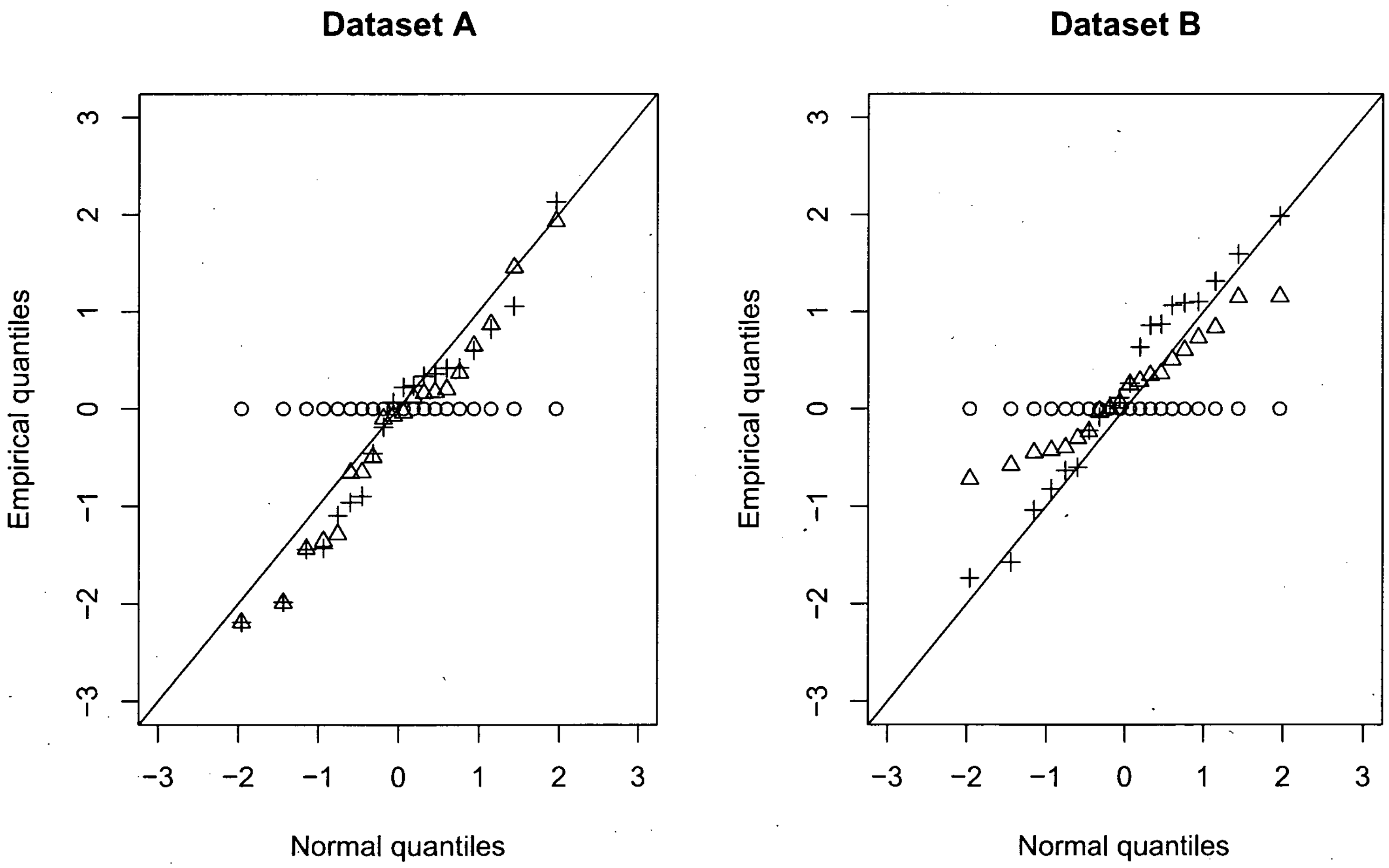


Figure 5.3 illustrates that adjustment for $\hat{Z}_{PSA}$ rather than $Z$ produces treatment and control groups that are more similar with respect to $C$. The circle ($\circ$) symbols lie right along the horizontal axes for both Datasets A and B. The quantities $\frac{\hat{\theta}_1}{\hat{SE}_1}, \frac{\hat{\theta}_2}{\hat{SE}_2}, \ldots, \frac{\hat{\theta}_{20}}{\hat{SE}_{20}}$ have small variance compared to a sample of size 20 from a standard normal distribution. In contrast, when we adjust for $Z$ we obtain the (+) symbols which, as expected, lie along the diagonal line indicating rough agreement with a standard normal density. Thus in Datasets A and B, stratifying on propensity scores estimated from PSA produces better covariate balance than stratifying on the true propensity scores. The empirical distributions of the confounders in treatment and control are more similar than would be expected if treatment were assigned at random. This does not mean that adjusting for $\hat{Z}_{PSA}$ is a more effective strategy for reducing confounding bias in either dataset. Given $Z$, the distribution of $C$ is identical in treatment and control, and we cannot have confounding. But PSA appears to reduce differences in the empirical distributions of $C$ that arise by chance.

