ESTIMATION OF HETEROSKEDASTIC LIMITED DEPENDENT VARIABLE MODELS

By

STEPHEN GEOFFREY DONALD

B. Ec. (Honours) University of Sydney

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

DEPARTMENT OF ECONOMICS

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

October 1990

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Economics

The University of British Columbia
Vancouver, Canada

Date 11. October 1990

Abstract

This thesis considers the problem of estimating limited dependent variable models when the latent residuals are heteroskedastic normally distributed random variables. Commonly used estimators are generally inconsistent in such situations. Two estimation methods that allow consistent estimation of the parameters of interest are presented and shown to be consistent when the latent residuals are heteroskedastic of unknown form. Both estimators use recent advances in the econometric literature on nonparametric estimation and deal with the unknown form of heteroskedasticity by approximating the variance using a Fourier series type approximation.

The first estimator is based on the method of maximum likelihood and involves maximising the likelihood function by choice of the parameters of the variance function approximation and the other parameters of interest. Consistency is shown to hold in the three most popular limited dependent variable models — the Probit, Tobit, and sample selection models — provided that the number of terms in the approximation increases with the sample size. The second estimator, which can be used to estimate the Tobit and sample selection models, is based on a two-step procedure, using Fourier series approximations in both steps. Consistency and asymptotic normality are proven under restrictions on the rate of increase of the number of parameters in the approximating functions.

Finally, a small Monte Carlo experiment is conducted to examine the small sample properties of the estimators. The results show that the estimators perform quite well and in many cases reduce the bias, relative to the commonly used estimators, with little or no efficiency loss.

Table of Contents

List of Tables

Acknowledgement

I would like to thank the members of my committee, John Cragg, Harry Paarsch and Brendan McCabe for comments on previous versions of this thesis. I would also like to thank John Cragg and Harry Paarsch for support and encouragement during my time at UBC. The comments of Terry Wales, Ken White, Margaret Slade, Bill Schworm, Jeremy Rudin and classmates are also appreciated. I would also like to thank my parents for their financial and moral support over the last four years.

Chapter 1

Introduction

The use of limited dependent variables in economics has increased dramatically in recent years. Such models often arise in micro-econometric contexts where data sets contain variables which may be qualitative or limited in nature. This has necessitated the explicit introduction of such possibilities into structural economic models and the development of statistical techniques for the estimation of such models. This approach often gives rise to a latent model and a rule determining what will be observed in the data. From this and an assumption governing the distribution of the unobservables in the model one can then proceed to estimate the parameters. A common assumption is that such unobservables are independent and identically distributed (i.i.d.) as normal random variables.

Because of the nature of limited dependent variable models this assumption is not harmless and so the situation differs markedly from other more standard situations such as the linear regression model. In particular, violation of either the normality assumption or the identicality assumption is known to cause estimator inconsistency (the nature of which is still not well understood). While recent attention has been focussed on the normality assumption and, in particular, on obtaining consistent estimators without explicit knowledge of the distribution of the error terms, the identicality assumption has been relatively neglected. This thesis examines this assumption while maintaining the normality assumption, and proposes estimation techniques for some popular limited dependent variable models which yield consistent estimators in situations where there is quite general heteroskedasticty in the unobserved components of the model.

Chapter 2 provides a brief survey of the literature on the estimation of limited dependent variable models (LDVMs) and other theoretical work related to that considered in this paper. It will be apparent that while there have been important advances in our understanding of LDVMs and in estimation technology, much remains to be done before our understanding of LDVMs approaches that of the regression model. It is hoped that this thesis will remedy some of the deficiencies in this literature.

LDVMs are typically estimated (under normality) using either maximum likelihood or some two-step method (which is computationally simpler). Chapter 3 considers the problem of estimating the three workhorse LDVMs: the Probit, Tobit, and sample selection models, using a maximum likelihood type estimator. The estimator proposed in this case allows for heteroskedasticity of unknown form in the latent residual by approximating the variance using a Fourier series type approximation. Consistency results for these three models are proven under the assumption that the number of terms in the approximation increases with the sample size. This technique has been developed by Gallant and Nychka (1987) and is referred to as semi-non-parametric estimation. Similar estimation techniques are referred to in the statistics literature as "method of sieves", see Geman and Hwang (1982). One attractive feature of the estimation technique is ease of computation relative to some other estimators proposed in the literature.

In Chapter 4 we consider two-step type estimators for the Tobit and sample selection models, which are similar to those proposed by Heckman (1979), but which explicitly allow for the fact that there is heteroskedasticity of unknown form in the latent errors. It is shown that heteroskedasticity shows up in the conditional mean of the dependent variable in two places and lends a structure similar to that of the semiparametric regression model. The added complication arises from a first step estimation error. The estimator proposed is similar to estimators belonging to a class of semiparametric estimators, recently proposed by Andrews (1989a), and again use series approximations to the

unknown functions. Results regarding the $\sqrt{N}$ consistent estimation and asymptotic normality of the proposed two-step estimator are proven under fairly weak assumptions and explicit maximal rates of increase of the number of parameters are derived. In addition, consistent estimators for the covariance matrix are proposed and Durbin-Wu-Hausman type specification test statistics are developed.

One nice feature of the estimators proposed in this thesis is that they are relatively easy to compute and may in fact be computed using standard packages such as SHAZAM or TSP. This makes it relatively simple to perform a small Monte Carlo experiment to examine how the estimators perform relative to the commonly used estimators in small samples. In addition, some evidence on the required number of terms for good performance can also be provided although the evidence cannot look at all possible situations. As is the case with other nonparametric type estimators, it is found that the bias can be reduced even with a small number of terms in the approximating functions. In the case of the two-step estimators, this usually comes at the cost of a loss in efficiency, but in the case of the MLE, allowing for heteroskedasticity may even improve the efficiency (in small samples) of the estimates. Chapter 6 concludes the thesis with a brief summary of the contributions and points to areas of future research.

Chapter 2

Literature Survey

In this chapter we survey briefly some of the literature concerning the estimation of limited dependent variable models, identifying important contributions and deficiencies in this area. We then identify some of the contributions to this literature made in this thesis. More complete surveys on LDVMs which also survey some of the applications of these types of models may be found in Maddala (1983), McFadden (1985), Amemiya (1985), Dhrymes (1986), and Maddala (1986). In addition, recent contributions to the theoretical literature are contained in Duncan (1986a) and Blundell (1987) and a survey of some specification testing literature is contained in Pagan and Vella (1988). This survey is by no means complete and concentrates on the econometric literature in this area, recognising that some of these models have a longer history in statistics.

The need for LDVMs in economics is quite obvious. Often, especially in microeconomic situations, one is confronted with data that are qualitative or limited in nature. Examples include: car or electrical appliance ownership; hours worked by an individual; expenditure on durables and so on. The first example of empirical work in econometrics using a model which takes into account this type of situation is Tobin (1958) in his analysis of household expenditure on durable goods. In this situation, one is confronted with a non-negligible portion of households with zero expenditure in a given year. Tobin noted that the usual least squares estimator in this situation is inconsistent, even if the random errors are iid normally distributed, because a negative expenditure is impossible (at least in observed data). Tobin proposed a maximum likelihood (MLE) estimator by

constructing the logarithm of the likelihood function under the assumption of iid normal latent errors. His model has become known as the "Tobit" model.

The next paper to consider this type of situation in economics appears to be Cragg (1971) who presented a number of more general models, some of which contained bivariate random variables and which were estimated under the same assumptions used by Tobin (1958). These can be considered as forerunners to the now popular 'sample selection model' attributed to Heckman (1974, 1976, 1979) and Gronau (1973). All of these models have been termed 'generalised' Tobit models by Amemiya (1985). An alternative method of estimation for these models based on a two-step procedure was proposed in Heckman (1976, 1979). Amemiya (1973) proved the consistency and asymptotic normality of the MLE of the Tobit model under the assumption of iid normal random errors. Amemiya (1985) provides a useful survey of Tobit type models and various applications.

In the case of discrete dependent variables, the first papers to appear in the econometric literature appear to be Theil (1969) and McFadden (1974) — the latter relating such models to choice decisions of individuals and hence incorporating utility functions. As in the case of the Tobit type models, the early contributions appear to be directed at generalisations of the simple zero-one case and, although in the case of discrete choice models, distributions apart from the normal one have been used to construct the likelihood functions (notably the logistic, giving rise to the Logit model) little attention has been paid to the possible mis-specification bias resulting from invalid distributional assumptions. McFadden (1985) provides a useful summary of the literature on discrete choice models.

Since these early contributions, researchers have begun to worry more about the extreme distributional assumptions needed for consistency and examined to what extent they can be relaxed. An important paper by Robinson (1982) showed that for the Tobit model the usual MLE remains consistent when observations are dependent. The only

problem is to obtain a consistent estimate of the covariance matrix of the MLE. In addition, the MLE is inefficient in this case. As Dagenais (1983) has shown, computing the true MLE, even in the simple case of an AR(1) error term can be computationally intractable. Parallel work in the case of the discrete choice or Probit model shows that the MLE in this case is also consistent with dependent data, see Gourieroux, Monfort and Trognon (1984) and Poirier and Ruud (1988). Estimation taking into account the dependence has been examined by Poirier and Ruud (1988).

Most work concerning robustness has been focussed on the normality assumption which is in general needed for consistency, see Ruud (1983). Due to the highly nonlinear nature of these models, the precise nature of the inconsistency is poorly understood. There has been some evidence provided in very simple cases by Arabmazar and Schmidt (1982) for the Tobit model, who show that the biases can be 'large' and are worse the higher the degree of censoring. Also Paarsch (1984) has shown that in finite samples the usual MLE may perform quite poorly when the true error distributions have fat tails.

Given that our understanding is so limited, it is unsurprising that much attention has been devoted to the problem of estimating LDVMs without making explicit distributional assumptions. Perhaps the first papers to do this are Manski (1975, 1985) who developed the Maximum Score (MS) estimator for the discrete choice model. This estimator chooses the parameter vector that maximises the number of correct predictions (regarding choice) and is consistent under quite general conditions, only requiring a median restriction. The estimator is robust not only to nonnormality, but also heteroskedasticity. One drawback of the estimator is that it is difficult to compute and that it is not asymptotically normally distributed. Moreover, it is less than $\sqrt{N}$ consistent (and is in fact only $N^{1/3}$ consistent). A smoothed version of the estimator has recently been proposed by Horowitz (1989) which is asymptotically normally distributed but is still less than $\sqrt{N}$ consistent.

Similar work for Tobit type models has been done by Powell (1984, 1986). The first

estimator proposed is similar to the MS estimator in that it is consistent under a restriction on the median of the latent errors. This estimator, known as the Censored Least Absolute Deviations (CLAD) estimator, works because the median is invariant to censoring provided that it is positive (assuming that censoring is at zero). A fundamental identification condition requires that a non-negligible part of the data comes from conditional densities that where this is true. Like the MS estimator the CLAD is robust to heteroskedasticity and the only impediment to its use is computation which is quite difficult. The second estimator proposed is known as the Symmetrically Censored Least Squares (SCLS) estimator. It is based on the observation that if one could censor the observations symmetrically about the mean (assuming symmetric distribution of the latent errors) then ordinary least squares applied to the symmetrically censored data would be consistent. An objective function which does this (asymptotically) is set up and estimation is based on a sort of reweighted least squares type algorithm. A similar approach works in the truncated regression model where trimming takes the place of censoring. As with the CLAD estimator these estimators are robust against heteroskedasticity and only require a symmetric and unimodal distribution.

Despite the attractiveness of the estimators of Manski and Powell they have not been used much in practice. This appears to be in part due to the computational difficulties and also the fact that they tend to be inefficient. Thus based on an MSE criteria they may not be much better than the usual estimators (see Paarsch (1984) and Powell (1986)). This has led to some other work which has been concentrated on obtaining estimators based on maximum likelihood type criteria. Work in this direction has been done by Cosslett (1983) for the discrete choice model and by Duncan (1986b) and Fernandez (1986) for the Tobit model. In these estimators, the likelihood function is maximised by choice of the parameters in the model and the density function of the latent errors. To make this a meaningful exercise some restrictions must be made in finite samples,

but consistency results obtain due to the fact that the restrictions are relaxed as the sample size grows. Duncan (1986b) uses splines to approximate the density of the errors, while Fernandez (1986) uses Fourier series approximations and estimation is performed by choice of the parameters of the approximation and the parameters in the model. Unfortunately, for these estimators only consistency results have been obtained thus far. Another drawback is that while the estimators are robust to nonnormality they maintain an identicality assumption and so are not robust to heteroskedasticity. The estimation method proposed by Matzkin (1988) which maximises by choice of both the distribution function and the regression function, may circumvent this problem, but the estimator seems difficult to employ and will give estimates that may be somewhat difficult to interpret.

Similar work to this has been done by Ichimura (1988) as well as Klein and Spady (1987) for the discrete choice model and Horowitz (1986, 1988) for the Tobit model. In these cases, however, the estimators have been shown to be consistent and asymptotically normal and in the case of Klein and Spady the estimator attains the semiparametric efficiency bound. These estimators use kernel estimation and exploit the single index nature of the moments of the dependent variable. Note too, Andrews (1988, 1989d) has shown recently that similar results will hold if series are used instead. As with the estimation methods based on the likelihood function, these estimators require the assumption that the latent error terms are identically distributed, or the restrictive assumption that the distribution depends on the indpendent variables only through the index (which is also the mean of the latent variable).

Thus far we have neglected any discussion of the sample selection model. This is because it differs in an important way from the other models in that it contains a bivariate random variable. This makes it impossible to employ estimators analogous to those of Manski and Powell since quantile restrictions in bivariate models are somewhat

problematic. This means that the likelihood based methods or some alternative methods have to be used. Gallant and Nychka (1987) have proposed a likelihood based method similar to the methods of Duncan and Fernandez, but approximate a bivariate density using hermite polynomial approximation. They also give quite general results for this type of situation. The other recently proposed method for the sample selection model is based on two-step methods similar to the Heckman (1979) estimator but which allow for an arbitrary density of the unobserved latent errors. The first paper to do this is Cosslett (1984) who derived consistency results. Also in this vein is the work of Powell (1987) and Newey (1988b) who use kernel estimation and polynomial approximation respectively. Both estimators have been shown to be $\sqrt{N}$ consistent and asymptotically normally distributed under certain conditions. In the first stage one must use a method such as that of Ichimura or Klein and Spady, which are $\sqrt{N}$ consistent and asymptotically normal. The distribution of the second step estimator will depend on the distribution of the first step estimator. The results in Andrews (1989d) also cover this situation as does the semiparametric regression results of Robinson (1988) although the identification conditions in his case are quite restrictive. An application of these methods by Newey, Powell, and Walker (1990) has shown that in the case of the Heckman (1974) model of labour supply (and the subsequent work of Mroz (1987)) the results do not differ much from those obtained under the assumption of normality.

Common in much of this work is the assumption that the latent errors are identically distributed and it seems likely that the estimators (apart from those of Powell and Manski) will be inconsistent in the presence of heteroskedasticity. As is the case with the normality assumption, little is known about the properties of the usual estimators when the homoskedasticity assumption is violated. Only Hurd (1979) and Arabmazar and Schmidt (1981) have considered the nature of the asymptotic bias caused by heteroskedasticity in very simple situations. Powell (1986) in his Monte Carlo experiments

has provided some finite sample evidence on the nature of the bias in the context of the Tobit model and was led to the conclusion that "...failure of the homoskedasticity assumption may have more serious consequences than failure of normality in censored regression models." Moreover, since these models are typically estimated using cross-sectional data where there is often considerable heterogeneity, it is somewhat surprising that more work on LDVMs with heteroskedasticity has not been done. The only contributions appear to be those of Powell (1984, 1986) and Manski (1975, 1985) which are limited to univariate models and have seen little application. In fact, it appears that no remedies for the problem of heteroskedasticity in the sample selection model have yet been proposed. The contribution of this thesis is an attempt to remedy this deficiency in the literature. In particular, while maintaining the normality assumption (in which case the heteroskedasticity shows up as changing variance across observations) we propose estimation methods for the three main LDVMs, which yield consistent and in some cases asymptotically normal estimators, when the errors are heteroskedastic of unknown form. One nice feature of the estimators proposed is that they are relatively easy to compute as evidenced by the ease with which we can perform a Monte Carlo experiment. As a by product we present in Chapter 5 some small sample evidence on the effect of heteroskedasticity on these models

We should also briefly mention some of the technical work that is used in this paper. The main tool used to overcome the unknown nature of the variance function is the Fourier series approximation method proposed in the econometric literature by Gallant (1981) for estimation of regression models. The basic idea is to substitute the Fourier series approximation for the variance and to choose the parameters based on the maximisation of some objective function. In Chapter 3 the objective will be the likelihood function, while in Chapter 4 the objective will be the sum of squared deviations of the dependent variable from its expected value. The proof of consistency in the first case is

similar to that of Gallant (1987) and Gallant and Nychka (1987), while in the second case we employ some of the results of Robinson (1988), Andrews (1988, 1989d) as well as Andrews and Whang (1989) although the situation is somewhat different from any of these. The methods used in the second case differ from those proposed in Andrews (1989a,b,c) which yield consistency and asymptotic normality for situations similar to that considered in this thesis.

Chapter 3

Maximum Likelihood Estimation Methods

3.1 Introduction

Likelihood based methods are now quite common in applications of LDVMs. This is especially the case in the Probit and Tobit models. Recent contributions by Cosslett (1983), Duncan (1986b) and Gallant and Nychka (1987) have examined generalising these methods to situations where the density of the latent errors in LDVMs is not restricted to a particular parametric family. Estimation proceeds by maximising the likelihood by choice of a density and the usual parameters of interest. The density is usually approximated by a function which is capable (in the limit) of approximating an arbitrary density function. Hence the terminology nonparametric maximum likelihood estimation. A common assumption is that the latent errors are identically distributed. The robustness of these types of estimators to heteroskedasticity is as yet unknown, but from a theoretical viewpoint it appears inconsistency will result.

In this chapter, we propose an estimator, which under the assumption of normality, provides consistent estimates of the parameters of interest in the presence of heteroskedasticity of arbitrary unknown form. We deal with this case because when there is normality, heterogeneity takes on a form which can clearly be examined. The particular type of heterogeneity to be examined is a dependence of the scale parameters on the exogenous variables in the model. This is much like heteroskedasticity in the regression model, but unlike the regression model it produces inconsistent estimates (although little

is known about the nature of this inconsistency).