The balancing properties of stratifying on true versus estimated propensity scores are investigated by Rosenbaum and Rubin [3] and by Rubin and Thomas [12, 13]. Rubin and Thomas [12, 13] study matched sampling in observational studies when matching on $Z$ versus $\hat{Z}_{PSA}$. They derive analytic approximations of the variance of the difference of the sample means in the matching variables for treatment versus control in the case where $C$ is multivariate normal. They demonstrate that matching on $\hat{Z}_{PSA}$ can reduce the variance of the difference of the sample means by a factor of one half compared to matching on $Z$.

Figure 5.3 illustrates that regression adjustment for $\hat{Z}_{BPSA}$ produces greater differences in the distribution of $C$ in treatment versus control compared to adjustment for $\hat{Z}_{PSA}$. The ($\triangle$) symbols are much more variable than the ($\circ$) symbols. However, adjustment for $\hat{Z}_{BPSA}$ does not appear to be any worse than adjustment for $Z$. In

Datasets A and B, the variation of the ($\triangle$) symbols is not greater than that of the (+) symbols.

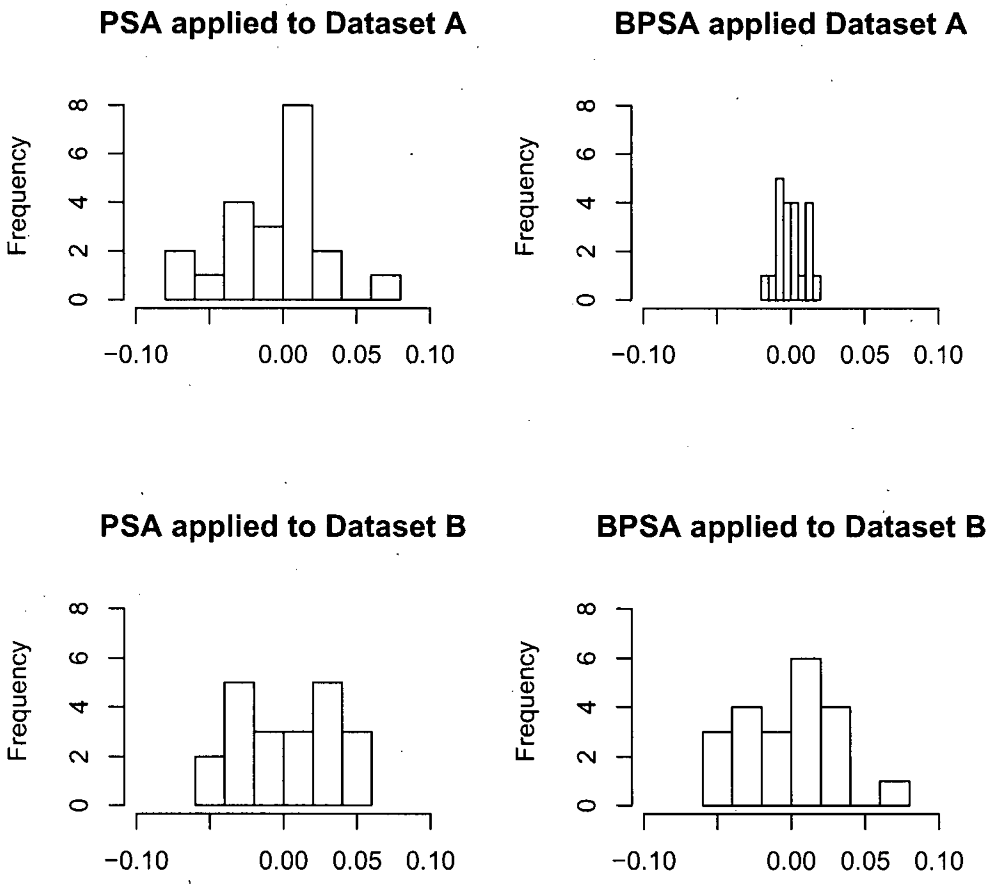
We now investigate the quality of the propensity score estimates from applying BPSA or PSA to the data. Recall that the parameter $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{20})$ models the propensity scores because it indexes the regression model

$$\begin{aligned} \text{logit}[P(X = 1|c)] &= \gamma' c \\ &= \gamma_0 + \gamma_1 c_1 + \gamma_2 c_2 + \ldots + \gamma_{20} c_{20}. \end{aligned}$$

Thus in Datasets A and B, the true propensity score for a subject with covariate vector $C$ is equal to $Z = \text{expit}(\gamma_0 + \gamma_1 C_1 + \gamma_2 C_2 + \ldots + \gamma_{20} C_{20})$. Figure 5.4 plots histograms of the quantities $(\hat{\gamma}_k - \gamma_k)$ for $k = 0, 1, \ldots, 20$ where the $\hat{\gamma}_k$ are point estimates obtained from BPSA or PSA applied to either Datasets A or B. For example, the top left figure has the heading "PSA applied to Dataset A". This means that we apply PSA to Dataset A, obtain point estimates of $\gamma_1, \gamma_2, \ldots, \gamma_{20}$ denoted $\hat{\gamma}_1, \hat{\gamma}_2, \ldots, \hat{\gamma}_{20}$, calculate the quantities $(\hat{\gamma}_k - \gamma_k)$ for $k = 0, 1, \ldots, 20$, and make a histogram of the result. If the histograms have small variance and are centered at zero, then this indicates that the quantities $\hat{\gamma}_k$ are high quality estimates of $\gamma_k$. The histogram does not give us information about bias or efficiency because they are generated from the analysis of a single dataset.

As expected, Figure 5.4 illustrates that in the Dataset A, BPSA provides better estimates of $\gamma$ compared to PSA. The histogram is visibly concentrated near zero because the point estimates of the components of $\gamma$ have small MSE compared to PSA. This fits well with the simulation results of Section 5.1 which indicate that BPSA performs favorably when the model for $P(Y = 1|x, c)$ is correct. In Dataset B, the model for $P(Y = 1|x, c)$ is misspecified, and Figure 5.4 indicates that BPSA point

Figure 5.4: Histograms of $(\hat{\gamma}_k - \gamma_k)$ for $k = 0, 1, \ldots, 20$ based on either BPSA of PSA analyses of either Dataset A or B.



estimates do not appear to have any clear advantage over PSA. The variances of the histograms are roughly the same for both methods. This is also expected because the simulations in Section 5.2 illustrate that misspecification of $P(Y = 1|x, c)$ adversely affects the performance of estimates of $\gamma$.

We can also study the quality of the propensity score estimates by estimating prediction error for treatment. Let $(X^*, C^*)$ denote the data for a study unit drawn from the density function $f(x^*|c^*)f(c^*)$, which is the same for Datasets A and B. We can study the performance of competing estimates for $P(X^* = 1|C^*)$. Two estimates are the quantities

$$\begin{aligned} \hat{Z}^*_{PSA} &= \text{expit}(\hat{\gamma}'C^*) \\ \hat{Z}^*_{BPSA} &= \int \text{expit}(\gamma'C^*)f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c})d\gamma, \end{aligned}$$

which are the estimated propensity scores for $(X^*, C^*)$ calculated from the dataset used to build the predictive model for treatment. We study performance using the quantities

$$\begin{aligned} \hat{\phi}_{BPSA} &= \frac{1}{N}\sum_{i=1}^{N} L(\hat{Z}^*_{BPSA,\, i}, X^*_i) \\ \hat{\phi}_{PSA} &= \frac{1}{N}\sum_{i=1}^{N} L(\hat{Z}^*_{PSA,\, i}, X^*_i) \\ \hat{\phi}_{TRUE} &= \frac{1}{N}\sum_{i=1}^{N} L(Z_i, X^*_i), \end{aligned}$$

where $L(Z, X^*)$ is the predictive log score defined in Section 5.3, and indexing over $i$ denotes observations in a test dataset of sample size $N$.

Table 5.12 gives the values of $\hat{\phi}_{BPSA}$, $\hat{\phi}_{PSA}$ and $\hat{\phi}_{TRUE}$ calculated from the predictive models obtained by analyzing Datasets A and B. For Dataset A, the quantity

Table 5.12: The quantities $\hat{\phi}_{BPSA}, \hat{\phi}_{PSA}$ and $\hat{\phi}_{TRUE}$ based on predictive models for treatment calculated from Datasets A or B.

| Dataset | $\hat{\phi}_{BPSA}$ | $\hat{\phi}_{PSA}$ | $\hat{\phi}_{TRUE}$ |
|---|---|---|---|
| A | 0.572 | 0.574 | 0.572 |
| B | 0.574 | 0.574 | 0.572 |

SE's < 0.001

$\hat{\phi}_{BPSA}$ is less than $\hat{\phi}_{PSA}$. This implies that propensity scores estimates calculated from BPSA predict treatment assignment of future observations with smaller error than PSA. This is consistent with the results of Figure 5.3, and also the simulation results. In contrast, for Dataset B we see no improvement in prediction of treatment for BPSA compared to PSA.

### *Discussion*

The motivation for this section was to understand the seemingly contradictory results of Chapter 4 and 5. The simulations of Section 5.1 and 5.2, indicate that BPSA propensity score estimates perform favorably compared to PSA. We would expect that adjustment for propensity scores estimated from BPSA should be an effective approach to reducing confounding bias. But in Chapter 4, we applied BPSA the statin data and observed that the treatment and control groups differed systematically with respect to important outcome risk factors.

The results of this investigation demonstrate that better estimation of the propensity scores does not imply better covariate balance between treatment versus control groups. When we fit the regression models in equation (5.4), adjustment for $\hat{Z}_{PSA}$ rather than $Z$ produces a greater level of similarity in the distribution of $C$ in treatment versus control. Conditional on $Z$, $X$ and $C$ are independent, but the empirical distributions of $C$ given $X$ will differ from one dataset to the next. In contrast, stratifying on $\hat{Z}_{PSA}$ reduces differences in the empirical distributions that emerge by chance. A detailed discussion of this characteristic of PSA is given by Rubin and Thomas [12, 13].

Intuitively, BPSA may yield results which are a compromise between regression adjustment for $\hat{Z}_{PSA}$ versus adjustment for $Z$. The quantities $\hat{Z}_{BPSA}$ are the better estimates of $Z$. But in Datasets A and B, we appear to pay a price in terms of worse comparability in the distribution of $C$ in treatment versus control. This may shed some light into the apparently poor performance of BPSA when applied to the statin data. The imbalances in Table 4.6 may be a consequence of better estimation of the propensity scores. To fully address this question, it would be useful to study the quality of the propensity score estimates in the statin data, perhaps using the cross-validation approach outlined above. Furthermore, the results of this section may shed light into why BPSA point estimates of $\beta$ are less efficient in the simulations of Section 5.1 and 5.2 than those of PSA.