The basic idea of the estimator is to approximate the scale parameters with the Fourier Flexible Functional (FFF) form, and then to maximise the likelihood function with respect to the parameters of interest and the parameters of the approximating function. Consistency obtains because as the sample size increases, the number of terms in the approximation increases and the approximation error reduces. The advantage of this approach is that the estimates can be obtained using standard maximum likelihood methods. This approach is similar to recent work by Gallant and Nychka (1987) (hereafter GN) who approximate the shape of the density function with a Hermite polynomial series. The proof of consistency in this chapter and GN (and most papers considering consistency of optimization type estimators) is essentially adapted from the classical proof of the consistency of the Maximum Likelihood Estimator (MLE) in the independent and identically distributed (iid) case by Wald (1949).

The major drawback of the approach is the assumption of normality which is maintained. Future work could examine the possibility of consistent estimation with arbitrary distributional shape and heterogeneity. The estimation technique does, however, appear to be useful in situations where there is an infinite dimensional nuisance parameter which must be taken into account before consistent estimates of the parameters of interest can be obtained.

The remainder of the chapter is organised as follows. Section 2 gives a heuristic discussion of the estimator and gives a general result on consistency which is adapted from GN. The remaining sections present the specific models of interest and provide assumptions on these models which guarantee that the conditions of the theorem are satisfied. The chapter concludes with a discussion of the practical implications of the chapter and suggests areas where further work is needed.

3.2 Heuristics

The type of situation we have in mind is the following. The underlying unobservables in the model are generated according to the following density function $f(u|\sigma(x))$ where f is a known density function and $\sigma(x)$ is an unknown function of a vector of exogenous variables x belonging to on some set X, which is a subset of Euclidean space. Typically f will be chosen to be normal so that $\sigma^2(x)$ is the variance parameter. This would give rise to the familiar heteroskedastic regression model if u were an additive disturbance in the usual regression model.

Observations of the dependent variable z, which is possibly a vector, are presumed to come from some parametrically specified model with parameter vector β but are observed with error because of some, possibly discontinuous, dependence on the unobserved errors u. It is also presumed that we are dealing with a situation in which the MLE of β assuming $\sigma(x)$ is a constant, is inconsistent (unlike the case of the linear regression model). This is generally the case with the LDVMs considered in this chapter. The parametric assumptions and the density f give rise to a density of the observed variables z (conditional on x) which we denote by $p(z|x,\beta,\sigma(x))$ and a sample average logarithm of the likelihood function (assuming independence of observations) the negative of which is given by

$$s_n(\beta,\sigma) = (-1)(1/n)\sum_t \log p(z_t|x_t,\beta,\sigma)$$

The true values of β and σ will be denoted β^* and σ^*. The objective is to estimate β^* consistently by minimising this function by choice of β, belonging to some set B, and the function σ^*, assuming that σ belongs to some set of functions denoted Σ.

The arguments used to achieve this end follow the classical lines of the proof of consistency of the MLE by Wald (1949) which have been used repeatedly since. Our particular problem bears some resemblance to recent work by Gallant and Nychka (1987) who also

follow Wald's methods, but they are interested in maximising the likelihood function over β and the density f, under the maintained assumption that the u variables are iid random variables. A whole host of other types of semi-(non)-parametric MLEs can be cited (Matzkin (1988), Cosslett (1984) *etc.*) but it appears that most have been concerned with shape rather than with heterogeneity across observations. (Matzkin (1988), however, can deal with both for the discrete choice model.)

The intuition behind the approach taken in this chapter is clear. Since the set Σ will be infinite dimensional, the exercise of minimising s_n by choice of β and the function σ will not provide estimators which converge to the true values in any meaningful sense. What is done instead, is to minimise over some subset of Σ, denoted Σ_K, where the restriction is such that there may be only a finite number of parameters, which we denote K. For example Σ_K could contain trigonometric polynomial functions of x and K parameters, and one would proceed by estimating β and the parameters of the polynomial. In our case this will represent an approximation to the σ function. The consistency result, for the estimators of β^* and σ^*, which will denote the true values, is obtained by letting the number of parameters, K increase with the sample size, provided that the set Σ_K becomes "close" to Σ, in some sense. The sense in which thay must be close is that as K increases, there should exist functions in Σ_K which are capable of approximating the true function in Σ as measured by some norm denoted $\|.\|$. This corresponds to the denseness condition in the theorem presented below, which is a slight modification of Theorem 0 of GN. Note that since K will be required to increase with n, we index it by n as K_n.

Theorem 1: *Let $\hat{\beta}_n$ and $\hat{\sigma}_n$ minimise $s_n(\beta, \sigma)$ over $B \times \Sigma_{K_n}$ where $B \subset R^k$ is compact in the Euclidean norm $\|.\|_E$ and Σ_{K_n} is a subset of Σ on which is defined a norm $\|\sigma\|$. Suppose the following conditions hold;*

(a) Compactness: The closure of Σ w.r.t $\|.\|$ is compact in the relative topology generated by $\|.\|$,

(b) Denseness: $\bigcup_{n=1}^{\infty} \Sigma_{K_n}$ *is a dense subset of* Σ *w.r.t* $\|.\|$,

(c) Uniform Convergence: There are points (β^*, σ^*) *in* $B \times \Sigma$ *and there is a function* $\bar{s}(\beta, \sigma, \beta^*, \sigma^*)$ *that is continuous in* (β, σ) *w.r.t.* $\|\beta\|_E + \|\sigma\|$ *and*

$$\lim_{n\to\infty} \sup_{B\times\bar{\Sigma}} |s_n(\beta, \sigma) - \bar{s}(\beta, \sigma, \beta^*, \sigma^*)| = 0$$

(d) Identification: Any point (β^0, σ^0) *in* $B \times \bar{\Sigma}$ *with* $\bar{s}(\beta^0, \sigma^0, \beta^*, \sigma^*) \leq \bar{s}(\beta^*, \sigma^*, \beta^*, \sigma^*)$ *has* $\|\beta^0 - \beta^*\|_E = 0$ *and* $\|\sigma^0 - \sigma^*\| = 0$.

Then $\lim_{n\to\infty} \|\hat{\beta}_n - \beta^*\|_E = 0$ *a.s. and* $\lim_{n\to\infty} \|\hat{\sigma}_n - \sigma^*\| = 0$ *a.s. provided that* $\lim_{n\to\infty} K_n = \infty$.

Proof: For the proof of this result and all others in Chapter 3, see Appendix A.

Our approach is somewhat similar to GN, in that we obtain consistency by using an approximation of the unknown function. In our case we are concerned with the scale parameter σ, whereas GN were concerned with the density function parameters. In addition, the nature of the approximation in our case will be very different from that used in GN.

The conditions (a), (c) and (d) are fairly standard conditions for consistency results involving extremum type estimators. Condition (b) is the major modification, and as mentioned earlier, essentially requires that the approximation for σ be capable of becoming arbitrarily "good" as the number of terms increases. In the next section, we present the three models to be examined in this chapter. The subsequent sections then make primitive assumptions on these models that guarantee that the above conditions for Theorem 1 are satisfied. The first two conditions can be examined in fairly general terms without reference to any particular model. The other two conditions must be verified specifically for each model. It will become apparent that one set of primitive assumptions will guarantee conditions (c) and (d) for almost any LDV model that may arise in practice (at least when normality is maintained).

3.3 Models of Interest

The structures of the Probit and Tobit models are fairly similar as is the way in which they fit into the framework of section 2. Consider some latent variable y generated by the following (where t subscripts have been suppressed for notational convenience),

$$y = x\beta + u.$$

In the Probit model, one observes the following endogenous variable

$$z = I(y \geq 0)$$

(where $I(A)$ is the indicator function for the event A) while in the Tobit model one observes

$$z_1 = I(y \geq 0)$$

$$z_2 = yI(y \geq 0).$$

The densities of the observables are respectively

$$p(z|x, \beta, \sigma) = F(x\beta|\sigma)^z \left[F(x\beta|\sigma)\right]^{(1-z)}$$

$$p(z_1, z_2|x, \beta, \sigma) = f(z_2 - x\beta|\sigma)^{z_1} \left[1 - F(x\beta|\sigma)\right]^{1-z_1}$$

where

$$F(x\beta|\sigma) = \int_{-\infty}^{x\beta} f(u|\sigma(x))du.$$

In both these models f is typically chosen to be a normal density with variance $\sigma(x)$ so we can write

$$\begin{aligned} f(u|\sigma) &= \frac{1}{\sigma(x)}\phi(\frac{u}{\sigma(x)}) \\ F(x\beta|\sigma) &= \Phi(\frac{x\beta}{\sigma(x)}) \end{aligned}$$

where ϕ and Φ are the standard normal density and distribution functions.

In both these cases, we will be minimising $s_n(\beta, \sigma)$ by choice of β and the R^1 valued function σ where $\sigma > 0$. More details on the required properties of this function will be given in Section 4.

The sample selection type model involves a bivariate normal density function and the two latent variables given below

$$y_1 = x\beta_1 + u_1$$

$$y_2 = x\beta_2 + u_2.$$

The observed endogenous variables in this case are

$$z_1 = I(y_1 \geq 0)$$

$$z_2 = z_1 y_2.$$

In this case, the vector (u_1, u_2) has the bivariate normal density with covariance matrix Ω which can be written as follows

$$\Omega(x) = \begin{pmatrix} \sigma_1^2(x) & \sigma_{12}(x) \\ \sigma_{12}(x) & \sigma_2^2(x) \end{pmatrix}$$

Because this must be positive definite for all the observations, it will be convenient to impose some structure on it and to redefine the covariance matrix. In particular, we assume that it can be written as follows

$$\Omega(x) = \begin{pmatrix} \sigma_1^2 q_1(x)^2 & \sigma_{12} q_1(x) q_2(x) \\ \sigma_{12} q_1(x) q_2(x) & \sigma_2^2 q_2(x)^2 \end{pmatrix}$$

where the two functions q_1 and q_2 will be restricted to be positive. Positive definiteness can then be imposed by making sure that the constant matrix

$$\Psi = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

is positive definite. Note that this structure allows covariance terms to be heteroskedastic but imposes constancy on the correlation coefficient. Also no restriction is placed on the sign of this correlation. Such a structure would arise if

$$u_1 = q_1(x)\epsilon_1$$

$$u_2 = q_2(x)\epsilon_2$$

with (ϵ_1, ϵ_2) having a homoskedastic bivariate normal density with covariance matrix Ψ. The density in this case is complicated and is given below where the above assumptions have been imposed. The parameter ρ_{12} is the constant correlation coefficient.

$$p(z|\Omega(x), x, \beta) = \left[1 - \Phi(\frac{x\beta_1}{\sigma_1 q_1(x)})\right]^{(1-z_1)} \times$$

$$\left[\Phi((\frac{x\beta_1}{\sigma_1 q_1(x)} + \frac{\rho_{12}}{\sigma_2 q_2(x)}(z_2 - x\beta_2))(1-\rho_{12}^2)^{-1/2})\frac{1}{\sigma_2 q_2(x)}\phi(\frac{z_2 - x\beta_2}{\sigma_2 q_2(x)})\right]^{z_1}$$

Note that it is impossible to identify σ_1 and σ_2, unless q_1 and q_2 are normalised, although this is unnecessary since the main concern is with the β coefficients. Also note that in this case and in the Probit case some normalization of β_1 will be required to help identify the other parameters. This is also obviously the case in the homoskedastic models where it is typically assumed that $\sigma_1 = 1$.

In the sample selection case, we minimise the objective function by choice of the two q functions, the two sets of coefficients β_1 and β_2 and the correlation coefficient ρ_{12}. The restrictions will be that the q functions are positive and that $|\rho_{12}| \leq 1 - \epsilon$ for some $\epsilon > 0$. The second of these restrictions is easy to impose. The problem of imposing positivity on the approximating functions in all these models will be discussed in Section 7 where we look more closely at the practical implications of this work. The next section considers in detail the technical side of conditions (a) and (b) with regard to these models and discusses the nature of the approximating functions.

3.4 Approximating the Variance

In this section, we discuss the way in which the unknown variance and covariance functions will be approximated. The functions we wish to approximate are σ for the Probit and Tobit models and q_1 and q_2 for the sample selection model. All the discussion will apply to each of these functions separately. Before deciding on a method of approximation, we must first decide what minimal set of conditions these functions must satisfy. We will refer to the function of interest generically as g. Then we must decide the sense in which the approximating function approximates g. This is done with reference to some norm. This will then be termed the consistency norm for the function g. We denote this norm by $\|.\|$ and for the moment we define it by

$$\|g\| = \sup_{x \in X} |g(x)|$$

where g is defined on the set X. We will assume that X is an open bounded subset of k-dimensional Euclidean space where, without loss of generality, we assume that the closure of X is contained in the k-dimensional open cube defined by the intervals $(0, 2\pi)$. This assumption is motivated by the fact that the type of approximation to be used is based on Fourier series which are periodic of period 2π. Since g will not be assumed to be periodic, this type of approximation will only be able to approximate g on X if cl(X) is in the open cube, see Gallant (1981).

The following notation will be useful in what follows. Let

$$D^{\lambda} g(x) = \frac{\partial^{|\lambda|}}{\partial x_1^{\lambda_1} ..... \partial x_k^{\lambda_k}} g(x)$$

where $|\lambda| = \sum_{i=1}^{k} \lambda_i$ and λ_i are nonnegative integers. If $\lambda = 0$ then $D^0 g(x) = g(x)$. We make use of the following pseudonorms (which are discussed fully in Adams (1975))

$$\|g\|_{m,\infty,\mu} = \max_{|\lambda| \leq m} \operatorname{ess\,sup}_{x} |\mathrm{D}^{\lambda} \mathrm{g(x)}|$$

where the essential supremum is with respect to the probability measure μ defined on X. That is we assume that x is a random vector on X, with probability measure given by μ. These norms are termed Sobolev norms and give rise to Sobolev spaces which are subspaces of the more familiar L^p spaces.

In discussing the issues of compactness and denseness, we follow the same arguments as in Elbadawi, Gallant, and Souza (1983) (hereafter EGS). We define the set of g functions as G. Assumption 1 gives the conditions we wish the g functions to satisfy.

Assumption 1: *G consists of all functions g defined on X such that i)$D^\lambda g(x)$ is continuous on X for all λ such that $|\lambda| \leq 3$; ii) $\sup_x |D^\lambda g(x)| < \infty$ for $|\lambda| = 3$; iii)$\|g\|_{2,\infty,\mu} \leq b < \infty$; iv) $\inf_x g(x) \geq \delta > 0$.*

Note that these assumptioñs imply that up to third order partial derivatives are continuous (i), and that certain smoothness conditions are satisfied by the function. The last assumption (iv) ensures that the variance terms are bounded away from 0 uniformly in x. The type of approximating function that we use is a variant of a multivariate Fourier series function. It may contain linear and quadratic terms in x, but the remaining terms in the function will be trigonometric. In talking about the number of terms of the approximating function going to infinity, we will be refering to the trigonometric terms. If the function were to contain quadratic terms, then it would coincide with the Fourier Flexible Functional (FFF) form of Gallant (1980). The addition of quadratic terms to the Fourier terms is meant to improve the approximation so that fewer trigonometric terms are needed to obtain a "good" approximation. These additional terms, however, will not affect any of the theory regarding the approximation of a function by Fourier Series.

The FFF is given by

$$g_K(x) = a + b'x + \frac{1}{2}x'Cx +$$

$$\sum_{\alpha=1}^{A}(u_{0\alpha} + \sum_{j=1}^{J}(u_{j\alpha}\cos(jk'_{\alpha}x) - v_{j\alpha}\sin(jk'_{\alpha}x)))$$

with a, b, C and u_{ij} being parameters, and where the k_{α} are a sequence of multi-indices (which are integer valued k-dimensional vectors) given in Gallant (1981) and

$$C = -\sum_{\alpha=1}^{A} u_{0\alpha}k_{\alpha}k'_{\alpha}$$

The degree of the trigonometric polynomial will be denoted by K and satisfies

$$j\sum_{i=1}^{k}|k_{i\alpha}| \leq K$$

for all $\alpha \in \{1, 2, ..., A\}$ and $j \in \{1, 2, ..., J\}$. As noted by EGS, A and J will depend on K. For a given sample size n, we let the degree of the polynomial be K_n. In order to obtain a consistent estimate for the function, we will need the number of parameters in the approximating function to increase with the sample size. That is, we need that $K_n \to \infty$. As noted by Gallant (1980), the rule determining the degree K may be adaptive (data dependent) or fixed.

A convenient way of writing the approximating function is

$$g_K(x|\theta) = \sum_{j=1}^{K}\theta_j\psi_j(x)$$

Here the θ_j are coefficients and the ψ_j are the elementary functions used (a constant will be included). The attractive feature of these types of approximations is that as K goes to infinity the above function becomes capable of approximating g in a sense that will now be discussed. In particular, letting θ be an infinite dimensional vector of parameters and given Assumption 1 and the corollary to Theorem 1 in Gallant (1981) then there exists a vector θ^* such that

$$\lim_{K\to\infty}\|g - g_K(\theta^*)\|_{2,\infty,\mu} = 0. \tag{3.1}$$

In practice, it would appear that there would be little problem in just substituting the approximating function into the objective functions defined in Section 4 and minimising over β and the θ parameters using standard maximum likelihood type procedures. To obtain consistency, however, we must overcome a number of technical difficulties to ensure that the conditions of Theorem 1 are satisfied. We follow the line of argument of EGS with some modification.

From above, we know that there is a vector θ^* such that the corresponding approximating function represents g in the sense defined above. We will define the pointwise limit of this approximating function by g_∞. We would like to be able to just replace the g function with the corresponding g_∞ function and proceed from there. There are, however, a few problems we must overcome. First, we want to make sure that g_∞ is at least continuous. Second we would like the set of these g_∞ functions, corresponding to the g functions satisfying Assumption 1, to be contained in a set which is compact with respect to $\|.\|$. We introduce some more notation. Let $\|g_\infty\|_{2,\infty,\mu} = \|\theta\|_{2,\infty,\mu}$. Define the set Θ by

$$\Theta = \{\theta : \lim_{K\to\infty} \max_{|\lambda|\leq 2} \sup_x |\sum_{j=1}^{P_k} \theta_j D^\lambda \psi_j(x)| < \infty\}$$

and the set Θ_0 by

$$\Theta_0 = \{\theta \in \Theta : \|\theta\|_{2,\infty,\mu} \leq b, \inf_x \sum_{j=1}^{\infty} \theta_j \psi_j(x) \geq \delta\}$$

Now by (3.1) for every g in G there is a θ in Θ_0 which corresponds to it in the sense of $\|.\|_{2,\infty,\mu}$. The the following lemma is the first part of Lemma 2 of EGS and is useful in obtaining the needed compactness and continuity discussed above.

Lemma 1: *There is a set Θ^* such that for all θ in Θ^* $g_\infty(x|\theta)$ is continuously differentiable to order 1 and there is a continuous mapping from the weak* closure of Θ_0, denoted by $\bar{\Theta}_0$, onto Θ^*. Moreover the set Θ^* is compact in the relative topology generated by $\|.\|_{1,\infty,\mu}$.*

Proof: See Elbadawi, Gallant, and Souza (1983).

This lemma provides us with the desired continuity and compactness. We can now represent the g function by some θ in Θ^* as $g_\infty(x|\theta)$. This g_∞ function will be continuous and so is identical to the g function in the consistency norm.

Next, since we have that all the g_∞ are continuous then

$$\|g_\infty\| = \|g_\infty\|_{0,\infty,\mu} \leq \|g_\infty\|_{1,\infty,\mu}$$

where the weak inequality is obvious from the definition of the Sobolev norm given above. This result then gives the result that convergence in $\|.\|_{1,\infty,\mu}$ implies convergence in $\|.\|$ so that Θ^* will also be compact in the topology generated by the consistency norm.

Next, we discuss the estimation space. Since it is clearly infeasible in finite samples to minimise the objective function by choice of the complete vector θ we must truncate the vector at some point say K_n which may depend on the sample size n. In terms of the above we represent this as minimising over the subset of θ vectors in $\Theta^* \cap \Theta$ which have the form $(\theta_1, \ldots, \theta_{K_n}, 0, \ldots)$ There will clearly not be any problem of emptiness here since if $b > \delta$ then the element which has all zeroes except the first and that is between b and δ, is clearly of the appropriate form and is also in $\Theta^* \cap \Theta$. We denote the set of vectors with K_n nonzero elements by $\Theta^*_{K_n}$. Clearly, we will then have that $\bigcup_{n=1}^{\infty} \Theta^*_{K_n} = \Theta^* \cap \Theta$ provided that $K_n \to \infty$ so that condition (b) of Theorem 1 is satisfied. In actual practice, when performing the optimisation we will have two sets of constraints that will force the function to be in $\Theta_0 \cap \Theta \subset \Theta^* \cap \Theta$. These will be that

$$\sup_x |D^\lambda g_{K_n}(x|\theta)| \leq b$$

for all $|\lambda| \leq 2$, and

$$\inf_x g_{K_n}(x|\theta) \geq \delta > 0.$$

The required continuity will be clearly satisfied. EGS note that the fact that we are optimising over a subset of the function space is not likely to be of much importance

in practice. As far as imposing the above restrictions goes, the first does not seem to pose any practical problems since we can always make b quite large.The non-negativity constraint, however, does appear to pose some problem in practice. More on this will be taken up later.

So far we have only considered the case of a single variance parameter. In the case of a bivariate model, we must estimate two functions. This is easily handled by the above discussion since we can treat each function as a separate parameter and consider compactness and denseness separately. The norm of the variance covariance matrix can just be defined as the sum of the norms of the two functions. The product space $\Theta^* \cap \Theta \times \Theta^* \cap \Theta$ will then be compact in this norm. Minimisation will then be done over two sets of θ parameters.

3.5 Uniform Convergence

In this section we provide conditions which guarantee that (c) of Theorem 1 is satisfied for the three models of interest. That is, we verify that there exists a function $\bar{s}(\beta, \sigma, \beta^*, \sigma^*)$ that is continuous in (β, σ) with respect to the norm $||(\beta, \sigma)||$ and

$$\lim_{n \to \infty} \sup_{B \times \Sigma} |s_n(\beta, \sigma) - \bar{s}(\beta, \sigma, \beta^*, \sigma^*)| = 0$$

(Note that in the sample selection model the ρ_{12} parameter should be included with the β parameters and will be restricted to lie in $[-1+\epsilon, 1-\epsilon]$). In addition, to our Assumption 1, we make the following assumption on the exogenous variables.

Assumption 2: *x_t is a Cesaro sum generator with respect to $\mu(x)$ on X and (u_t, x_t) is a Cesaro sum generator with respect to $f(u|\sigma^*(x))dud\mu(x)$ where σ^* is the true variance parameter(s).*

The meaning of this assumption is that limits of Cesaro sums of the random variables of interest can be computed as integrals. Moreover, the same applies to any continuous

function which is dominated by an integrable function. This allows for completely exogenous variables (GN), so no identicality of distributions is needed. Essentially, all that is needed is that some law of large numbers is applicable.

The following three lemmata give the required results for the three models of interest. In addition to assumptions 1 and 2, we use the fact that B is compact. This along with assumption 1 put bounds on the regression function part $x\beta$ and keeps σ away from 0 or ∞ and these facts are used heavily in the proofs of the lemmata. Also, convergence in the metric $\|.\|$ implies pointwise convergence of the function and this is exploited.

Lemma 2: *For the Probit model, given Assumptions 1 and 2 and compactness of B, (c) of Theorem 1 is satisfied.*

Lemma 3: *For the Tobit model, given Assumptions 1 and 2 and compactness of B, (c) of Theorem 1 is satisfied.*

Lemma 4: *For the sample selection model, given Assumptions 1 and 2 and compactness of B, (c) of Theorem 1 is satisfied.*

The proofs are somewhat simpler than the proof of the result in GN which is for the sample selection model. That proof is complicated by the fact that they were interested in approximating the shape of the density function. Note that in the proofs heavy use is made of the fact that the x variables are bounded. This comes about because of the type of approximation we have used, which will only work for a function defined on a bounded set. If we were able to choose an approximating function that was able to approximate the variance terms over the whole real plane then some assumptions about the integrability of certain functions of x would be required.

3.6 Identification

In this section, we examine the identification of the three models of interest. This is done by verifying that in the class of functions $C(X)$ and the set B, that any two parameters which are different from the true parameters in the sense of the norm, give rise to different probability models in a sense that will be defined below. Then it will be shown that this implies condition (d) of Theorem 1. Instead of examining the narrow class of function Θ^*, we concentrate here on the class $C(X)$ which contains Θ^*. Identification in $C(X)$ will imply identification in Θ^*. Note that two elements are considered different if the norm of their difference is nonzero.

We make use of the following definition of identifiable uniqueness of a probability model in a set of semi-non-parametric probability models. It is a generalisation of the definition typically used in the literature (see Wald (1949) Assumption 4 for example).

Definition: *The parameter vector (β^*, σ^*) in $B \times C(X)$ is identifiably unique in $B \times C(X)$ if for any (β, σ) not equal to (β^*, σ^*) there is a set D in X with $\mu(D) > 0$ and a set E in the support of z, such that $Pr(z \in E|x; \beta, \sigma) \neq Pr(z \in E|x; \beta^*, \sigma^*)$ for all x in D.*

Note that E must have positive probability under at least one of the two parameter vectors. In examining whether the models satisfy this, we make a number of assumptions which are typical of this type of work.

Assumption 3: *The probability measure μ on X has the property that at least one co-ordinate of x has everywhere (on its domain in $(0, 2\pi)$) positive Lebesgue density conditional on the other components.*

This assumption is common in work on nonparametric approaches to estimation and identification (see Matzkin (1989) for example). It allows one to obtain an open ball around a given point in X which has positive probability with respect to μ. and on

which the densities of z differ for some nonnull set in the support of z.

Assumption 4: *For the Probit model and the Tobit model σ is in $C(X)$. For the sample selection model the functions q_1 and q_2 are in $C(X)$.*

Assumption 5: *x in R^k satisfies a full rank condition that $\mu(x : x(\beta - \beta^*) = 0) = 1$ implies that $\beta = \beta^*$.*

Assumption 6: *In the Probit model and the dichotomous part of the sample selection model (when only the variable $I(y \geq 0)$ is observed) it is known that the coefficients in the corresponding equation satisfy $\|\beta\|_E = 1$.*

Assumption 4 is a minimal smoothness condition which is satisfied by the set of functions of interest, discussed in the previous sections. Assumption 5 rules out perfect multicollinearity among the x variables and allows identification of β given that $x\beta$ is identified. Assumption 6 takes care of the fact that when the discrete variable is observed the contribution to the likelihood function is invariant to monotonic transformations to the β and the σ coefficients. Since it can be shown that $x\beta/\sigma$ is identified and hence the sign of $x\beta$ is identified, we can then identify the β up to scale. This is similar to the situation of Manski (1975, 1985). To be able to obtain similar results we make the following assumption where we have x partitioned as $x = (x_k, x_l)$ with x_k being the continuously distributed variable, and the domain of x_k being (e, f) (which is strictly contained in $(0, 2\pi)$).

Assumption 7: *It is known that in the equation for the variable y for which $I(y \geq 0)$ is observed, the β corresponding to the continuously distributed x, β_k^*, is nonzero. In addition if $\beta_k^* > 0$ then*

$$e\beta_k^* + x_l\beta_l^* < 0 < f\beta_k^* + x_l\beta_l^*$$

for a (measurable) set of $x_l \in C$ with $P(C) > 0$ and $P(x_k \in (c, d)|C) > 0$ (where these probability measures are derived from the measure μ) for any $(c, d) \subset (e, f)$. If $\beta_k^* < 0$

then the inequality above is reversed.

This provides a sufficient condition for the identification of the parameters of the equation generating the dichotomous variable. They are slightly stronger than those of Manski (1985) and are necessitated by the fact that in our case the x variables are bounded, which implies that the values of $x\beta$ will be bounded. Hence, more stringent restrictions on the β vector are required. Essentially, the above assumption guarantees that there is a non-trivial subset of X on which the sign of the $x\beta$ changes. This then allows identification of the β^* parameters. The simplest example of the type of restrictions being imposed is the case where x_l is a constant. Then we require that the intercept term and the slope coefficient β_k^* differ in sign (assuming that the x variables have been scaled to be positive, as is required for use of the Fourier series approximation) and be of magnitudes which allow the above inequality to hold.

Lemma 5: *Given Assumptions 3-7 the parameter vector* (β^*, σ^*) *is identifiably unique for the Probit model.*

For the Tobit model we are able to identify β and σ without the use of normalisations. This is the same as the usual case without heteroskedasticity. It results because at least for some observations we observe the value of y. Moreover, for any x value there is a positive probability that y is nonzero. Identification then follows from the properties of the normal density and the continuity built into the model.

Lemma 6: *Under Assumptions 3-5* (β^*, σ^*) *is identifiably unique for the Tobit model.*

The sample selection model contains elements of both the Probit and Tobit models and hence it is unsurprising that β_1 is only identified up to scale. It is important, however, to note that ρ_{12} is identified although σ_1 and σ_2 are not, their influences on the density being mixed into the functions q_1 and q_2 which are identified up to scale.

Lemma 7: *Given Assumptions 3-7* $(\beta_1^*, \sigma_1^* q_1^*, \beta_2^*, \sigma_2^* q_2^*, \rho_{12}^*)$ *is identifiably unique.*

We can now verify condition (d) of Theorem 1. To do this we use the approach of Wald (1949) which exploits Jensen's inequality, identifiable uniqueness, and the integrability of functions of x. We first note that the densities in all cases may be written as

$$s(u, x, \beta^0, \sigma^0, \beta^*) = -\log p(z|x, \beta^0, \sigma^0)$$

and that

$$\int_X \int_{R^2} s(u, x, \beta^0, \sigma^0, \beta^*) \frac{1}{\sigma^*(x)} \phi(\frac{u}{\sigma^*(x)}) du d\mu(x) =$$
$$\int_X \int_Z -\log p(z|x, \beta^0, \sigma^0) p(z|x, \beta, \sigma) d\nu(z) d\mu(x)$$

for any function h, where $\nu(z)$ is the measure with respect to which

$$p(z|x, \beta^0, \sigma^0)$$

is a density. (To get this we use a change of variables argument.) For example, in the Probit model

$$\nu(z) = \delta_0(z) + \delta_1(z)$$

where δ_i gives measure to the set where $z = i$ and in this case $Z = \{0, 1\}$. Using this notation, we have that for all three models the following lemma holds.

Lemma 8: *The identification condition (d) of Theorem 1 is satisfied in all three models.*

We have now verified that the four conditions of Theorem 1 are satisfied by the three models. Although in this section we only used $C(X)$, it should be noted that this will imply identification in the actual function space used. This is because the function space (Σ in Theorem 1 and Θ^* in Section 4) is contained in $C(X)$.

3.7 Consistency and Practical Implications

We can now state the main results of the chapter relating to consistency of the maximum likelihood estimators for the three models. These are presented below.

Proposition 1: *Given Assumptions 1-7, the maximum likelihood estimators $\hat{\beta}$ and $\hat{\sigma}$ of the Probit model converge almost surely to the true values.*

Proposition 2: *Given Assumptions 1-5 the MLE of the Tobit model converges almost surely to the true values.*

Proposition 3: *Given Assumptions 1-7 the MLE of the sample selection model converges almost surely to the true values.*

These propositions are direct applications of Theorem 1 where the preceding lemmata have verified the conditions of the Theorem.

In practice, we could face a number of problems. The first is related to the use of the FFF. As can be seen from the representation given earlier, there are likely to be a large number of parameters especially if x contains a large number of variables. This could lead to difficulties when trying to maximise the likelihood function. In principle, the maximisation can be done, but in practice difficulties may arise. One solution may be to limit the number of x variables appearing in the function. For example, if one had prior information on the variables most likely to be related to heterogeneity then this could be used. One such example could be that in the case of durable expenditure one would expect there to be more expenditure variability at the high income levels. Information of this sort may be useful in simplifying the estimation.

A second problem concerns the imposition of the non-negativity constraint on the variance approximations. As the method has been presented this is likely to be nontrivial. One way to get around this problem is to approximate some transformation of the variance, with the transformation ensuring that the variance is nonnegative. For example if one notes that

$$g(x) = \exp(\log(g(x)))$$

it would be possible to impose non-negativity on g by approximating $\log(g(x))$ and

ensuring that this is bounded away from $-\infty$ and ∞. This will keep g bounded away from 0 and ∞ as is desired. All of the previous analysis can be applied to the parameter defined by $\log \sigma(x)$. Since exp is a uniformly continuous function on bounded intervals all the identification results will trivially be satisfied.

It was noted earlier that we have assumed that X is a bounded open set. Amemiya (1973) and Dhrymes (1987) both prove consistency in the iid case for models of the type discussed in this chapter under the assumption of bounded x variables. In our case, this was necessitated by the fact that we chose to use the FFF. In practice, this is not likely to cause problems since all actual data sets can be contained in some bounded open set, and hence can be scaled so that all x variables are in $(0, 2\pi)$. An alternative to this boundedness assumption would be to use an alternative type of approximating function and some other metric for measuring the goodness of the approximation to the true function. This would require the imposition of restrictions on the moments of the x variables, and the integrability of certain functions of x. Moreover, the type of convergence one could achieve would likely be weaker than the intuitively appealing uniform convergence achieved in this chapter. Further work could examine the possibility of using other conditions and assumptions.

3.8 Conclusions

This chapter has considered the consistent estimation of heteroskedastic LDV models with particular reference to the Probit, Tobit, and sample selection models. Although interest in these models is typically centered on the parameters of the means of the latent variables (the β's), it was noted that mis-specification of the variance parameters, and in particular assuming homoskedastic latent residuals when there is in fact heteroskedasticity, will lead to inconsistent estimates of the parameters of interest (unlike the case of

the linear regression model). The method proposed in this chapter allows one to obtain consistent estimates of both the variance parameters and the parameters of interest. The basic idea was to approximate the unknown variance functions using a Fourier series type approximation and to let the number of terms in the approximation increase with the sample size. Given certain smoothness conditions on the variance functions, the approximation becomes capable of being arbitrarily good (in the sense defined above). In addition, and of more interest, the corresponding estimates of the β's are consistent.

Since the method of estimation is based on maximum likelihood, the estimation should be fairly straightforward to perform in practice and can be done using any of the standard optimisation algorithms. The only difference lies in the fact that one must optimise with respect to both the β's and the parameters of the approximation to the variance parameter. This makes the method attractive relative to the computationally burdensome Maximum Score and Least Absolute Deviations estimators for the discrete choice and censored regression models respectively. The other advantage of the method is that it is more generally applicable unlike the techniques mentioned above which are based on quantile restrictions and do not generalise in an obvious way to multivariate contexts. In particular, for the sample selection model or any other bivariate model, no other technique is as yet known which is consistent in the presence of heteroskedasticity of an unknown form (at least without substantial identification assumptions as would be required in a semi-parametric regression type technique). One could conjecture that the method may also be used in other multivariate models such as those listed in Amemiya (1985).

This chapter has been concerned only with consistency. Therefore, a useful area for further work is the asymptotic distribution theory for the estimator. There are two possible avenues to explore in this context. The first is to view the approximation as a parametric assumption and to consider the problem of making inferences on the resulting

Quasi MLE. In this case, the methods of White (1982) and others may be of use. The second approach is to attempt to find the distribution of the $\hat{\beta}$ and or $\hat{\rho}_{12}$ (note that we are most interested in inferences on β and ρ_{12} rather than the σ terms) using recent results on asymptotic distribution theory for these types of estimators (see Andrews (1989) for example).

Chapter 4

Two Step Estimation Methods

4.1 Introduction

In this chapter we consider the problem of using two-step methods to estimate the sample selection and Tobit models when there is heteroskedasticity of an unknown form in the latent residuals. Such methods are quite common in practice since they are somewhat less burdensome computationally than the maximum likelihood estimator (MLE). These methods were first suggested by Heckman (1976, 1979), for these models and one rarely finds examples where the maximum likelihood estimator has been computed (see Donald (1985)). Another reason for considering these methods is that the regression structure they possess is easier to handle than the MLE, which is usually defined as the implicit solution of a set of nonlinear equations, so that when one is concerned with the distribution of the estimators the analysis may be somewhat simpler. In addition, the analysis may aid in the development of an asymptotic normality result for the MLEs considered in the preceding chapter.