# Chapter 6

# Conclusion

## 6.1 Summary

We have proposed an approach for combining propensity score methods with Bayesian inference for control of confounding bias in observational studies with a dichotomous outcome, dichotomous treatment, and measured confounders. The method models the propensity score as a latent variable and uses Bayes theorem to integrate the latent variable out of the posterior distribution for model parameters. This estimation strategy is common in data analysis applications with missing data or latent structure. We model the joint distribution for the data, parameters and latent quantity, and we then study the marginal posterior distribution for model parameters. For BPSA, the posterior distribution for the treatment effect incorporates uncertainty about the propensity scores for each subject. In simulations and in the analysis of the statin data, we demonstrate that BPSA yields interval estimates for the treatment effect parameter which are longer on average compared to PSA. This is a consequence of propagating uncertainty in the propensity scores through the analysis.

Intuition says that BPSA and PSA should give similar answers. The methods use the same models. Asymptotically, uncertainty in the propensity scores should be small. Nonetheless, PSA is a two step procedure. It involves first estimating propensity scores and then including the estimates as a covariate in a regression model for the outcome. BPSA does both steps simultaneously. The method exploits prior information about the relationship between the outcome and propensity score within treatment groups. The MCMC computational scheme involves iteratively updating the propensity scores and then fitting a complete data step from the results. During the updating, the algorithm is likely to yield propensity score estimates which cluster patients into groupings based on the outcome risk. The conditional distribution for the $\gamma$ parameter, given the data and $(\beta, \xi)$, contains a contribution from the model for the outcome variable. As we learn about $\xi$ and $\beta$, information flows back through the algorithm to affect estimation of $\gamma$. The outcome variable is ignored by PSA when estimating the propensity scores. To put it another way, PSA estimates $\gamma$ from the marginal density $f(x|c)$ whereas BPSA uses the joint model for $f(y, x|c)$. The methods use the same models, but handle information in different ways.

We demonstrate BPSA in an observational study of the effectiveness of statin therapy in Ontario patients discharged alive from hospital following acute myocardial infarction. Austin and Mamdani previously used this dataset to conduct a detailed case-study of propensity score methods [2]. We apply BPSA and compare the results to PSA. While treatment effect estimates are similar, we see large differences in point and interval estimates of $\xi$. Further examination reveals that the differences can be attributed to differences in the characteristics of patients classified to propensity score bins for either method. Patients with high propensity scores are healthier in general because physicians are known to prescribe statins to patients who are young with fewer comorbidities [19, 20]. When we apply BPSA, study subjects in the upper bins with high propensity score have low prevalences of mortality risk factors. We see the reverse effect in lower propensity score bins. BPSA aggregates subjects into bins depending on how sick they are, whereas PSA only considers the relationship between the treatment variable and confounders.

In the Monte Carlo simulations of Chapter 5, we demonstrate that BPSA may yield more efficient estimates of $\gamma$ and therefore the propensity scores. When the model for the outcome variable follows equation (3.1), BPSA point estimates are more efficient relative to PSA and have lower mean squared error. Interval estimates for $\gamma$ have shorter average length and retain roughly nominal coverage probability. Point and interval estimates of $\xi$ calculated from BPSA also appear to have better performance. Improved estimation of $\gamma$ may allow more accurate classification of study subjects into propensity score bins, improving estimation of $\xi$.

While BPSA and PSA use the same models, BPSA makes stronger assumptions in some sense. We might expect BPSA to be less robust to modelling assumptions for for $P(Y = 1|x, c)$. We study this issue in Section 5.2. Synthetic data are simulated for the case when the distribution of the outcome variable follows a conventional regression model of $Y$ on $X$ and $C$ given in equation (5.1). We show that in this case the performance advantage of BPSA breaks down. Point and interval estimates of $\gamma$ and $\xi$ have fairly severe bias, and this increases MSE and harms interval estimation.

Furthermore, stratifying on propensity score bins estimated from BPSA appears to produce treatment and control groups which are not particularly comparable. For the PSA methodology, it is standard practice to assess the quality of the propensity score model by looking for systematic differences in the covariate distributions for treatment versus control within each of the bins. In the statin data, PSA eliminates much of the confounding bias, whereas BPSA produces treatment and control groups which are not particularly comparable. This suggests that the BPSA method does not effectively reduce confounding in the statin data example. These findings are surprising in light of the simulations of Section 5.1 and 5.2. To explore this phenomenon further, Section 5.4 conducts a detailed analysis of two synthetic datasets which closely approximate the statin data in terms of sample size, covariates, and underlying parameter values. We show that BPSA may yield better estimation of the propensity scores which occurs simultaneously with greater dissimilarity in the covariate distributions for treatment and control. The poor comparability observed in Table 4.6 for the statin data may be a consequence of better estimation of the propensity scores.

## 6.2 Future Research

*Implementing BPSA using the MCEM algorithm.*

There is nothing inherently Bayesian about fitting regression models for $P(Y = 1|x, c)$ and $P(X = 1|c)$ simultaneously. In principle, it should be possible to use other methods of estimation, such as maximum likelihood. An advantage is that this would yield likelihood-based estimates which might have similar performance to estimates computed from BPSA. In this thesis, we consider the case where the likelihood for the data is given by

$$\begin{aligned} L(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c}) &= E\left[f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma)f(\mathbf{x}|\mathbf{c}, \gamma)\right] \\ &= \int f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma)f(\mathbf{x}|\mathbf{c}, \gamma)f(\gamma)d\gamma, \end{aligned}$$

The quantity $f(\gamma)$ is a prior distribution for $\gamma$. This expression states that the likelihood is equal to the likelihood given the propensity score, averaged over uncertainty in propensity score. Because the density $f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma)$ depends on $\gamma$ only via the linear predictor $g(C, \gamma)$, given in equation (3.1), this amounts to saying that $L(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c})$ is the average of the likelihoods when we account for uncertainty in the propensity score bin classification for the study units.

To maximize this likelihood with respect to $\beta$ and $\xi$, we can use the Monte Carlo Expectation Maximization (MCEM) algorithm [46]. The algorithm permits us to maximize $L(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c})$ while only working with the quantity $f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma)$, often called the complete data likelihood, and the conditional probability density function

$$f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi) = \frac{f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma) f(\mathbf{x}|\mathbf{c}, \gamma) f(\gamma)}{L(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c})}.$$

In motivating the MCEM algorithm, Wei and Tanner [47] note that for fixed $\beta^{(0)}$ and $\xi^{(0)}$, we can write

$$\log L(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c}) = E[\log f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma) f(\mathbf{x}|\mathbf{c}, \gamma)] - E[\log f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi)],$$

where the expectations are with respect to $f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c}, \beta^{(0)}, \xi^{(0)})$. To maximize $\log L(\beta, \xi|\mathbf{y}, \mathbf{x}, \mathbf{c})$, we need only consider the first term on the RHS of the above equation. We use the following algorithm:

1. Initialize $\beta^{(0)}$ and $\xi^{(0)}$.

2. For $t = 1, 2, \ldots,$

   - Expectation Step (E-step):
     Draw $\gamma^{(t,1)}, \gamma^{(t,2)}, \ldots, \gamma^{(t,M)}$ from $f(\gamma|\mathbf{y}, \mathbf{x}, \mathbf{c}, \beta^{(t-1)}, \xi^{(t-1)})$.

   - Maximization Step (M-step):
     Assign $(\beta^{t}, \xi^{t}) \leftarrow \arg\max_{\theta} \frac{1}{M} \sum_{i=1}^{M} \log f(\mathbf{y}|\mathbf{x}, \mathbf{c}, \beta, \xi, \gamma^{(t,i)})$.

The theoretical details are given by Wei and Tanner [47] who show that the sequence $(\beta^{(1)}, \xi^{(1)}), (\beta^{(2)}, \xi^{(2)}), \ldots$ converges to the maximum likelihood estimator.

The MCEM algorithm has close parallels to the Metropolis Hastings algorithm detailed in Section 3.2 for posterior simulation for BPSA. The E-step imputes the missing data, in this case, the propensity scores for study units which are modelled by the parameter $\gamma$. The M-step maximizes the resulting mixture of likelihood functions.

This is conceptually similar to sampling for the posterior distribution of $\beta$ and $\xi$ given the data and $\gamma$. When implementing the MCEM algorithm, we cannot draw from the conditional density $f(\gamma|\mathbf{y},\mathbf{x},\mathbf{c},\beta^{(t-1)},\xi^{(t-1)})$ directly because it is of unknown form. But we can sample from it using the Metropolis Hastings algorithm. Moreover, we have already identified suitable proposal distributions in Section 3.3.

We can compute standard errors for the resulting estimates of $\beta$ and $\xi$ using the approach of Oakes [48]. The variance of the maximum likelihood estimator for $(\beta,\xi)$ is approximated by $\frac{\partial^2}{\partial(\beta,\xi)^2}\log L(\beta,\xi|\mathbf{y},\mathbf{x},\mathbf{c})$, and we can write

$$\begin{aligned}\frac{\partial^2}{\partial(\beta,\xi)^2}\log L(\beta,\xi|\mathbf{y},\mathbf{x},\mathbf{c}) &= E\left[\frac{\partial^2}{\partial(\beta,\xi)^2}\log L(\beta,\xi|\mathbf{y},\mathbf{x},\mathbf{c},\gamma)\right]+\\ &\quad Var\left[\frac{\partial}{\partial(\beta,\xi)}\log L(\beta,\xi|\mathbf{y},\mathbf{x},\mathbf{c},\gamma)\right].\end{aligned}$$

This yields expression for computing standard errors.

$$\begin{aligned}\hat{E}\left[\frac{\partial^2}{\partial(\beta,\xi)^2}L(\beta,\xi|\mathbf{y},\mathbf{x},\mathbf{c},\gamma)\right] &= \frac{1}{M}\sum_{i=1}^{M}\frac{\partial^2}{\partial(\beta,\xi)^2}\log f(\mathbf{y}|\mathbf{x},\mathbf{c},\hat{\beta},\hat{\xi},\gamma^{(t,i)})\\ \hat{V}ar\left[\frac{\partial}{\partial(\beta,\xi)}\log L(\beta,\xi|\mathbf{y},\mathbf{x},\mathbf{c},\gamma)\right] &= \frac{1}{M}\sum_{i=1}^{M}\left[\frac{\partial}{\partial(\beta,\xi)}\log f(\mathbf{y}|\mathbf{x},\mathbf{c},\hat{\beta},\hat{\xi},\gamma^{(t,i)})\right.\\ &\quad\left.-\frac{1}{M}\sum_{i'=1}^{M}\frac{\partial}{\partial(\beta,\xi)}f(\mathbf{y}|\mathbf{x},\mathbf{c},\hat{\beta},\hat{\xi},\gamma^{(t,i')})\right]^2.\end{aligned}$$