The work in this chapter differs in an important way from the work of Newey (1988), Powell (1988), and others who have considered two-step estimators which are consistent and asymptotically normal when the latent residuals are independent and identically distributed (iid) but with unknown density function. The models in this case have a double index form which in general will be lost if there is heteroskedasticity which is not

a function of the indexes. In fact, the robustness of these methods to the heteroskedasticty is unknown and is an important topic for future research since most data used to estimate these models are heteroskedastic in nature (being typically large cross sections with heterogeneity a common problem — see Phillips (1988)). Also, in a recent paper Newey, Powell, and Walker (1990) have estimated the Heckman (1974) model using these methods and found that the results do not differ much from the results obtained under Gaussian assumptions, so at least for that particular problem distibutional assumptions regarding the shape do not seem to matter too much. They argue that other forms of mis-specification may be more important (at least with regard to that model and data). Heteroskedasticity is one form of mis-specification that has been relatively neglected, and may actually be important. In addition, the results of this chapter cover the case of the Tobit model and so offer an alternative estimator to the censored least absolute deviations and symmetrically censored least squares estimators of Powell (1984, 1986).

This chapter first considers the nature of the problem we are dealing with and, in particular, shows that the conditional means which are typically considered bear some resemblance to some other models that are currently being used in econometrics, namely semiparametric regression models. We then go on to present a two-stage estimation procedure similar to the standard Heckman (1979) procedure, but which allows for fairly general forms of heteroskedasticity in the latent error terms. Under the various quite general assumptions presented, the estimator is shown to be consistent and asymptotically normally distributed with the usual rate of convergence.

The semiparametric method we use is based on series type nonparametric estimators which are quite easy to use and quite generally applicable. In particular, since the conditional mean regression function is likely to have heteroskedastic error terms, as noted in Robinson (1988), (and since we would like to be able to cope with heterogeneous observations in any case) we would like a nonparametric technique which copes with quite

general heteroskedasticity. The series approaches cope with this situation quite easily whereas other techniques such as kernel regressions appear to not be able to cope (see White and Stinchcombe (1989)). For this reason the results of Robinson (1988) are not directly applicable since he maintains the assumption of independent and identically distributed (i.i.d) errors (although he claims that heteroskedasticity can be dealt with using his technique). For this reason we first provide results analogous to those of Robinson (1988) but using series estimators and then go on to adapt this framework to the problem at hand. The main complication that arises and which causes the model to differ from a pure semiparametric regression model is due to the fact that the usual inverse Mills ratio is estimated (consistently and asymptotically normally) so there will be some estimation error involved which will complicate the nonparametric estimation and the covariance matrix. This chapter deals with these by the application of a result on the $\sqrt{N}$ asymptotic normality of averages of nonparametric point estimates.

Obtaining a consistent and asymptotically normal estimator with the usual rate of convergence ($\sqrt{N}$) is quite useful. First, if one wishes to obtain an estimator that is best among all estimators which allow for arbitrary heteroskedasticity, then one would like to at least know that an estimator exists that is consistent at this rate so that it must then follow that the best possible estimator must also be consistent at this rate. That is, the semiparametric efficiency bound is surely smaller than the variance-covariance matrix for the estimator obtained in this chapter. Second, the result allows one to provide a basis for a specification test for heteroskedasticity in these types of models which can be based on the difference between the estimator obtained under homoskedasticity and that obtained under general heteroskedasticity in this chapter measured in the metric provided by the variance-covariance matrix of this difference. Without both estimators being asymptotically normal at the same rate it is difficult to see how such a test could be constructed.

The chapter is structured as follows. Section 2 discusses the model of interest and shows how it is related to a semiparametric regression model and also what complicates it relative to a semiparametric regression model. Some issues regarding identification of the model are also discussed. Section 3 presents in fairly general terms the type of nonparametric regression technique to be used throughout the chapter and presents a few results regarding convergence rates of the mean squared error and asymptotic normality of weighted averages. Section 4 presents results analogous to Robinson (1988) but allows for heteroskedasticity and uses the series regression estimator. This allows the derivation of a $\sqrt{N}$ consistent asymptotically normal estimator in the unrealistic situation that the inverse Mills ratio were known. Section 5 discusses two possible ways of estimating the discrete choice model depending on the assumptions one makes regarding the error term in the discrete part of the model (which will only be relevant in the case of the sample selection model) and of obtaining an estimator of the inverse mills ratio. A linearisation analogous to the δ method is used to put the estimator in a form that can be handled when the distribution of the second stage estimator is examined. Section 6 presents the main result regarding $\sqrt{N}$ consistent and asymptotically normal (RNCAN) estimation of the full two stage estimator. Section 7 discusses the estimation of the variance-covariance matrix of the estimator. Section 8 presents a Hausman (1978) type specification test for heteroskedasticity which is shown to be asymptotically chi-squared. Section 9 concludes the chapter.

4.2 The Model

The type of model we consider involves a bivariate normal density function and the two latent variables given below

$$y_{1t}^{*} = x_{1t}\beta_1 + u_{1t} \tag{4.2}$$

$$y_{2t}^{*} = x_{2t}\beta_2 + u_{2t}. \tag{4.3}$$

for $t = 1, 2...N$. The observed endogenous variables in this case are

$$y_1 = I(y_1^{*} \geq 0) \tag{4.4}$$

$$y_2 = y_1 y_2^{*}. \tag{4.5}$$

That is, we observe a discrete variable y_1, and conditional on this being unity we observe the value of the continuous latent variable y_2^{*}. An example of this situation would be that we observe whether or not someone works and then conditional on them working we observe the number of hours worked. Also we observe the x variables for all the observations.

Since we will be assuming that the observations are independent, we can order the data such that the first N_1 have $y_1 = 1$ and the last N_0 have $y_1 = 0$ and $N = N_1 + N_0$. The vector (u_1, u_2) has the bivariate normal density with covariance matrix Ω which can be written as follows

$$\Omega(x) = \begin{pmatrix} \sigma_1^2(x) & \sigma_{12}(x) \\ \sigma_{12}(x) & \sigma_2^2(x) \end{pmatrix} \tag{4.6}$$

where $x = (x_1, x_2)$. Note that if $\beta_1 = \beta_2$, $x_1 = x_2$ and $u_1 = u_2$ then the model reduces to the standard Tobit or censored model.

The estimation problem arises because unobservables determining the value of the discrete variable (*i.e.*, u_1), may be correlated with those determining the value of the continuous variable. Thus employing OLS to estimate β_2 using only observations for which $y_1 = 1$ may be inconsistent. This is best seen by taking conditional expectations of y_2 given that $y_1 = 1$.

$$E(y_{2t}|y_{1t} = 1) = x_{2t}\beta_2 + \frac{\sigma_{12}(x_t)}{\sigma_1(x_t)}\lambda\left(\frac{x_{1t}\beta_1}{\sigma_1(x_t)}\right) \tag{4.7}$$

where

$$\lambda\left(\frac{x_{1t}\beta_1}{\sigma_1(x_t)}\right) = \frac{\phi(\frac{x_{1t}\beta_1}{\sigma_1(x_t)})}{\Phi(\frac{x_{1t}\beta_1}{\sigma_1(x_t)})} \tag{4.8}$$

with ϕ and Φ being the standard normal density and distribution functions respectively. When the covariance term is nonzero, then the possible correlation between the second term and the first will in general lead to inconsistent estimates of β_2.

The above representation also provides motivation for a two-step estimation procedure where in the first stage we estimate β_1 by Probit maximum likelihood on the first equation in the two equation model and then use it to obtain a correction term $\hat{\lambda}$. Then the second stage consists of performing OLS for the N_1 observations with y_1 with the correction term included as a regressor. This is almost always done under the assumption that the covariance matrix is constant across all the observations. If this is false, then problems may arise both in the estimation of λ and in the second stage regression where the coefficient of λ may in fact be a function of x. In general, this will result in the estimator being inconsistent (the degree of this inconsistency has not to my knowledge been considered in the literature). This chapter will discuss a method for dealing with these sources of inconsistency.

First since the main concern here is in estimating the β_2 parameter vector, the argument of λ, $\frac{x_{1t}\beta_1}{\sigma_1(x_t)}$, can be treated as some unknown function $h(x_{1t})$. We are implicitly assuming that the variance of the first latent residual u_1 is a function of only those x variables appearing in the mean (*i.e.*, x_1). This assumption greatly simplifies notation and does not seem too extreme. Writing it in this way does not affect any of the main results of the chapter since for the asymptotics the whole function must be consistently estimated. Note also that we can write the function $\frac{\sigma_{12}(x)}{\sigma_1(x)}$ as $g(x)$. Using these adjustments the model can be written as

$$y_{2t} = x_{2t}\beta_2 + g(x_t)\lambda(h(x_{1t})) + \xi_t \tag{4.9}$$

where ξ_t is an heteroskedastic error term with mean zero. In fact, the variance of ξ_t is

given by

$$V(\xi_t|x_t) = \sigma_2^2(x_t) - \frac{\sigma_{12}^2(x_t)}{\sigma_1^2(x_t)}(x_{1t}\beta_1\lambda(h(x_{1t})) + \lambda(h(x_{1t}))^2). \tag{4.10}$$

The model above bears some resemblance to some models that are currently being examined in econometrics — namely, semiparametric regression models, see Robinson (1988). Suppose for the moment that the h function is known and is bounded on its domain so that λ is bounded away from 0 and ∞. Then one can transform the model into the semiparametric regression form given below

$$w_t = z_t\beta_2 + g(x_t) + \eta_t \tag{4.11}$$

where $w_t = y_t/\lambda_t$, $z_t = x_t/\lambda_t$, $\eta_t = \xi_t/\lambda_t$ and $\lambda_t = \lambda(h(x_{1t}))$. This is now in the familiar semiparametric regression form with g being an unknown function. So given that h is known we can apply recent results, for example Andrews (1988, 1989d), to obtain $\sqrt{N}$ consistent estimates of β_2 for this model. Note, however, that since the error term is heteroskedastic the results of Robinson (1988) are not directly applicable. Section 4 presents a method of dealing with this.

In practice, however, h will be unknown and must be estimated. The model will have the form

$$\hat{w}_t = \hat{z}_t\beta_2 + g(x_t) + \hat{\eta}_t + \vartheta_t \tag{4.12}$$

with ϑ_t being related to the estimation error inherent in estimating the λ_t, and where estimates are used in place of the true values. In analysing this model, we must also worry about the estimation error inherent in the estimation of the λ_t. In addition, one must worry about the identification of the model since the unrestricted nature of the g function may cause problems unless there are certain exclusion restrictions (referring to exclusion of certain of the x variables from the different functions involved).

4.2.1 Intuition Behind the Estimator

Before we discuss the intuition behind the estimation method to be used, we outline some of the basic assumptions on the model. For ease of exposition denote the vector of x variables included as arguments of the covariance matrix elements and hence arguments of g by x_3 and remember that λ_t is a function of x_{1t}. The complete set of exogenous variables will be referred to as x. Then, as noted above, we can write the conditional mean of y_{2t} as

$$y_{2t} = x_{2t}\beta_2 + g(x_{3t})\lambda_t + \xi_t \tag{4.13}$$

with $E(\xi_t) = 0$ and $V(\xi_t) = \sigma_t^2$ for all t and it will be assumed that the model is such that

$$0 < \inf_t \sigma_t^2 < \sup_t \sigma_t^2 < \infty \tag{4.14}$$

and

$$0 < \inf_t \lambda_t < \sup_t \lambda_t < \infty. \tag{4.15}$$

As mentioned above this allows us to transform the model so that

$$w_t = z_t\beta_2 + g(x_{3t}) + \eta_t \tag{4.16}$$

with the transformed residual, η_t, having properties similar to those of ξ_t. The population motivation for the type of estimator to be used in this chapter is the following. Take expectations of (4.16) conditional on x_{3t} which gives

$$E(w_t|x_{3t}) = E(z_t|x_{3t})\beta_2 + g(x_{3t}) \tag{4.17}$$

and subtract this from (4.16) to give

$$w_t - E(w_t|x_{3t}) = (z_t - E(z_t|x_{3t}))\beta_2 + \eta_t \tag{4.18}$$

Note that we are implicitly conditioning on the fact that $y_{1t} = 1$. Since this is common to all the observations, the inclusion of this as an argument is redundant provided that

x_3 contains a constant. This will be important further on in this chapter, so one should remember that y_{1t} is, in fact, an argument of these conditional expectations, but that it is omitted for notational simplicity. Note also that we are assuming that g is a common function for all the observations and that it does not depend on y_{1t}. This is no different from the usual case with homoskedasticity.

It would follow that if we knew the expectations in (4.16) then we could potentially estimate β_2. There is, however, the question of identification to be dealt with since if $z_t = E(z_t|x_{3t})$ then (4.18) would no longer contain information on β_2 due to the cancellation. To avoid problems of non-identification in this regression model we make the following assumptions

Assumption 2.1: *For each element of the vector z_t we have that $z_{it} \neq E(z_{it}|x_{3t})$.*

Since in our case we have that $z_t = \frac{x_{2t}}{\lambda(h(x_{1t}))}$ then we will require that x_1 contain at least one variable that is not contained in the vector x_3. Moreover, we require that this variable is not measurable with respect to the σ-field generated by x_3. If all the elements of the variance-covariance matrix are functions of x_3 then this would require that there be one variable in x_1 that is not in x_3. Our next assumption is a standard full rank assumption on the regressors. Define the matrix of the z_t vectors as Z and the matrix of conditional expectations as T_3.

Assumption 2.2: *Assume that*

$$\text{plim}\frac{1}{N_1}(Z - T_3)'(Z - T_3) = \bar{A}_1$$

which is a finite, positive definite matrix.

Note that we avoid complications such as in White (1984), where the limit may not exist. This simplifies the notation appreciably without diminishing the worth of the results. In addition to the preceding assumptions, we make the assumption of boundedness of all the exogenous variables. This is obviously not a desirable assumption, but is quite

common in theoretical work on limited dependent variables and in nonparametrics involving series type estimators (see White and Stinchcombe (1989), for example). Finally, it will be assumed that the errors possess sufficient moments, so we can apply a central limit theorem to the estimators considered in this chapter.

Assumption 2.3: *The data x_t are uniformly bounded independent random variables and any needed probability limits are assumed to exist. In addition*

$$E(\xi_t^4|x) < \Delta < \infty$$

for all t.

Note that this allows the data to be heterogeneous but does not allow any sort of dependence. This independence assumption may not be needed, as recent work by White and Stinchcombe (1989) has shown.

Define the population regression estimator of β_2 by

$$\bar{\beta}_2 = ((Z - T_3)'(Z - T_3))^{-1}(Z - T_3)'(W - E(W|x_3)). \tag{4.19}$$

We now can prove the following simple result on the asymptotic normality and consistency of the above estimator.

Theorem 1: *Given Assumptions 2.1-2.3 we have*

$$\sqrt{N_1}(\bar{\beta}_2 - \beta_2) \to N(0, \bar{V}_2)$$

where

$$\bar{V}_2 = (\bar{A}_1)^{-1}\text{plim}(1/N_1)E((Z - T_3)'D_3^{-1}\Gamma D_3^{-1}(Z - T_3))(\bar{A}_1)^{-1}.$$

Proof: For the proof of this and all other results in this chapter, see Appendix B.

The matrix Γ is the diagonal variance covariance matrix of the error terms ξ_t and the D_3 matrix is given by $D_3 = \text{diag}\lambda_t$. The result in Theorem 1 is not operational and the estimator $\bar{\beta}_2$ is infeasible, but it does provide motivation for the approach in this chapter which is essentially to replace the unknown expectations by consistent estimates.

4.3 Nonparametric Series Regression results

In this section, we examine the series approach to the estimation of conditional expectations as are used to make the estimator considered above a feasible estimator. In doing this it is useful to refer to a generic type of situation given by the following nonparametric regression model

$$y_t = g(x_t) + u_t \tag{4.20}$$

where g is the function to be estimated (note that these variables are not necessarily the same as those included in the preceding analysis). The error term has expectation zero conditional on x_t so that the regression function has the conditional mean interpretation. A series type estimator for g is obtained by replacing g with a truncated series or approximation to g which involves parameters and basis functions denoted by $\psi_i(x_t)$ which may be polynomials or trigonometric functions. The type of approximation that will be used is based on the Fourier Flexible Functional Form (FFF), developed by Gallant (1981). In that case the basis functions include quadratic terms involving x, and trigonometric terms of the form

$$\cos(jk'_\alpha x), \sin(jk'_\alpha x)$$

where the k_α are multi-indices, some of which are presented in Gallant (1981), and j are constant integers. The number of terms in the FFF is increased by adding more trigonometric terms. For a given number of terms, one is then working with following regression

$$y_t = \sum_{i=1}^{K} \theta_i \psi_i(x_t) + u_t + g(x_t) - \sum_{i=1}^{K} \theta_i \psi_i(x_t) \tag{4.21}$$

where the object is to estimate the θ_i parameters, K denoting the number of terms in the approximation. Note that there are now two sources of error — one is due to sampling error (the u_t) and the other stems from the fact that the approximation may

not necessarily equal the true function. Typically, one can eliminate this specification error asymptotically by choosing basis functions that are capable of approximating the unknown g function and by increasing the number of terms K with the sample size. Results on the consistency (see Gallant (1987)) and asymptotic normality (see Andrews (1988, 1989d)) are now available for these types of estimators. Our interest will be in considering the rates of convergence of MSE for these types of estimators in situations where the sampling error is possibly heteroskedastic, and also in the asymptotic normality of weighted averages of these estimates.

For this purpose the estimator for g can be written

$$\begin{aligned} \hat{g} &= \Psi(x)(\Psi(x)'\Psi(x))^{-1}\Psi(x)'y \\ &= P_x y \end{aligned}$$

where $\Psi(x)$ is a $N \times K(N)$ matrix of basis functions and $K(N)$ is the number of terms in the approximation and will depend on the sample size N. P_x will be used to denote the projection matrix formed by the basis functions of the FFF constructed using the vector x. When the inverse does not exist then some generalised inverse may be used or else some terms may be omitted. Interest here will centre on the behaviour of the $MSE(g, \hat{g})$ given by

$$MSE(g, \hat{g}) = E\frac{1}{N}\|g - P_x y\|^2 \tag{4.22}$$

where $\|.\|$ will refer to the Euclidean norm of the argument.