The quantities $\hat{\beta}$ and $\hat{\xi}$ are the maximum likelihood estimators, and the first and second derivatives of $f(\mathbf{y}|\mathbf{x},\mathbf{c},\hat{\beta},\hat{\xi},\gamma)$ are obtained from the standard output from logistic regression of $\mathbf{y}$ on $\mathbf{x}$ and $\mathbf{z}$ as determined by the choice of $\gamma$. The expression for the large sample variance has a similar form to the variance decomposition given in equation (4.2) from Section 4.4.

Further study of the BPSA method could involve a detailed investigation of maximum likelihood estimators for $\beta$ and $\xi$ computed using the MCEM algorithm. We could compare such estimates to those from PSA in much the same manner as the investigations of Sections 5.1 and 5.2. Because we have already studied computational algorithms for BPSA, little work would be needed to implement the MCEM algorithm.

### *BPSA for hierarchically structured data*

The BPSA method has the advantage that it allows for the incorporation of standard Bayesian machinery in data analysis settings using propensity scores. We can use flexible modelling strategies, such as hierarchical models based on exchangeability assumptions, or incorporation of prior information from expert opinion or external validation data. BPSA also permits the use of Markov chain Monte Carlo methods for computing point and interval estimates. This gives inferences which do not depend on asymptotic approximations, as is the case for PSA.

One extension for BPSA would be to apply it in settings involving hierarchically structured data. Propensity score methods have generally been developed in settings involving cross-sectional data with individual-level covariates. However, epidemiological data often have a hierarchical structure. In the statin data example, prescribing practices might be driven by hospital-level covariates. For example, teaching hospitals may provide different quality services than hospitals in rural areas. Study subjects from the same hospital may have treatment levels which are correlated.

Suppose that $r$ is a $m \times 1$ vector of dummy variables indicating hospital membership for each study unit. We could model the propensity scores as

$$\text{logit}[P(X = 1|c, r)] = \gamma' c + \tau' r.$$

where the parameter $\tau$ models the hospital effects. Since the components of $\tau$ are likely to be similar, particularly if $m$ is large, we could model them hierarchically

$$\tau_1, \tau_2, \ldots, \tau_m \sim N(0, \sigma^2),$$

where $\sigma^2$ is a hyperparameter. If hospital level covariates are available, then they can be incorporated into modelling variability in $\tau$.

Sampling from the posterior distribution $f(\beta, \xi, \gamma, \tau | \mathbf{y}, \mathbf{x}, \mathbf{c})$ is an extension of the MCMC algorithm of Section 3.1. To fit the models

$$\begin{aligned} \text{logit}[P(Y = 1|x, c)] &= \beta x + g(z(c, \gamma))'\xi \\ \text{logit}[P(X = 1|c)] &= \gamma' c + \tau' r \\ \tau &\sim N(0, \sigma^2 I), \end{aligned}$$

with suitable prior distributions for $\beta$, $\xi$, $\gamma$ and $\tau$, we update sequentially from the conditional densities

$$\begin{gathered} f(\beta, \xi | \mathbf{y}, \mathbf{x}, \mathbf{c}, \gamma, \tau) \\ f(\gamma | \mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi, \tau) \\ f(\tau | \mathbf{y}, \mathbf{x}, \mathbf{c}, \beta, \xi, \gamma) \end{gathered}$$

The first conditional density does not depend on $\tau$ and is just the posterior step from the MCMC algorithm of Section 3.1. The second conditional distribution is identical to the imputation step of Section 3.1, but uses a $N(0, \tau^2)$ prior for $\gamma$ rather than a diffuse Gaussian prior. The final conditional distribution is conditionally conjugate under an inverse gamma prior for the hyperparameter $\sigma^2$.

In summary, BPSA seems to be a sensible alternative to PSA for reducing confounding in observational studies. While BPSA assumes more than PSA in estimating propensity scores, the method is merely exploiting the modelling assumptions that are built into PSA when selecting a linear predictor $g(.)$ for modelling the outcome variable. PSA has the advantage that inferences may be more robust to model misspecification. But in many applications, this estimation strategy may be overly pessimistic in the sense that prior information is available. Furthermore, PSA implicitly uses modelling assumptions for the outcome when computing parameter estimates and standard errors. Our investigation helps to shed light on validity and relevance of modelling assumptions that underlie propensity score methods.

# Bibliography

[1] P.R. Rosenbaum and D.B. Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–57, 1983.

[2] P.C. Austin and M.M. Mamdani. A comparison of propensity score methods: A case-study estimating the effectiveness of post-AMI statin use. *Stat Med*, 25:2084–106, 2005.

[3] P.R. Rosenbaum and D.B. Rubin. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc*, 79:516–24, 1984.

[4] J.K. Lunceford and M. Davidian. Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Stat Med*, 23:2937–60, 2004.

[5] J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky. Bayesian model averaging: A tutorial. *Statist Sci*, 14:382–417, 1999.