The following assumptions will be used:

Assumption 3.1: *For the regression equation (4.20) the observations are independent and*

$$E(u_t|x_t) = 0, E(u_t^2|x_t) = \sigma_t^2$$

where

$$0 < \inf_t \sigma_t^2 < \sup_t \sigma_t^2 < \infty.$$

Next define the following Sobolev norm (see Adams (1975)) of the function h by

$$\|h\|_{q,\infty,X} = \max_{|\lambda| \leq q} \sup_{x \in X} |D^\lambda h(x)| \tag{4.23}$$

where D denotes the generalised partial derivative as defined in Chapter 2 and the Sobolev smoothness index (see Andrews (1988)) of the function h by

$$S(h) = \max_q \{q : \|h\|_{q,\infty,X} < \infty\} \tag{4.24}$$

where X is the set to which x belongs and $X \in R^d$, where d is the dimension of the x_t vector. The following assumption concerns the possibility of approximating the g function of interest.

Assumption 3.2: *For any g there is a sequence*

$$\theta_m = (\theta_{m1}, .., \theta_{mm})' \in R^m$$

such that

$$m^\alpha \| \sum_{i=1}^m \theta_{mi} z_i(x) - g(x) \|_{0,\infty,X} \to 0$$

as $m \to \infty$ for all α such that $0 \leq \alpha < \frac{S(g)}{d}$.

This assumption says that it is possible to find a sequence of uniformly convergent approximating functions. Moreover, the approximation error goes to zero at a rate that is larger the larger is the smoothness of the g function relative to the number of variables in x.

There are a number of circumstances when this assumption will be satisfied. The one of interest to us is given below.

Assumption 3.3: *The density functions of the x_t is everywhere bounded below by some small positive number. In addition it is assumed that X is open and bounded and closure$(X) \subset \times_{i=1}^{d}(0, 2\pi)$, and the basis functions are given by the trigonometric functions as used in a fourier series.*

Note that any bounded set X can be made to satisfy this assumption by suitable linear transformations, without affecting, in any meaningful way, the results. Many of the results of this chapter will also hold for polynomial series estimators under a similar condition. Note that the results also cover the case where the x are fixed provided that the empirical distribution converges to one like that above. The following result which is similar to that of Andrews and Whang (1989) can now be proved where we condition on the x vector to find the MSE and determine its rate of convergence.

Lemma 1: *Given Assumptions 3.1 and 3.2 (i) If $\frac{K(N)}{N} \to 0$ and $K(N) \to \infty$ then $MSE(g, \hat{g}) \to 0$. (ii) If $K(N) = O(N^r)$ then $MSE(g, \hat{g}) = O(N^{-b})$ where $b = \min\{1 - r, 2\alpha r\}$ (iii) If $\frac{S(g)}{d} > \frac{1}{2q}$ and $K(N) = O(N^{q-\gamma})$ for $q \leq \frac{1}{2}$, and $0 < \gamma < q - \frac{d}{2S(g)}$, then $\sqrt{N}\| \sum_{i=1}^{K} \theta_i z_i(x_t) - g\| \to 0$ for some sequence of parameters, and if $q \leq \frac{1}{2}$, then in addition b in (ii) is such that $b > (1 - q)$.*

These results show that the rate of convergence depends on three parameters relating to the smoothness of the unknown function, the dimension of its argument and the rate of increase of the number of terms in the approximation. Generally, the rate increases when the function is smooth, when the dimension of x is reduced, and when K increases more slowly. The final result concerns bias reduction and will be crucial in the results of the chapter. In particular, it appears that $\sqrt{N}$ consistency will not obtain unless the bias can be reduced at faster than $\sqrt{N}$ (as one may anticipate). This is achieved by altering both $S(g)$ and K — the rate of increase on K used in (iii) will also be needed in the main results.

As it stands, the result does not allow for discrete x variables. However if the discrete

x variables have finite support then g may be written as a finite sum of g_i functions of continuous x variables, one for each point in the support of the discrete x's. Then the result would hold provided that each g_i function satisfied the conditions of the Lemma. The remainder of the chapter avoids this complication without loss of generality.

A second result that will prove useful in the remainder of this chapter concerns the asymptotic distribution of weighted averages of estimated g functions — *i.e.* $\frac{1}{N}c'\hat{g}$ for some vector of constants c. In particular it is possible to show that under certain conditions such averages are $\sqrt{N}$ consistent and asymptotically normal.

Theorem 2: *Under Assumptions 3.1-3.2 and (i)* $S(g) \geq d+1$, *(ii)* $K(N) = O(N^{1/2-\gamma})$, *and (iii)* $c_t = c(x_t)$ *a bounded sequence of known functions of* x_t *with* $S(c) > 0$ *we have*

$$\sqrt{N}(\frac{1}{N}c'P_x y - \frac{1}{N}c'g) = \frac{1}{\sqrt{N}}c'u + o_p(1)$$

and hence

$$\sqrt{N}(\frac{1}{N}c'P_x y - \frac{1}{N}c'g) \rightarrow N(0, lim\frac{1}{N}E(c'\Psi_u c))$$

where $\Psi_u = E(uu'|x)$ *is the diagonal covariance matrix of* u.

This result will prove useful in Section 6 when we deal with the estimation error inherent in using an estimator of the λ_t. It will allow us to obtain the usual $\sqrt{N}$ convergence even though the individual $\hat{g}$ (or in our case $\hat{\lambda}_t$) elements are less than $\sqrt{N}$ consistent (see Andrews (1988, 1989d)). This type of result has been proved for kernel type estimators by Rilstone (1989) and Powell, Stock and Stoker (1989). A simple corollary to the above result arises when the c vector is just a vector of ones and a constant is included in the series functions (as is usually done). Denoting such a vector by ι then since $P_x\iota = \iota$ (because P_x is a projection matrix), the result is immediate.

Corollary 1: *Given Assumptions 3.1-3.2 and hypotheses (i) and (ii) of Theorem 2*

then

$$\frac{1}{\sqrt{N}}(\iota' P_x y - \iota' g) \to N(0, \lim E(\frac{1}{N} tr\,\Psi))$$

Note that in both results the hypotheses (i) and (ii) are used to eliminate the bias at faster than $\sqrt{N}$ rate as in Lemma 1 (iii) shows. The third hypothesis ensures that $P_x c$ converges to c in MSE and uniformly. This is trivially true when c is, in fact, equal to one of the basis functions as in the case of ι as proved in Corollary 1. One could pursue this line of work further, and could in particular consider the estimation of average derivatives for the discrete choice model considered below in Section 5.

4.4 Estimation with known λ_t

The estimator of interest in this section is the same as $\bar{\beta}_2$, but with the unknown expectations replaced by estimates based on series regression discussed in the previous section. The analysis of this estimator is useful as an intermediate step toward the result for the estimator considered later, in which λ_t is estimated. In fact as will be shown, the completely general estimator can be split up into two parts. The first will have the same distribution as the estimator of this section, and the second will have a non-degenerate distribution that will depend on the preliminary estimator of λ_t.

Since the expectations in this case are conditional on x_3 we denote the relevant projection matrix by P_{x_3} and the truncation parameter by $K^3(N_1)$ or simply as K^3 where it will be understood that it increases with N and or N_1. The estimator can then be written as

$$\sqrt{N}_1(\hat{\beta}_2 - \beta_2) = \left[(\frac{1}{N_1})(Z - P_{x_3} Z)'(Z - P_{x_3} Z)\right]^{-1} \times \tag{4.25}$$

$$(\frac{1}{\sqrt{N_1}})(Z - P_{x_3} Z)'(g - P_{x_3} g + \eta - P_{x_3}\eta) \tag{4.26}$$

$$= \left[(\frac{1}{\sqrt{N_1}})(Z - P_{x_3}Z)'(Z - P_{x_3}Z) \right]^{-1} (\frac{1}{\sqrt{N_1}})(Z - P_{x_3}Z)'(g + \eta) \qquad (4.27)$$

using the expression for w and the fact that

$$(Z - P_{x_3}Z)'P_{x_3} = 0 \qquad (4.28)$$

since P_{x_3} is a projection matrix. The dimension of x_{jt} will be denoted d_j, so that the dimension of z_t is d_2 (in the case of the Tobit model $d_1 = d_2$) and each element of z_t will be referred to as z_{it}. The conditional expectations are denoted by

$$E(z_{it}|x_{3t}) = \tau_i(x_{3t}) \qquad (4.29)$$

and the matrix of these values by $T_3 = E(Z|x_3)$ which will be of the same dimension as Z. The symbol δ will denote a generic small strictly positive constant while Δ and C will be used as generic large finite positive constants.

The following assumptions will be required to prove that the estimator $\tilde{\beta}_2$ is RNCAN.

Assumption 4.1: Assume $E(z_{it}|x_{3t}) = \tau_i(x_{3t})$ *for each* i *with* $x_{3t} \in X^3 \subset R^{d_3}$ *and assume Assumption 3.3 applies to the distribution of* x_{3t} *and the set* X^3.

Assumption 4.2: *Defining* $u_{it} = z_{it} - \tau_i(x_{3t})$ *then* u_{it} *satisfies Assumption 3.1 for all* i.

We are now in a position to prove the main result of this section concerning the RNCAN of the estimator $\tilde{\beta}_2$.

Theorem 3: *Given Assumptions 2.1-2.3 and 4.1-4.2, if (i)* $\frac{S(g)}{d_3} > \frac{1}{2q}$ *for some* $q \leq 1$, *(ii)* $S(\tau_i) \geq 0$ *for all* i, *(iii)* $K^3(N_1) = O(N_1^{q-\gamma})$ *with* $0 < \gamma < q - \frac{d_3}{2S(g)}$ *then*

$$\sqrt{N_1}(\tilde{\beta}_2 - \beta_2) \rightarrow N(0, V_2)$$

where V_2 *is given in Theorem 1.*

Note that it is required that the unknown g function (which arises due to the heteroskedasticity in the latent model) be smooth and increasingly so the larger is the

dimension of its argument and the the slower the rate of increase of the K^3 parameter. This along with the third hypothesis guarantee that the bias term disappears at the required $\sqrt{N}$ rate as proved in Lemma 1. The τ_i functions related to the expectations are required only to possess one derivative in order for the bias in estimating them to disappear sufficiently fast as $K^3 \to \infty$.

Obviously, this result is generally inapplicable since λ_t is generally unknown, however it does provide a useful result en route to the more general result. Moreover, this result is of independent interest since it provides a proof of RNCAN for semiparametric regression models which allows for more general data generating processes than does Robinson (1988) (although he claims his result does generalise to this case). It also provides alternative conditions to those of Andrews (1989a) who uses empirical process methods to obtain a similar result under a different set of assumptions.

4.5 Estimation of the Discrete Choice Model

As noted above, the preceding analysis is generally inapplicable due to the fact that λ_t is generally unknown. This section details two approaches to the estimation of λ_t which are appropriate under different conditions, depending on whether one believes there is heteroskedasticity in the discrete part of the model or not. In the case of the Tobit model, both latent equations are the same so that only the method allowing for heteroskedasticity is of interest. Therefore, the method based on a homoskedastic discrete choice model assumption is only of interest in the context of the bivariate sample selection model.

The equation of interest is rewritten below

$$y_{1t}^{*} = x_{1t}\beta_1 + u_{1t} \tag{4.30}$$

where $E(u_{1t}) = 0$ and $V(u_{1t}) = \sigma_1^2(x_{1t})$ and u_{1t} are independent normal random variables.

The variable $y_{1t} = I(y_{1t}^{*} > 0)$ is all that is observed As noted earlier, λ_t is related to these variables by

$$\lambda_t = \frac{\phi(\frac{x_{1t}\beta_1}{\sigma_1})}{\Phi(\frac{x_{1t}\beta_1}{\sigma_1})} \tag{4.31}$$

so as in Heckman (1979) one can obtain a consistent estimator of λ_t by replacing the β_1 and σ_1 parameters with consistent estimators. There are two possibilities that will be considered here. The first is to assume that σ_1 is, in fact, a function of the x variables, remembering that we assumed that only the variables in x_1 may be included as arguments of the function. In this case, there is bivariate heteroskedasticity (this is the only interesting case for the Tobit model) so the problem can then be written as one of obtaining consistent estimates of $\frac{x_{1t}\beta_1}{\sigma_1(x_{1t})} = h(x_{1t})$ where the structure of h can be ignored without affecting the results (the convergence properties of an estimator of h will depend on both the estimators for β_1 and σ_1). Alternatively, one could treat σ_1 as being a constant and then without loss of generality one could normalise it to 1 and then use standard Probit maximum likelihood estimation. It should be re-emphasised that this second case is only of interest in the case of the bivariate sample selection model.

4.5.1 Heteroskedastic u_{1t}

The work of Heckman (1979) and the survey of Amemiya (1985) are useful in pointing out that the use of an estimated λ_t in a second stage regression will have some affect on the variance-covariance matrix of the parameters due to the pre-estimation error. In fact, the distribution of the estimator will depend on the distribution of the λ_t estimator. This makes it imperative that we choose a method for estimating the λ_t that allows us to get at the distribution of the estimator. Since in the form in which the estimator appears the estimated λ_t enter as a weighted linear average the results of section 3 will be of use with some adjustments. (Note that we are excluding the possibility of finding a super-efficient

estimator for λ_t in which case the pre-estimation error would disappear at a sufficiently fast rate so that the distribution of the second stage estimator would not depend on the distribution of the first stage estimator.) Unfortunately, as it stands the work in Chapter 3 is inapplicable due to the difficulty in obtaining appropriate expansions and distribution theory. In contrast, it appears that the regression based methods discussed in Section 3 may be more useful with some adjustment. In fact, we will show that the use of the δ method along with the results for series regression estimators allow us to obtain a consistent estimator for λ_t and an expansion or linearisation that will help in the derivation of the distribution of the second stage estimator.

To see why a regression based technique may work, consider the following expectation

$$E(y_{1t}|x_{1t}, x_{2t}, x_{3t}) = E(y_{1t}|x_{1t}) = \Phi(h(x_{1t})) \tag{4.32}$$

so that y_{1t} can be written in a regression form as

$$y_{1t} = \Phi(h(x_{1t})) + \epsilon_{1t} \tag{4.33}$$

with $E(\epsilon_{1t}|x_{1t}) = 0$ and $V(\epsilon_{1t}|x_{1t}) = \Phi(h(x_{1t}))(1 - \Phi(h(x_{1t})))$ which has a familiar regression type structure. Since h is an unknown function we can let $l(x_{1t}) = \Phi(h(x_{1t}))$ and then the regression equation can be written as

$$y_{1t} = l(x_{1t}) + \epsilon_{1t} \tag{4.34}$$

and then h can be recovered by the inverse transform $h = \Phi^{-1}(l)$ using the inverse function of the normal cumulative distribution function. The following assumptions regarding h and l will be made

Assumption 5.1: *$h : X^1 \to [-\Delta, \Delta]$ and hence $l : X^1 \to [\delta, 1-\delta]$ where $\delta = \Phi(-\Delta)$ and $1 - \delta = \Phi(\Delta)$ for some small $\delta > 0$. Also X^1 and the distribution of the x_1 satisfies Assumption 3.3.*

Again we rule out discrete exogenous variables in this but as mentioned earlier they can be dealt with without much trouble. Note that the conditional probabilities are kept away from 0 and 1. This also ensures that the conditional variances are bounded away from ∞ and 0 as is required for the results of Section 3 to be useful. These are required to avoid problems when we transform the variables in the second stage regression by division.

The technique used to estimate λ_t will be based first on estimation of the heteroskedastic regression model (4.34) which gives us estimates of $l_t = l(x_{1t})$ which are then are used to construct the λ_t using the fact that l and λ_t are related by

$$\lambda_t = \frac{\phi(\Phi^{-1}(l_t))}{l_t} \tag{4.35}$$

The estimator for λ_t is obtained by replacing the l_t in this relation with the $\hat{l}_t$ obtained by the series regression of y_{1t} on x_{1t}. Two important facts about the transformation are that the first two derivatives with respect to the l_t given by

$$\frac{d\lambda}{dl} = -(\phi(\Phi^{-1}(l)) + l\Phi^{-1}(l))l^{-2} = d_1(l) \tag{4.36}$$

$$\frac{d^2\lambda}{dl^2} = l^{-3}(\phi(\Phi^{-1}(l)) + l\Phi^{-1}(l)) - l^{-2}(\frac{l}{\phi(\Phi^{-1}(l))}) = d_2(l) \tag{4.37}$$

are both bounded functions of l on $[\delta, 1-\delta]$ for any small δ. This will prove useful in the expansions of $\hat{\lambda}_t$ about its true value λ_t. Note also that λ_t itself is bounded away from 0 and ∞

The estimator for l that we use is given by

$$\hat{l} = P_{x_1} y_1 \tag{4.38}$$

where P_{x_1} is the projection matrix formed by the the basis functions of x_1 as described in section 3, analogously to P_{x_3} and the number of terms in the approximation or the truncation parameter is denoted by $K^1(N)$ or K^1. Note that all the N observations are

used in the computation of the l unlike the second stage estimator which uses only the N_1 observations for which $y_{1t} = 1$. One complication that arises when one computes the $\hat{\lambda}_t$ estimator is that it is required that one compute $\Phi^{-1}(\hat{l})$ which is only defined if the $\hat{l}$ is between 0 and 1. Unfortunately, the use of the series regression estimator does not guarantee this so one must have some rule for dealing with such observations. Note that this analogous to the problem in the linear regression model with heteroskedasticity where one estimates the variance function using the squared residuals and performs weighted least squares using the estimated residual squared regression function for weights (as in Cragg (1988) for example). Depending on the method used, one can obtain negative predicted variances. These are usually dealt with by setting them to some small positive value. The solution proposed here will be similar. In particular the following rule is used to define the estimator $\hat{l}(x_{1t})$

$$\hat{l}(x_{1t}) = \begin{cases} \hat{l}(x_{1t}) & \text{if } \delta/2 < \hat{l}(x_{1t}) < 1 - \delta/2 \\ \delta/2 & \text{if } \hat{l}(x_1 t) < \delta/2 \\ 1 - \delta/2 & \text{if } \hat{l}(x_1 t) > 1 - \delta/2 \end{cases}$$

where the δ is as defined as in Assumption 5.1. In practice, the choice of δ is up to the discretion of the econometrician and typically some small number will be chosen.

The following lemma allows us to write linear combinations of $\hat{\lambda}_t - \lambda_t$ as linear combinations of $\hat{l}(x_{1t}) - l(x_{1t})$ plus a stochastic error that converges to zero as $N \to \infty$. Since such linear combinations of $\hat{\lambda}_t - \lambda_t$ appear in the the second stage estimator and since Lemma 2 allows us to find the distribution of linear combinations of $\hat{l}(x_{1t}) - l(x_{1t})$ this will allow us to show how the distribution of the second stage estimator depends on the distribution of this first stage estimator.

Lemma 2: *Given Assumption 5.2 and (i)* $S(l) \geq d_1 + 1$, *(ii)* $K^1(N) = O(N^{1/2-\gamma})$ *for*

$0 < \gamma < \frac{1}{2} - \frac{d_1}{2S(l)}$, *and (iii)* $c(x_{1t})$ *a sequence of bounded constants then*

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} c(x_{1t})(\hat{\lambda}_t - \lambda_t) =$$

$$\frac{1}{\sqrt{N}} \sum_{t=1}^{N} c(x_{1t}) d_1(l(x_{1t}))(\hat{l}(x_{1t}) - l(x_{1t})) + o_p(1)$$

Note that in this result all N observations are used. In the application of this result, this will also be the case so some condition on the relation between N_1 and N will be needed. The hypotheses of the result are slightly stronger than in previous results since in this case we need the MSE of the $\hat{l}$ to be $o_p(N^{-1/2})$ for the result to follow. Combined with Lemma 2 this result allows us to prove the following result concerning the RNCAN of weighted averages of $\hat{\lambda}_t - \lambda_t$ and as noted earlier is one of the key steps in proving the RNCAN of the second stage estimator.

Lemma 3: *Given Assumption 5.1* $S(c) > 0$ *and conditions (i),(ii) and (iii) of Lemma 2 then*

$$\sqrt{N}(\frac{1}{N} c'\hat{\lambda} - \frac{1}{N} c'\lambda) = \frac{1}{\sqrt{N}} c' D_1 \epsilon + o_p(1)$$

where $D_1 = \text{diag}\{d_1(l(x_{1t}))\}$ *and hence*

$$\sqrt{N}(\frac{1}{N} c'\hat{\lambda} - \frac{1}{N} c'\lambda) \rightarrow N(0, limE(\frac{1}{N} c' D_1 \Omega D_1 c))$$

This result bears some resemblance to a δ method type of result. In fact, it is a combination of the δ method, which is applied to linearise $\hat{\lambda}_t - \lambda_t$ for each t, and the result in Lemma 2 on the distribution of weighted averages of point estimates, which proves RNCAN for combinations of point estimates which are less than RNCAN as noted earlier.

4.5.2 Homoskedastic u_{1t}

As mentioned earlier in this case due to the assumed homoskedastic normality of the latent error u_{1t}, one may proceed with the Probit MLE to estimate β_1 with $\sigma_1 = 1$ imposed as a normalisation. Then the estimator for λ_t may be computed as

$$\hat{\lambda}_t = \frac{\phi(x_{1t}\hat{\beta}_1)}{\Phi(x_{1t}\hat{\beta}_1)} \tag{4.39}$$

Using standard results, as in Amemiya (1985) for example, one can then perform the same sort of linearisation of this about its true value so that

$$\hat{\lambda}_t - \lambda_t = \frac{\partial \lambda_t}{\partial \beta_1}(\hat{\beta}_1 - \beta_1) + O_p(N^{-1}) \tag{4.40}$$

which can be written in vector notation as

$$\hat{\lambda} - \lambda = -D_4 X_1(\hat{\beta}_1 - \beta_1) + O_p(N^{-1}) \tag{4.41}$$

where $D_4 = \text{diag}(x_{1t}\beta_1\lambda_t + \lambda_t^2)$ and X_1 is the $N \times d_1$ matrix of x_{1t} variables which appear after evaluating the derivative. Standard mean value expansions for the MLE $\hat{\beta}_1$ as in Amemiya (1985) (which are valid given our assumptions) yield

$$\sqrt{N}(\hat{\beta}_1 - \beta_1) \rightarrow N(0, \text{plim} N(X_1' D_2 X_1)^{-1}) \tag{4.42}$$

where

$$D_2 = \text{diag}(\Phi(x_{1t}\beta_1)^{-1}(1 - \Phi(x_{1t}\beta_1))^{-1}\phi(x_{1t}\beta_1)^2) \tag{4.43}$$

Using this notation, it then follows that

$$\frac{1}{\sqrt{N}}(c'\hat{\lambda} - c'\lambda) \rightarrow N(0, \text{plim}\frac{1}{N} c' D_3 X_1 (X_1' D_2 X_1)^{-1} X' D_3 c) \tag{4.44}$$

which is the usable equivalent to the result provided in Lemma 5. Much of this is standard and found in Amemiya (1985) or any other text dealing with this problem. The

use of the standard MLE Probit estimator avoids the need for trimming or truncating the observations in practice as is needed in the heteroskedastic case. An interesting question to consider in the finite sample analysis (using simulations) would be whether the effect of the trimming is such that even in the presence of mis-specified variance (*i.e.*, assuming homoskedasticity when it is false) the second approach is superior. In the case of heteroskedastic regression models, Cragg (1989) has found that the use of variance estimators where trimming (or more correctly truncation) is required, that in finite samples one may obtain more efficient estimates using a mis-specified variance function than the true one. The situation here may differ somewhat, however, since in the regression model unbiasedness and consistency obtain no matter what weighting function is used, whereas here mis-specification causes inconsistency (although the nature of this is still not very well understood).

4.6 Asymptotic Normality with Estimated λ_t

The preceding results allow us now to derive the distribution of the estimator given in Section 4 where λ_t is replaced by either of the two estimators developed in Section 5. We denote this estimator by $\tilde{\beta}_2$ and note that it can be written as follows.

$$\tilde{\beta}_2 = ((\hat{Z} - P_{x_3}\hat{Z})'(\hat{Z} - P_{x_3}\hat{Z}))^{-1}(\hat{Z} - P_{x_3}\hat{Z})'(\hat{w} - P_{x_3}\hat{w}) \tag{4.45}$$

Noting that

$$\hat{w}_t = \hat{z}_t\beta_2 + g(x_{3t}) + g(x_{3t})\frac{(\lambda_t - \hat{\lambda}_t)}{\hat{\lambda}_t} + \frac{\xi_t}{\hat{\lambda}_t} \tag{4.46}$$

this estimator can be rewritten as

$$\tilde{\beta}_2 = ((\hat{Z} - P_{x_3}\hat{Z})'(\hat{Z} - P_{x_3}\hat{Z}))^{-1}\times \tag{4.47}$$

$$(\hat{Z} - P_{x_3}\hat{Z})'((\hat{Z} - P_{x_3}\hat{Z})\beta_2 + g - P_{x_3}g + g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} - P_{x_3}g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} + \frac{\xi}{\hat{\lambda}} - P_{x_3}\frac{\xi}{\hat{\lambda}})$$

where $g\frac{(\lambda-\hat{\lambda})}{\hat{\lambda}}$ is a vector of the elements $g(x_{3t})\frac{(\lambda_t-\hat{\lambda}_t)}{\hat{\lambda}_t}$ and $\frac{\xi}{\hat{\lambda}}$ is the vector of $\frac{\xi_t}{\hat{\lambda}_t}$. Simplifying and noting that

$$(Z - P_{x_3}Z)'P_{x_3} = 0 \tag{4.48}$$

due to the fact that P_{x_3} is a projection matrix, we have

$$\sqrt{N_1}(\tilde{\beta}_2 - \beta_2) = ((\hat{Z} - P_{x_3}\hat{Z})'(\hat{Z} - P_{x_3}\hat{Z}))^{-1}\times \tag{4.49}$$

$$(\hat{Z} - P_{x_3}\hat{Z})'(g + g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} + \frac{\xi}{\hat{\lambda}})$$

which is now in a form that can be analysed. It should be noted that this estimator can be split into two parts. The first part involves g and ξ and is similar to the estimator analysed in Section 4.4. The only difference involves the fact that $\hat{Z}$ is used in place of Z. It will be shown in Theorem 4 below that this term and the estimator in Section 4.4 are asymptotically equivalent. The second term involves the pre-estimation error $\hat{\lambda_t} - \lambda_t$. Using the linearisation results of Section 4.5, it will be shown in Theorem 4 that the term will have an asymptotic distribution similar to the one in Lemma 3. Finally, since the two terms will be shown to be asymptotically uncorrelated, the asymptotic distribution of the sum will be normal with covariance matrix given by the sum of the covariance matrices of the two terms involved.

One technical complication that arise when we consider the asymptotic distribution of the term involving the estimation error is that all N observations are used to estimate the λ_t. The problem arises because only the estimates of λ_t for observations with y_{1t} appear in the estimator. After linearisation of the term corresponding to the pre-estimation error, we will end up with a sum of the following form,

$$\sum_{t=1}^{N} y_{1t}c(x_{1t})(\hat{l_t} - l_t).$$

The problem arises because the constants that appear in the sum involve y_{1t} and the results of Lemma 2 and 3 require the constants to be smooth functions of only x_{1t}. The

results in those Lemmata are obtained because of the properties of the projection matrix formed by the basis functions (of x_{1t}) and the fact that c is a smooth function of x_{1t} (see the proof of Theorem 2). To make this type of result apply we must break up the constants into a part which is the conditional expectation of the constants (given x_{1t}) and another part with conditional mean zero, and then make assumptions regarding these terms so that the previous results may apply.

To do this we introduce the following notation. Define the following variable ν_{it} by

$$\nu_{it} = E((z_{it} - \tau_i(x_{3t}))y_{1t}|x_{1t}) \tag{4.50}$$

which will be defined for all N observations, and where $z_{it} - \tau_i(x_{3t})$ appears as part of the constant in the sum above. It should be noted that as mentioned earlier $\tau(x_{3t})$ is actually a function of y_{1t} but the argument is suppressed for convenience. Note that we have only considered estimating this for observations with $y_{1t} = 1$. Since the ν terms will appear in the covariance matrix it would appear that we would need estimates of τ for all the observations. Fortunately this is not the case because when we estimate ν by projecting $(z_{it} - \tau_i(x_{3t}))y_{1t}$ onto the space spanned by the basis functions of x_{1t} the terms corresponding to $y_{1t} = 0$ will be zero so that τ need not be estimated for these observations. This avoids a complication when we come to estimate the covariance matrix.

Assumption 6.1: *The error term defined by* $z_{it} - \tau_i(x_{3t}) - \nu_{it}$ *satisfies Assumption 3.3 and in addition* $\text{plim}\frac{1}{N}\nu'\nu$ *is a finite positive definite matrix.*

More useful notation includes the matrices

$$D_3 = \text{diag}(\lambda_t) \tag{4.51}$$

$$D_5 = \text{diag}(d_1(l(x_{1t}))) \tag{4.52}$$

$$L = \text{diag}(l(x_{1t})) \tag{4.53}$$

and

$$G = \text{diag}(g(x_{3t})) \tag{4.54}$$

The main result of this section is the following

Theorem 4: *Given Assumptions 2.1-2.3, 4.1-4.2, 5.1, 6.1, and, (i)* $K^1(N) = O(N^{1/2-\gamma_1})$ *for* $0 < \gamma_1 < \frac{1}{2} - \frac{d_1}{2S(l)}$ *and* $S(l) \geq d_1 + 1$, *(ii)* $K^3(N_1) = O(N_1^{1/2-\gamma_3})$ *for* $0 < \gamma_3 < \frac{1}{2} - \frac{d_3}{2S(g)}$ *and* $S(g) \geq d_3 + 1$, *(iii)* $S(\tau_i) \geq d_3 + 1$ *for all* i *(iv)* $N_1 = O(N)$, $\frac{N_1}{N} \to p > 0$ *and* $\frac{N_1}{N} \geq \delta > 0$ *(v)* $x_3 \subset x_1$, *a proper subset and* $E(z_i|x_1)$ *has strictly positive Sobolev smoothness index for all* i *(vi)* $S(\nu_i) > 0$ *for all* i *then*

$$\sqrt{N_1}(\tilde{\beta}_2 - \beta_2) \to N(0, V_2)$$

where $V_2 = \bar{A}_1^{-1}(S_1 + S_2)\bar{A}_1^{-1}$ *with*

$$S_1 = \text{plim}\frac{1}{N_1}(Z - T_3)'D_3^{-1}\Gamma D_3^{-1}(Z - T_3)$$

and

$$S_2 = \frac{1}{p}\text{plim}\frac{1}{N}\nu' G D_3^{-1} D_5 L \Omega L D_5 D_3^{-1} G \nu$$

This result provides the basis for performing linear and non-linear Wald tests on the parameters for both sample selection and Tobit models under heteroskedasticity One should note that in this result we have strengthened the smoothness conditions on the τ_i functions which are required for convergence due to the fact that only estimated λ_t values are used which potentially reduces the rate of convergence of the estimated τ_i functions to the true values. This needs to be increased for RNCAN to hold and the way this is achieved is by requiring a stronger smoothness condition. Assumption (v) is required for the application of Theorem 2 to be valid. It seems likely that this may be unnecessary, but the proof, which is already quite complicated, would become even more

complicated. The implication is that the variance-covariance matrix is a function of a subset of the variables in the first latent equation (the discrete part), and in that case it would just require that not all of those explanatory variables appear in the variance function. Conditions (i) and (ii) are slightly stronger than in Theorem 3 due to the fact that there is estimation error and a faster rate of convergence of the MSE of the estimated expectations is required. Condition (iv) ensures that the number of observations with $y_{1t}^{\sim}$ always makes up a nontrivial proportion of the sample and seems fairly mild, while (vi) allows the application of Lemma 5.

Although the result does offer some guidance concerning the rates of increase for the truncation parameters, the actual choice of these in practice is somewhat more difficult to advocate. The results may suggest a rule of thumb where say for 100 observations one would pick $K < 10$ and so on, however in practice there is no reason to suggest that such a rule will be optimal in any sense. The results of a simulation exercise may aid in this regard. Alternatively, one could use cross validation type procedures or some testing procedures. More on this will be discussed in the following chapter.

One advantage of the estimator is that it is relatively simple to compute, requiring only regressions and transformations. In fact, in some sense it may even appear to be simpler than the usual Heckman procedure which involves the first stage maximum likelihood Probit estimator. In principle, any regression package, such as SHAZAM, LIMDEP or TSP, should be able to compute the estimator while the use of matrix commands should enable the computation of the covariance matrix (more on this will be taken up in the following section).

A corollary to the above result arises under the assumption that the discrete part of the model is homoskedastic and then the Probit MLE can be computed in the first stage to compute the λ_t for the second stage. The result corresponding to Theorem 4 is given for this case.

Corollary 2: *Given Assumptions 2.1-2.3, 4.1-4.2, then (i)* $K^3(N_1) = O(N_1^{1/2-\gamma_3})$ *for* $0 < \gamma_3 < \frac{1}{2} - \frac{d_3}{2S(g)}$ *and* $S(g) \geq d_3 + 1$, *(ii)* $S(\tau_i) \geq d_3 + 1$ *for all* i, *(iii)* $N_1 = O(N)$, $\frac{N_1}{N} \to p > 0$ *and* $\frac{N_1}{N} \geq \delta > 0$ *(iv) the* u_{1t} *are homoskedastic*

then

$$\sqrt{N_1}(\tilde{\beta}_2 - \beta_2) \to N(0, V_2)$$

where $V_2 = \bar{A}_1^{-1}(S_1 + S_3)\bar{A}_1^{-1}$ *with*

$$S_1 = \text{plim}\frac{1}{N_1}(Z - T_3)'D_3^{-1}\Gamma D_3^{-1}(Z - T_3)$$

and

$$S_2 = \frac{1}{p}\text{plim}\frac{1}{N}(Z - T_3)'D_3X_1(X_1'D_2X_1)^{-1}X'D_3(Z - T_3)$$

The estimator in Corollary 2 is slightly more difficult to compute but not overly so and most statistical packages should be able to do it. Clearly, the only difference between the two estimators lies in the part of the variance-covariance matrix related to the first stage estimator. In the case of the Probit estimator, there is no need for trimming since all predicted probabilities will be restricted to lie between 0 and 1. It will be interesting to see in the simulations whether the trimming has an adverse effect on the first estimator and whether the second performs better even in the presence of mis-specification.

4.6.1 Estimating the Covariance Matrices

The only thing that remains to make the results in Theorem 3 and Corollary 2 operational is to find ways of estimating (consistently) the variance-covariance matrices. This is needed to construct Wald type tests of restrictions on the β_2 parameters. A natural thing to do is to replace the unknown values appearing in the $\bar{A}_1$, S_i matrices with consistent estimates and sample analogues. This requires a decision on how to estimate each unknown.