[6] K. Hirano and G.W. Imbens. The propensity score with continuous treatments. In. *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages Ed. A. Gelman and X. Meng, New York, Wiley. (2004), 71–84.

[7] D.B. Rubin. Estimating causal effects from large datasets using propensity scores. *Ann Intern Med*, 15:757–63, 1997.

[8] J. Hahn. On the role of the propensity score for efficient estimation of the average treatment effects. *Econometrika*, 66:315–31, 1998.

[9] J.M. Robins and Newey W.K. Estimation of exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48:479–95, 1992.

[10] P.R. Rosenbaum. Model-based direct adjustment. *J Am Stat Assoc*, 82:387–94, 1987.

[11] K. Hirano, G.W. Imbens, and G. Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrika*, 71:1161–89, 2003.

[12] D.B. Rubin and N. Thomas. Matching using estimated propensity scores: Relating theory to practice. *Biometrics*, 52(1):249–264, 1996.

[13] D.B. Rubin and N. Thomas. Characterizing the effect of matching using linear propensity score methods with normal distributions. *Biometrika*, 79:797–809, 1992.

[14] D.B. Rubin. The use of propensity scores in applied Bayesian inference. *Bayesian Statistics*, 2:463–72, 1985.

[15] J.M. Robins and Y. Ritov. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Stat Med*, 16:285–318, 1997.

[16] J.C. LaRosa, J. He, and S. Vupputuri. Effect of statins on risk of coronary disease: A meta-analysis of randomized controlled trials. *J Am Med Assoc*, 282:2340–2346, 1999.

[17] H.D. Aronow, E.J. Topol, M.T. Roe, and et al. Effect of lipid-lowering therapy on early mortality after acute coronary syndromes: An observational study. *Lancet*, 357:1063–68, 2001.

[18] U. Stenestrand and L. Wallentin. Early statin treatment following acute myocardial infarction and 1-year survival. *J Am Med Assoc*, 285:430–36, 2001.

[19] D.T. Ko, M. Mamdani, and D.A. Alter. Lipid-lowering therapy with statins in high-risk elderly patients: The treatment-risk paradox. *J Am Med Assoc*, 291:1864–70, 2004.

[20] R.J. Glynn, E.L. Knight, R. Levin, and J. Avorn. Paradoxical relations of drug treatment with mortality in older persons. *Epidemiol*, 12:682–689, 2001.

[21] K.J. Rothman and S. Greenland. *Modern Epidemiology, 2nd ed.* Lippincott, Philadelphia, 1998.

[22] J. Pearl. *Causality, models reasoning and inference.* Cambridge University Press, New York, 1999.

[23] S. Greenland, J. Pearl, and J.M. Robins. Confounding and collapsibility in causal inference. *Statist Sci*, 14:29–46, 1999.

[24] M.A. Hernán, S. Hernández-Diaz, M.M. Werler, and Mitchell A.A. Causal knowledge as a prerequisite for confounding evaluation: An application to birth defects epidemiology. *Am J Epidemiol*, 155:176–84, 2002.

[25] D.B. Rubin. Practical implications of modes of statistical inference for causal effects and the critical role of the assignment mechanism. *Biometrics*, 47:1213–34, 1991.

[26] L. Wasserman. *All of Statistics.* Springer, New York, 2004.

[27] Rosenbaum PR. Propensity score. In. *Encyclopedia of Biostatistics*, pages Ed. P. Armitage and T. Colton, vol. 5. New York, Wiley. (1998), 3551–3555.

[28] P.C. Austin and G.M. Normand, S.T.and Anderson. Conditioning on the propensity score can result in biased estimation of common measures of treatment effect: A Monte Carlo study. *Stat Med*, 26:754–68, 2006.

[29] P.C. Austin. The performance of different propensity score methods for estimating marginal odds ratios. *Stat Med*, 26:3078–94, 2006.

[30] T. Stümer, K.J. Rothman, and R.J. Glynn. Insights into different results from different causal contrasts in the presence of effect-measure modification. *Pharmacoepidemiol Drug Saf*, 15:698–709, 2006.

[31] T. Kurth, A.M. Walker, R.J. Glynn, K.A. Chan, J.M. Gaziano, K. Berger, and J.M. Robins. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*, 163:262–270, 2006.

[32] M.H. Gail, S. Wieand, and S. Piantadosi. Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates. *Biometrika*, 71:431–44, 1984.

[33] T. Stümer, S. Schneeweiss, M.A. Brookhart, K.J. Rothman, J. Avorn, and R.J. Glynn. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: Nonsteroidal anti-inflammatory drugs and short-term mortality in the elderly. *Am J Epidemiol*, 161:891–9, 2005.

[34] M.S. Cepeda, J.T. Boston, R.and Farrar, and B.L. Strom. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. *Am J Epidemiol*, 158:280–7, 2003.

[35] C. Drake. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics*, 49:1231–1236, 1993.

[36] D.B. Rubin and R.P. Waterman. Estimating the causal effects of marketing interventions using propensity score methodology. *Stat Sci*, 21:206–222, 2006.

[37] H. Zheng and R.J.A. Little. Penalized spline model-based estimation of the finite populations total from probability-proportional-to-size samples. *J Off Stat*, 19:99–107, 2003.

[38] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis, 2nd edition.* Chapman Hall/CRC, New York, 2004.