We first deal with S_1. In the case of the Γ matrix, a natural estimator is to take the squares of the estimated residuals given by

$$\hat{\xi}_t = y_{2t} - x_{2t}\tilde{\beta}_2 - \hat{g}(x_{3t})\hat{\lambda}_t \tag{4.55}$$

A natural estimator of Γ is

$$\hat{\Gamma} = \text{diag}\{\hat{\xi}_t^2\} \tag{4.56}$$

and given (4.16) a natural estimator of the vector g is

$$\hat{g} = P_{x3}(\dot{w} - \hat{Z}\tilde{\beta}_2) \tag{4.57}$$

while the obvious estimator for $Z - T_3$ is $\hat{Z} - P_{x_3}\hat{Z}$. Finally, D_3 can be estimated by $\hat{D}_3 = \text{diag}\{\hat{\lambda}_t\}$. Define the variance estimator so constructed by

$$\hat{S}_1 = \frac{1}{N_1}(\hat{Z} - P_{x_3}\hat{Z})'\hat{D}_3^{-1}\hat{\Gamma}\hat{D}_3^{-1}(\hat{Z} - P_{x_3}\hat{Z}) \tag{4.58}$$

Similarly, in the case of S_2 define

$$\hat{\Omega} = \text{diag}\{\hat{l}(x_{1t})(1 - \hat{l}(x_{1t}))\} \tag{4.59}$$

$$\hat{G} = \text{diag}\{\hat{g}(x_{3t})\} \tag{4.60}$$

$$\hat{D}_5 = \text{diag}\{d_1(\hat{l}(x_{1t})\} \tag{4.61}$$

$$\hat{L} = \text{diag}\{\hat{l}(x_{1t})\} \tag{4.62}$$

and the natural estimator of ν can be constructed by

$$\hat{\nu} = P_{x_1}((\hat{Z} - P_{x_3}\hat{Z})'0')' \tag{4.63}$$

where 0 here denotes a $N_0 \times d_2$ matrix of zeroes which arises due to the definition of ν_i and the fact that for these observations $y_{1t} = 0$. A natural estimator for S_2 is then given by

$$\hat{S}_2 = \frac{N}{N_1}\frac{1}{N}\hat{\nu}'\hat{G}\hat{D}_3^{-1}\hat{D}_5\hat{L}\hat{\Omega}\hat{L}\hat{D}_5\hat{D}_3^{-1}\hat{G}\hat{\nu} \tag{4.64}$$

Finally, define the estimator of $\bar{A}_1$ by

$$\hat{A}_1 = \frac{1}{N_1}(\hat{Z} - P_{x_3}\hat{Z})'(\hat{Z} - P_{x_3}\hat{Z}) \tag{4.65}$$

Note that G contains estimates of g for all the observations and for observations with $y_{1t} = 0$ one must take the estimated g function and recover the estimated parameters and evaluate the function at the values of x_{3t} for the observations. Defining the basis functions for these N_0 functions as Ψ_{30} and the basis functions for the other observations as Ψ_3 the estimated vector of g functions for these observations is given by

$$\hat{g} = \Psi_{30}(\Psi_3'\Psi_3)^{-1}\Psi_3'(\hat{w} - \hat{Z}\tilde{\beta}_2). \tag{4.66}$$

Unfortunately, due to technical difficulties it was necessary to alter some of the conditions of Theorem 4 to obtain consistency of these estimators. In particular, thus far we have avoided worrying about the rank of the matrix formed by the basis functions and just used a generalised inverse. To prove consistency of the covariance matrix estimator we need that the inverse exists. This necessitates an additional assumption and stronger smoothness conditions on the functions of interest. This assumption is based on the work of Andrews (1988, 1989d) and ensures that provided the number of parameters does not increase too fast (the explicit rate must be no faster than $N^{1/4}$) and that the distribution of the x satisfies the assumptions made earlier, then the matrix of basis functions will have full rank. In addition, a condition on the projection matrices which ensures that the diagonal elements go to zero will follow. The following assumption in conjunction with those made previously allows the application of Theorem B-1 of Andrews (1988) which in turn allows the circumvention of the technical difficulties.

Assumption 6.2: *The smallest eigenvalues of the matrices*

$$\frac{1}{N_1}E(\Psi_3'\Psi_3)$$

and

$$\frac{1}{N_1}E(\Psi_1'\Psi_1)$$

have strictly positive limit infimum.

This along with a slower rate of increase of the $K^i(N)$ parameters allows the use of the results of Andrews which gives the following fact that is used in the proof of the consistency of the covariance matrix estimators:

$$\max_{t\leq N}\psi_{it}'(\Psi_i'\Psi_i)^{-1}\psi_{it}\to 0 \tag{4.67}$$

for both $i = 1$ and 3. The following Theorem provides the necessary result for the consistency of the estimated covariance matrix estimator

Theorem 5: *Given Assumptions 2.1-2.3, 4.1-4.2, 5.1, 6.1, and 6.2 and given (i)* $K^1(N) = O(N^{1/4-\gamma_1})$ *for* $0 < \gamma_1 < \frac{1}{4} - \frac{d_1}{2S(l)}$ *and* $S(l) \geq 2d_1+1$, *(ii)* $K^3(N_1) = O(N_1^{1/4-\gamma_3})$ *for* $0 < \gamma_3 < \frac{1}{4} - \frac{d_3}{2S(g)}$ *and* $S(g) \geq 2d_3+1$, *(iii)* $S(\tau_i) \geq 2d_3+1$ *for all i (iv)* $N_1 = O(N)$, $\frac{N_1}{N} \to p > 0$ *and* $\frac{N_1}{N} \geq \delta > 0$ *(v)* $x_3 \subset x_1$, *a proper subset and* $E(z_i|x_1)$ *has strictly positive Sobolev smoothness index for all i (vi)* $S(\nu_i) \geq 2d_1 + 10$ *for all i then the result of Theorem 4 holds and*

$$\hat{S}_1 - S_1 = o_p(1)$$

$$\hat{S}_2 - S_2 = o_p(1)$$

$$\hat{A}_1 = \bar{A}_1 + o_p(1)$$

and hence the estimator of V_1 *defined by*

$$\hat{V}_1 = (\hat{A}_1)^{-1}(\hat{S}_1 + \hat{S}_2)(\hat{A}_1)^{-1}$$

satisfies $\hat{V}_1 = V_1 + o_p(1)$.

The work of Andrews (1988, 1989d) indicates that in the case of a polynomial series estimator a slower maximal rate of increase of the parameters may be necessary. Also in

the case of the FFF series approximation the fact that the functions are not orthonormal may cause problems with Assumption 6.2. Andrews suggests that some of the functions may have to be deleted to ensure that the condition holds. In practice, there is unlikely to be any problem of invertibility. The $N^{1/4}$ should therefore be interpreted as a maximal rate.

Note that in constructing the matrix Γ we use a similar estimator to that of White (1980) for heteroskedastic regression models. As in the regression case this estimator may be improved on by use of the jackknife estimator proposed by MacKinnon and White (1985). This does not affect the consistency of the estimator but is likely to improve the finite sample performance of the estimator.

The equivalent result for the model with homoskedastic discrete component is given in Corollary 3 below where the only difference lies in the estimation of the S_3 matrix. Define the estimator of the S_3 matrix by

$$\hat{S}_3 = \frac{N}{N_1}\frac{1}{N}(\hat{Z} - P_{x_3}\hat{Z}))'\hat{D}_3 X_1 (X_1'\hat{D}_2 X_1)^{-1} X_1'\hat{D}_3(\hat{Z} - P_{x_3}\hat{Z}) \tag{4.68}$$

and define

$$\hat{V}_2 = (\hat{A}_1)^{-1}(\hat{S}_1 + \hat{S}_3)(\hat{A}_1)^{-1} \tag{4.69}$$

Corollary 3: *Under the conditions of Theorem 5, and with the assumption of a homoskedastic discrete part, then the result of Corollary 2 holds and*

$$\hat{V}_2 = V_2 + o_p(1).$$

4.7 Specification Tests

This section considers two forms of specification tests based on the estimator developed in this chapter. Both tests are of the Durbin-Wu-Hausman form and compare the estimator

of this chapter to other estimators which are generally preferable under certain restrictions due to improved efficiency. The metric of comparison is the variance-covariance matrix of the difference of the estimators so the tests will have a χ^2 limiting distribution under the respective null hypotheses. The value in performing the test is that under the null the two estimators should have the same limit (the true value), while under the alternative the estimator developed in this chapter will converge to the true value while the alternative estimator will not in general. Thus, the tests may be considered consistency tests on the alternative estimators, where it is the case that the alternative estimator is that which is generally used in practice.

Since in our case the alternative estimator need not be the MLE (and may include the two-step estimation method of Heckman (1979) which is not efficient) the simplification of the variance-covariance matrix noted by Hausman (1978) is not generally applicable. This simplification relies on one estimator being asymptotically efficient (in the sense that it attains the Cramer-Rao lower bound), in which case Hausman (1978) has shown that it must be uncorrelated with any other consistent and asymptotically normal estimator. This greatly simplifies the computation of the covariance matrix of the difference between the two estimators. This will not generally be the case here, so that one must deal with the correlation between the two estimators in forming the variance-covariance matrix. This is not overly difficult, however, as can be seen by the various applications of this testing framework proposed by White (1980a), for example. In particular, if one can show that two estimators, say b_1 and b_2, satisfy

$$b_1 = W_1' u_1 + o_p(1)$$

$$b_2 = W_2' u_2 + o_p(1)$$

where W_i are non-random matrices, and u_i are random error terms, then the variance-covariance matrix of the difference between b_1 and b_2 can be found provided that one

can find the covariance between u_1 and u_2. This will be the case with the estimators considered in this chapter.

In addition, there may be a number of alternative estimators and the tests will be developed for the most commonly used ones only. The two forms of mis-specification dealt with are sample selectivity bias (which is only relevant in the case of the bivariate sample selection model) and a test for heteroskedasticity in the latent model (which can be performed in the case of both types of models considered in this chapter).

4.7.1 A Test for Sample Selectivity Bias

A common test in empirical applications of the bivariate sample selection model is for selectivity bias. In the homoskedastic case, this is done by performing a t-test on the estimated coefficient of $\hat{\lambda}_t$ in the second step regression. Since in this case the coefficient is $\frac{\sigma_{12}}{\sigma_2}$ the test corresponds to a null hypothesis that $\sigma_{12} = 0$, and this being the case one may use OLS without the correction term, since it will then yield consistent and efficient estimates of the β_2 parameters.

There are two reasons why this form of test (a Wald type test of $\sigma_{12} = 0$) will not be proposed. The first relates more generally to why such a test is of interest in any case. The second reason will have to do with the nature of this type of test in the heteroskedastic case.

In both the heteroskedastic and homoskedastic cases, the fact of $\sigma_{12} \neq 0$ is insufficient for the parameters β_2 to be biased by the neglect of the λ_t regressor in the second stage regression. In particular, if λ_t is uncorrelated with the other regressors then no bias will be evident, so if one is mainly interested in obtaining "good" estimates of β_2 then the OLS estimator may be preferred — even if there is a small correlation then OLS may even be preferable on a MSE criterion. Of course, if the main concern is to obtain estimates of the $E(y_2|y_1 = 1)$ then the correction term will be included in any case. This

suggests that the use of a Durbin-Wu-Hausman type test may be of use since it examines directly whether the exclusion of the correction term causes significant bias (or change) in the parameter estimates for the β_2 vector. Such a test would be based on testing the significance of the difference of the estimates obtained when the correction term is included and when it is not.

The second reason why a Wald type test may not be advisable has to do more specifically with the nature of the model in the heteroskedastic case. The null in this case would be $\sigma_{12}(x_3) = 0$ or in the context of the notation and form of the model as developed in this chapter, $H_0{:}g(x_3) = 0$. The null, therefore, requires that the function be identically 0. In the instance where we know the parametric form of g up to a finite number of parameters, one could write the null as a test involving restrictions on these finite parameters that ensure that g is 0. Standard Wald type tests would be applicable as there would be a finite number of restrictions, and the parameter estimates themselves would be asymptotically normal with the usual rate of convergence. The more general nonparametric case is more problematic. First, there is the problem that the number of parameters in the g function is not known and the approximation involves an increasing number of parameters. So even if one could ascertain the distribution of each parameter (and the results of this chapter do not deal with this problem) then one has the problem of testing an increasing number of restrictions, where there is no bound to the number. To consider this difficulty more closely, suppose that one were able to observe the true value of β_2. Then a natural estimator for g would be given by

$$\hat{g} = P_{x_3}(w - Z\beta_2) = P_{x_3}\xi \tag{4.70}$$

(where the restriction $g = 0$ has been imposed). Using the nature of the P_{x_3} matrix one may expect that a test statistic of the form

$$W = \xi'\Psi(\Psi'\Gamma\Psi)\Psi\xi \tag{4.71}$$

could provide the appropriate form of the test. The degrees of freedom, however, are unbounded since the dimension of Ψ increase with the sample size.

One alternative to this type of procedure would be to consider a finite number of the functions in Ψ. In this case, a Lagrange Multiplier (LM) type test could be constructed where the ξ could be replaced by the estimated values from OLS, which are consistent under the null hypothesis. Thus, one is testing whether the OLS residuals are uncorrelated with the candidates x_3. This type of conditional moment testing procedure has been surveyed and developed by Pagan and Vella (1989). The situation here, however, involves an infinite number of conditional moment restrictions. Alternatively one could pick a finite number of points in X^3 space at which to evaluate the g function and use the results of Andrews (1988, 1989d) to obtain an appropriate test statistic. These procedures are likely to possess some power, but are not entirely satisfactory since they do not test the entirety of the null hypothesis and there is likely to be some degree of arbitrariness in the selection of moment conditions.

These two factors make it seem preferable to construct a test of the null H_0:$g = 0$ based on the consequences for β_2 estimators, of violations of the null. A natural comparison is between the $\tilde{\beta}_2$ estimator developed in this chapter and an OLS (or more generally some weighted least squares (WLS)) estimator, denoted $\hat{\beta}_2^w$). Although we have allowed quite general heteroskedasticity we will not rule out the possibility of obtaining an efficient estimator. The main test result, however, will not be developed under this assumption . Note that recent work by Robinson (1988) and White and Stinchcombe (1989) has shown that efficient estimation in the heteroskedastic regression model is possible under quite general conditions — in our case since the residuals are maintained to be normally distributed, such an estimator will attain the Cramer – Rao lower bound which will make the test statistic developed in this section somewhat simpler than it appears, due to a simplification noted by Hausman (1978). We can write the two estimators for comparison

in the useful form

$$\begin{aligned}\sqrt{N_1}(\hat{\beta}_2^w - \beta_2) &= (\frac{1}{N_1}X_2'WX_2)^{-1}\frac{1}{\sqrt{N_1}}X_2'W\xi + o_p(1) \\ &= A_1B_1'\xi + o_p(1) \end{aligned} \tag{4.72}$$

and

$$\begin{aligned}\sqrt{N_1}(\tilde{\beta}_2 - \beta_2) &= \sqrt{N_1}((Z - T_3)'(Z - T_3))^{-1}(Z - T_3)'D_3\xi + o_p(1) \\ &= A_2B_2'\xi + o_p(1) \end{aligned} \tag{4.73}$$

where W is a matrix of weights and the simplifictaion in $\tilde{\beta}_2$ arises from Theorem 3 and the fact that under the null $g = 0$. The rationale for considering a test statistic based on the difference between these two estimators is that if $g = 0$ then the estimators should have the same probability limits (β_2) whereas if $g \neq 0$ then the estimators will have different limits. A significant difference will be interpreted as evidence that the exclusion of the correction term results in selectivity bias or some other form of mis-specification. The following result provides the basis for such a test

Theorem 6: *Under H_0:$g = 0$ and given the conditions of Theorem 5,*

$$N_1(\hat{\beta}_2^w - \tilde{\beta}_2)'\hat{V}_{12}^{-1}(\hat{\beta}_2^w - \tilde{\beta}_2) \rightarrow \chi^2(d_2)$$

where $\hat{V}_{12}$ is a consistent estimator of V_{12} which is given by

$$\text{plim}A_1B_1'\Gamma B_1A_1 + A_2B_2'\Gamma B_2A_2$$

$$-A_1B_1'\Gamma B_2A_2 - A_2B_2'\Gamma B_1A_1$$

which is assumed to be a finite positive definite matrix.

A consistent estimator for V_{12} can easily be obtained by replacing unknowns with estimates and population moments with averages as was done previously. There are also

a number of alternatives that can be used. One which is guaranteed to be positive definite is given by using an estimator for A_1 given by

$$\hat{A}_1 = (\frac{1}{N_1} X_2' \hat{W} X_2)^{-1} \tag{4.74}$$

$$\hat{B}_1' = \frac{1}{\sqrt{N_1}} X_2' \hat{W} \tag{4.75}$$

where $\hat{W}$ are possibly estimated weights, and for the second estimator

$$\hat{A}_2 = (\frac{1}{N_1} (\hat{Z} - P_{x_3} \hat{Z})' (\hat{Z} - P_{x_3} \hat{Z}))^{-1} \tag{4.76}$$

$$\hat{B}_2' = \frac{1}{\sqrt{N_1}} (\hat{Z} - P_{x_3} \hat{Z})' \hat{D}_3 \tag{4.77}$$

while the variance-covariance matrix of ξ can be estimated

$$\hat{\Gamma}^1 = \text{diag}\{\hat{\xi}_1^2\} \tag{4.78}$$

$$\hat{\Gamma}^2 = \text{diag}\{\hat{\xi}_2^2\} \tag{4.79}$$

$$\hat{\Gamma}^{12} = \text{diag}\{\hat{\xi}_1 \hat{\xi}_1\} \tag{4.80}$$

where the matrices are used respectively in the first, second and last two terms in V_{12}. (We ignore the need for consistency results as they follow from Theorem 5 and work in White (1984)). The weights appearing in the first estimator may be any weights (any quasi-Aitken estimator), however in the situation where one is able to compute the asymptotically efficient estimator of White and Stinchcombe (1989) the variance-covariance matrix appearing in the test may be written as

$$V_{12} = \text{plim} A_2 B_2' \Gamma B_2 A_2 - AV \tag{4.81}$$

where AV is the Cramer-Rao lower bound, and this is positive definite although it is difficult to see how one could ensure an estimator in finite samples was positive definite without using the more complicated form above. The test based on the efficient estimator

may be expected to be more powerful (asymptotically) than the more general WLS test, although the relative performance of any of these tests in finite samples is open to question.

4.7.2 A Test for Heteroskedasticity

Similar principles point to the possibility of a test for heteroskedasticity in both sample selection and Tobit models, based on the estimator developed in this chapter. In this case, the null hypothesis of homoskedasticity not only imposes a restriction on the g function, but also imposes a restriction on the λ_t function. That is, it depends on x_1 via $x_1\beta_1$ only. This makes a Wald type test problematic. The same principle of a conditional moment test as discussed in Pagan and Vella (1989) is applicable in this case. Since violation of the hypothesis of homoskedasticity causes mis-specification of the conditional mean of y_2 one would expect that a test could be constructed using the residuals from the model estimated under homoskedasticity. In particular, the null of homoskedasticity would require that these residuals be uncorrelated with any function of x_t. As in the case of a test for selectivity bias there is some degree of arbitrariness in the selection of these conditions and it seems unlikely that such a test would be powerful against all forms of heteroskedasticity. As mentioned in the previous section, a fully robust procedure would be based on an increasing number of such moments (although it is difficult to see how such a procedure could be very powerful in practice).

This suggests that a Hausman type test may have some merit since it will be based on the premise of the null being rejected only when allowing for heteroskedasticity makes a difference in practice, and testing whether the consistency of the estimator is questionable. Such a test would be based on the difference between the preferred estimator under the null and the heteroskedasticity robust estimator developed in this chapter. The preferred estimator may be either the two-step estimator or the MLE. As noted earlier, this

will generally mean the MLE for the Tobit and the two-step estimator for the sample selection model. As noted previously when the preferred estimator is the MLE, the test statistic may be somewhat easier to compute. Here we develop the test for the case where the preferred estimator is the two-step estimator.