[39] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing:Vienna, 2004. ISBN 3-900051-00-3. URL http://www.R-project.org.

[40] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences. *Stat Sci*, 7:457–511, 1992.

[41] R.J.A. Little. To model or not to model? Competing modes of inference for finite population sampling. *J Am Stat Assoc*, 99:546–56, 2004.

[42] P. Gustafson and B. Clarke. Decomposing posterior variance. *J Stat Plan Inference*, 119:311–27, 2004.

[43] P.C. Austin, P. Grootendorst, and G.M. Anderson. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. *Stat Med*, 2006.

[44] B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *Am Stat*, 37(1):36–48, 1983.

[45] K. Imai and D.A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *J Am Stat Assoc*, 99:854–866, 2004.

[46] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods.* Springer, New York, 2004.

[47] G.C.G. Wei and M.A. Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J Am Stat Assoc*, 85:699–704, 1990.

[48] D. Oakes. Direct calculation of the information matrix via the EM algorithm. *J R Stat Soc Ser B*, 61:479–482, 1999.

# Appendix A

# Computing the limiting estimates for PSA.

In Section 5.2, we apply PSA to synthetic datasets where $P(Y=1|x,c)$ follows the model given in equation (5.1). The PSA method estimates the quantities $\beta^*$ and $\xi^*$ which solve the estimating equation

$$\Omega(\theta) = E\{R[Y - \text{expit}(R'\theta)]\} = \mathbf{0},$$

where $R$ is a $6 \times 1$ vector with the first component equal to $X$, and the last five components equal to $g(z(C,\gamma))$ and where $\theta^* = [\beta^*, \xi_1^*, \xi_2^*, \xi_3^*, \xi_4^*, \xi_5^*]$. For PSA we have $\gamma^*$ equal to $\gamma$, the true values used to generate the data.

We outline a method for calculating the quantity $\theta^*$ numerically for a given simulation design. We may write

$$\begin{aligned} \Omega(\theta^*) &= E\{R[Y - \text{expit}(R'\theta^*)]\} \\ &= E\{R[E[Y|R] - \text{expit}(R'\theta^*)]\} \\ &= \sum_r r[E[Y|r] - \text{expit}(r'\theta^*)]f(r) \\ &= \mathbf{0} \end{aligned}$$

where $r$ is a realization of $R$ and the summation is over the support of $R$. The quantity $f(r)$ is the probability density function of $R$.

Because $R$ is a discrete random variable, we can calculate the quantities $E[Y|r]$ and $f(r)$ numerically by Monte Carlo. To illustrate, consider the realization $r = [1, 1, 0, 0, 1, 0]$, meaning that we have $X = 1$ and $g(z(C, \gamma)) = [1, 0, 0, 1, 0]$. To calculate $E[Y|r]$ and $f(r)$,

1. Draw a large sample of vectors $C$ from the simulation design.

2. Retain the observations such that $g(z(\gamma, c))' = [1, 0, 0, 1, 0]$, and denote this sample as $c_1, c_2, \ldots, c_n$.

3. Compute

$$\begin{aligned}\hat{E}[Y|r] &= \frac{1}{n}\sum_{i=1}^{n} \text{expit}(\tau \times 1 + c_i'\rho) \\ \hat{f}(r) &= \hat{P}\Big(X = 1|g(z(C,\gamma)) = [1,0,0,1,0]\Big) \times \\ &\quad P\Big(g(z(C,\gamma)) = [1,0,0,1,0]\Big) \\ &= \frac{1}{n}\sum_{i=1}^{n} \text{expit}(c_i'\gamma) \times \frac{1}{5}.\end{aligned}$$

where $\tau$, $\rho$ and $\gamma$ are the true parameter values from the simulation design of interest.

Here we have $P\Big(g(z(C,\gamma)) = [1,0,0,1,0]\Big) = \frac{1}{5}$ because of the way the data are simulated. Replicates of $C$ have 20% probability of lying in any specific propensity score bin.

Having computed $E[Y|r]$ and $f(r)$, we can find the solution to the equation $\Omega(\theta^*) = \mathbf{0}$ using the Newton Raphson method. The derivative of $\Omega(\theta^*)$ is given by

$$\frac{\partial}{\partial \theta^*}\Omega(\theta^*) = \sum_r [\text{rexpit}(r'\theta^*)][\text{rexpit}(-r'\theta^*)]' f(r),$$

where the summation is over the support of $R$. Thus we can calculate the $6 \times 1$ vector $\Omega(\theta^*)$ and $6 \times 6$ matrix $\frac{\partial}{\partial \theta^*}\Omega(\theta^*)$ as a function of $\theta^*$. To solve $\Omega(\theta^*) = \mathbf{0}$, we apply the following algorithm:

```
1. Initialize $\theta^*$, for example with $\theta^* = [0, 0, 0, 0, 0, 0]$.

2. Set $inc \leftarrow [1, 1, 1, 1, 1, 1]$

3. While $\text{Max}\{|inc_1|, |inc_2|, |inc_3|, |inc_4|, |inc_5|, |inc_6|\} \geq 10^{-5}$   {
   Set $inc \leftarrow \Omega(\theta^*) \left[\frac{\partial}{\partial \theta^*}\Omega(\theta^*)\right]^{-1}$
   Set $\theta^* \leftarrow \theta^* + inc$
   }

4. Return $\theta^*$
```