In this case, the second stage regression will have the form

$$y_{2t} = x_{2t}\beta_2 + \frac{\sigma_{12}}{\sigma_2}\hat{\lambda}_t + \xi_t + \frac{\sigma_{12}}{\sigma_2}(\lambda_t - \hat{\lambda}_t) \tag{4.82}$$

where $\hat{\lambda}_t$ is estimated using the homoskedastic estimator developed in Section 5. Since the test will be based on the β_2 estimates, we are only interested in this part of the parameter vector and since under the null any weighted least squares estimator of the above is consistent a useful weight could be $\frac{1}{\hat{\lambda}_t}$ in which case the regression equation becomes

$$\hat{w}_t = \hat{z}_t\beta_2 + \frac{\sigma_{12}}{\sigma_2} + \frac{\xi_t}{\hat{\lambda}_t} + \frac{\sigma_{12}}{\sigma_2}\frac{(\lambda_t - \hat{\lambda}_t)}{\hat{\lambda}_t} \tag{4.83}$$

which is similar in form to the heteroskedastic model. The estimator in this case may be written as

$$\hat{\beta}_2 = ((\hat{Z} - E(\hat{Z}))'(\hat{Z} - E(\hat{Z})))^{-1}(\hat{Z} - E(\hat{Z}))(\hat{w} - E(\hat{w})) \tag{4.84}$$

where

$$E(\hat{Z}) = \iota\frac{\iota'\hat{Z}}{N_1} \tag{4.85}$$

which are the averages of each of the regressors. This is similar in form to $\tilde{\beta}_2$. From Theorem 4 we know that these estimators can be written under the null as

$$\sqrt{N_1}(\tilde{\beta}_2 - \beta_2) = (\frac{1}{N_1}(Z - T_3)'(Z - T_3))^{-1} \tag{4.86}$$

$$\begin{aligned}&(\frac{1}{\sqrt{N_1}}((Z - T_3)'D_3^{-1}\xi + \nu'GD_3^{-1}D_5L\epsilon) + o_p(1)\\ &\qquad = A_1(B_{11}'\xi + B_{12}'\epsilon) + o_p(1)\end{aligned} \tag{4.87}$$

where $T_1 = Z$ in the case of the Tobit model, and

$$\sqrt{N_1}(\hat{\beta}_2 - \beta_2) = (\frac{1}{N_1}(Z - E(Z))'(Z - E(Z)))^{-1} \times \tag{4.88}$$

$$(\frac{1}{\sqrt{N_1}}((Z - E(Z)'D_3^{-1}\xi +$$

$$(Z - E(Z))'GD_4X_1(X_1'D_5X_1)^{-1}X_1D_5D_6\epsilon) + o_p(1)$$

$$= A_3(B_{31}'\xi + B_{32}'\epsilon) + o_p(1). \tag{4.89}$$

using the expansion for $\hat{\beta}_1$ given in Section 5. Note that the values in these representations are the true values under the null so that G is proportional to the identity matrix. Using the fact that ξ and ϵ are uncorrelated vectors we now have the following test of mis-specification.

Theorem 7: *Under the null that the underlying model is homoskedastic, and given the conditions of Theorem 4,*

$$N_1(\tilde{\beta}_2 - \hat{\beta}_2)(\hat{V}_{13})^{-1}(\tilde{\beta}_2 - \hat{\beta}_2) \rightarrow \chi^2(d_2)$$

where $\hat{V}_{13}$ is a consistent estimator of the variance-covariance matrix V_{13} given by

$$\text{plim}A_1(B_{11}'\Gamma B_{11} + B_{22}'\Omega B_{22})A_1 + A_3(B_{31}'\Gamma B_{31} + B_{32}'\Omega B_{32})A_3$$

$$-A_1(B_{21}'\Gamma B_{31} + B_{22}'\Omega B_{32})A_3 - A_3(B_{31}'\Gamma B_{21} + B_{32}'\Omega B_{22})A_1$$

which is assumed to be positive definite.

As in the previous case, all one is required to do is to obtain a consistent estimator for the variance-covariance matrix, which can easily be done by replacing unknowns with consistent estimates. Also, one may obtain an estimator that is bound to be positive definite.

Again, if the preferred estimator was the MLE then the test would have the form

$$N_1(\tilde{\beta}_2 - \hat{\beta}_2)'(\hat{V}_2 - \hat{V}^{MLE})^{-1}(\tilde{\beta}_2 - \hat{\beta}_2) \tag{4.90}$$

where the covariance matrix is somewhat easier to compute although it may not be positive definite in finite samples. This may be the desired form of the test in the case of the Tobit model. As noted above, this form of the test is likely to have power against quite arbitrary forms of heteroskedasticity, at least to the extent that such heteroskedasticity causes inconsistency. In addition this form of the test may be powerful against other forms of mis-specification.

As noted earlier, one of the advantages of obtaining RNCAN estimators for these models is that they allow the construction of tests of mis-specification of the sort considered here. Not only does one have a relatively simple estimator to use in the event of model failure, but one also has a relatively nice form of test of model adequacy based directly on the consequences of mis-specification for the parameters of interest.

4.8 Conclusions

In this chapter, we have developed two-step estimation procedures for both the sample selection model and the Tobit model, when there is heteroskedasticity of unknown form in the latent errors. The estimators were shown to be consistent and asymptotically normal under fairly mild regularity conditions. In addition, estimation of the covariance matrix was considered along with specification tests for heteroskedasticity and selectivity bias. One of the nice features of the estimator is the ease of computation being solely based on regressions. A useful result regarding the distribution of averages of nonparametric estimates was developed and used to find the distribution of the second stage estimator, where the problem of pre-estimation error must be dealt with.

Compared to the MLE developed in Chapter 3 the estimator differs in a couple of ways. First, the identification of the model requires that some restrictions must be satisfied by the data generating process. In the case of the Tobit model, this was simply that

at least one of the variables in the mean of the latent variable not appear as an argument of the variance function. Such restrictions were not required to identify the model in the case of the MLE. Another difference is that in the case of the two-step estimator we do not have to worry about the variance-covariance matrix being positive definite, which is required for the MLE method to be well defined. Also the two step estimation method is likely to be somewhat easier to compute in practice. This is evidenced in the Monte Carlo experimentation that appears in Chapter 5.

Chapter 5

Small Sample Properties

5.1 Introduction

In this chapter, we examine the small sample properties of the estimators developed in the preceding chapters. There are a number of reasons for doing this. First, the arguments used to justify the use of the estimators have been asymptotic in nature and in no way guarantee that the estimators will have any merit in samples of the size typically used in practice. Second, the consistency results rely on being able to increase the number of terms in the approximating functions with the sample size. Although in practice a sample size will be forced on an applied econometrician, the number of terms in the approximating functions is a choice variable, so it will be useful to see what number of terms may be required in some simple situations to make the estimators perform well.

As in most simulation work, simple models will be employed and it should be stressed that we make no attempt to examine every possible aspect of the performance of the estimators. Also, we make no attempt to choose the optimal number of terms in the approximations. The experiments will be conducted using a few different number of terms (truncation parameters) which will be chosen *a priori* and could be related loosely to the asymptotics. One interesting thing to consider is how well a simple quadratic approximation performs and whether there is much need to add on the Fourier terms.

The simulation work also gives some guide as to parameterisation of the variance function, which is obviously useful for practical purposes. In particular, in the case of

the MLE estimator, it is required that the variance function be restricted to be positive and the parameterisation should be chosen so that this will be true, plus it should be such that the estimator is relatively easily computed (*i.e.*, convergence is quick and no numerical problems arise). In addition, the ease with which the estimators can be computed will be brought out. For example, standard packages that have nonlinear optimisation routines will be sufficient to compute the estimators. There is in fact no need to program the derivatives (provided that the package has reasonable facilities for computing the derivatives as in SHAZAM and TSP).

The chapter has four remaining sections. In Section 2 we consider the Probit model and compare the usual MLE with the heteroskedasticity adjusted MLE in Chapter 3. Section 3 deals with the Tobit model and compares the usual Tobit MLE with the heteroskedastic MLE of Chapter 3 and the two-step estimator of Chapter 4. Finally, we consider the sample selection model and compare the usual Heckman two-step estimator with variants on this developed in Chapter 4. We do not examine the performance of the MLE in this case due to the fact that it is somewhat more expensive to compute than the other MLEs. In addition, as noted by Newey, Powell, and Walker (1990) the likelihood function in this case may be ill-conditioned making it somewhat difficult to perform a simulation exercise using it. Finally, in Section 5 we summarise the main findings.

5.2 Probit Model

In this section, we consider the Probit model and compare the performance of the usual (homoskedastic) MLE with the heteroskedastic MLE proposed in Chapter 3. The model we consider is based on the following equation for the latent variable

$$y_t^* = a + bx_t + u_t \tag{5.91}$$

where we assume that the x_t are independent and uniformly distributed on $(0.1, 6.1)$ which ensures that that they are contained in the $(0, 2\pi)$ interval required for the use of the Fourier series approximation. We assume that $a = -3$ and $b = 1$ so that approximately one half of the observations on the latent variable y^* will be positive. The discrete dependent variable is determined according to the rule $y_t = I(y_t^* > 0)$. The latent residuals will be generated as independent normal random variables with mean zero and variance given by $\sigma^2(x_t)$ where σ varies across the different experiments. Assuming that we proceed to estimation assuming that the standard deviation of u_t is given by $f(x_t)$ then the contribution to the likelihood for observation t is

$$y_t \log \Phi\left(\frac{a + bx_t}{f(x_t)}\right) + (1 - y_t)\log\left(1 - \Phi\left(\frac{a + bx_t}{f(x_t)}\right)\right) \tag{5.92}$$

Obviously, in the standard case f is assumed constant. In the heteroskedastic case, we must substitute in an approximation based on the arguments of Chapter 3. When doing this the approximating function must be kept away from zero. One way that seems to work well computationally is to let

$$f(x_t) = (\sum_{j=1}^{K} \theta_j \psi_j(x_t))^{-2} \tag{5.93}$$

where the θ_j are parameters that will be estimated along with a and b, and $\psi_j(x_t)$ are the basis functions, or the functions that comprise the Fourier Flexible Functional Form (FFF), discussed in Chapter 3. Since the x_t are scalars these functions will be (in order),

$$1, x, x^2, \sin(x), \cos(x), \sin(2x), \cos(2x), ... \tag{5.94}$$

As noted in Chapter 3, some normalisation of the coefficients a and b will be required to identify the parameters. The normalisation used in chapter 3 may not be easy to impose computationally, but an equivalent normalisation can be imposed by restricting one of the coefficients in the likelihood to a particular value. There are obviously many alternatives,

but the one chosen, and which resulted in the quickest convergence was to restrict $b = 1$ and to maximise over a and the θ_j parameters. To make the results comparable across the different experiments we normalise the resultant estimates so that they lie on the unit circle (*i.e.* $a^2 + b^2 = 1$ holds), and then compare these for the different estimators Note that the true values of a and b correspond to values of -0.94868 and 0.31623. The exponential parameterisation suggested in Chapter 3 was also tried and although it gave sensible results most of the time there were instances when the optimisation procedure failed due to exponential overflow. The parameterisation chosen here was better behaved.

In all the experiments apart from the above data generating process, we also keep constant the sample size, which is 200, and maintain the same set of x_t variables across all experiments. There are 500 replications in each experiment. We conduct five different experiments, one for each different assumed $\sigma(x_t)$ and compare the bias, standard deviation and quartiles for the different estimators. The variance functions will be chosen so that the average variance is constant across the experiments. The variance functions that we use are:

$$\begin{aligned}
(1)\ \sigma^2(x) &= 1 \\
(2)\ \sigma^2(x) &= c_2 * x^2 \\
(3)\ \sigma^2(x) &= c_3 \exp(0.1x)\exp(\exp(0.1x)) \\
(4)\ \sigma^2(x) &= c_4 \exp(-x)\exp(\exp(-x)) \\
(5)\ \sigma^2(x) &= c_5(5(x-3)^4 + 1)
\end{aligned}$$

where the constants are chosen so that the average variance in all experiments is 1. This is done because in LDVMs the performance of the estimators depends on the magnitude of the scale parameter (the σ). This enables one to be able to attribute whatever behaviour is observed for the different estimators, to the fact of changing variance function, as opposed to changing scale. These functions are chosen because it is easy to impose the

No. of Terms	Measure	Exp2	Exp3	Exp4	Exp5
1	L_2	2.6	0.01	0.59	0.65
	L_∞	7.4	0.19	1.6	1.5
3	L_2	0.5	0.00	0.0005	0.147
	L_∞	4.0	0.0001	0.066	0.723
5	L_2	0.17	0.00	0.00001	0.023
	L_∞	2.4	0.00002	0.013	0.371

Table 5.1: Approximate Specification Errors - Probit

constant average variance assumption and because they correspond to different types of heteroskedasticity. The first is homoskedasticity which is used to give some idea of the cost of using the estimators when the usual Probit estimator is valid. The second and third correspond to increasing functions of x, the fourth is a decreasing function of x, while the fifth is not monotonic. To gain some idea of the extent to which the FFF is capable of approximating these functions, we provide, in Table 5.1, different measures of fit for the four heteroskedastic experiments and three different approximations used. Since the normalisation is such that $\frac{1}{\sqrt{\sigma}}$ is approximated by the FFF, we regress these values on the terms in the FFF to see how well the FFF fits. Two measures of fit are provided. The first, denoted by L_2, is the mean squared error, while the second, denoted L_∞, is the maximum difference in the sample, of the function from the corresponding FFF approximation. These measures are computed for the three cases considered in the experiments below.

The rows corresponding to one term can be loosely interpreted as measuring the degree of heteroskedasticity, while the other rows give some indication of how well the FFF fits and hence the degree of mis-specification. As one would expect the FFF fit improves as more terms are added using either measure.

In each of the five experiments, we consider the properties of the following three likelihood based estimators which are all contained in the general likelihood above. Accordingly we identify each estimator with value of K. The first is the usual Probit estimator denoted by $K = 1$. The second is based on a quadratic approximation, and is denoted $K = 3$. The two other estimator corresponds to $K = 5$. Because the optimisation becomes increasingly time consuming the more terms that are included, we neglected any further estimators. The estimators are computed using the Newton-Raphson algorithm in TSP, which uses analytic derivatives. The number of iterations to convergence is generally two to three times the number that are usually required in the ordinary Probit estimation. The results are contained in Tables 5.2 to 5.6.

Estimator	Variable	Bias	SD	LQ	Median	UQ
MLE	one	0.00012	0.00450	-0.95151	-0.94832	-0.94535
	x	0.00006	0.01341	0.30762	0.31732	0.32606
HMLE3	one	0.00045	0.00738	-0.95255	-0.94859	-0.94375
	x	0.00053	0.02190	0.30440	0.31652	0.33066
HMLE5	one	0.00101	0.00765	-0.95310	-0.94820	-0.94234
	x	0.00230	0.02271	0.30266	0.31768	0.33466

Table 5.2: Probit experiment 1

Estimator	Variable	Bias	SD	LQ	Median	UQ
MLE	one	-0.00069	0.00252	-0.95114	-0.94934	-0.94776
	x	-0.00216	0.00761	0.30876	0.31425	0.31899
HMLE3	one	0.00013	0.00446	-0.95244	-0.94824	-0.94491
	x	0.00009	0.01340	0.30471	0.31756	0.32732
HMLE5	one	-0.00017	0.00473	-0.95302	-0.94896	-0.94523
	x	-0.00089	0.01423	0.30292	0.31540	0.32640

Table 5.3: Probit experiment 2

Estimator	Variable	Bias	SD	LQ	Median	UQ
MLE	one	-0.00041	0.00395	-0.95557	-0.95312	-0.95021
	x	-0.00130	0.01230	0.29477	0.30258	0.31160
HMLE3	one	-0.00005	0.00630	-0.95184	-0.94782	-0.94240
	x	-0.00050	0.01848	0.30611	0.31882	0.33449
HMLE5	one	-0.00024	0.00704	-0.95437	-0.94931	-0.94529
	x	-0.00110	0.02104	0.29863	0.31433	0.32624

Table 5.4: Probit experiment 3

Estimator	Variable	Bias	SD	LQ	Median	UQ
MLE	one	0.01260	0.00459	-0.93922	-0.93643	-0.93312
	x	0.03500	0.01215	0.34331	0.35086	0.35956
HMLE3	one	0.00099	0.00333	-0.95013	-0.94796	-0.94574
	x	0.00280	0.00985	0.31860	0.31840	0.32493
HMLE5	one	0.00034	0.00460	-0.95277	-0.94969	-0.94659
	x	0.00130	0.01345	0.30525	0.31318	0.32243

Table 5.5: Probit experiment 4

Estimator	variable	mean	SD	LQ	Median	UQ
MLE	one	-0.00420	0.00343	-0.95521	-0.95309	-0.95074
	x	-0.01300	0.01076	0.29594	0.30268	0.31000
HMLE3	one	-0.00112	0.00223	-0.95100	-0.94975	-0.94854
	x	-0.00346	0.00700	0.30920	0.31301	0.31666
HMLE5	one	-0.00102	0.00424	-0.95028	-0.94884	-0.94721
	x	-0.00337	0.01331	0.31139	0.31576	0.32061

Table 5.6: Probit experiment 5

The results indicate that the estimators that correct for heteroskedasticity do reduce the bias relative to that of the usual Probit estimator. In all the heteroskedastic cases (experiments 2 to 5) the bias is significantly (at the 0.01 level) smaller for both the heteroskedastic MLEs. In experiment 3 and 4 the estimator based on five terms is significantly (at the 0.01 level) less biased than the quadratic approximation. In experiments 2 and 5, HMLE3 and HMLE5 do not have signicantly different biases. In experiments 2 and 3, the bias reduction comes at the cost of a loss of efficiency, with the standard deviation being slightly less than twice as big as in the ordinary Probit case. In experiments 4 and 5, however, the estimator based on a quadratic approximation is less biased and more efficient than the usual Probit estimator and the estimator based on five terms is only slightly less efficient than the Probit MLE. These results indicate that the need for some method of choosing the number of approximating terms may be useful. With regard to the effect of heteroskedasticity on the Probit estimator it seems that the increasing function (experiments 2 and 3) result in an downward biases in both the coefficients. The decreasing variance function results in biases in the opposite direction. The use of HMLE3 and HMLE5 in homoskedastic situations may be costly, as one might expect, although only estimates based on HMLE5 are significantly biased. The order of the biases are small in the experiments but are statistically significant (at the 0.01 level) in all cases except the first. The small order of the biases is obviously related to the experimental design, but it is noticable that the degree of bias did not seem to be related to the degree of mis-specification, as measured in Table 5.1.

5.3 The Tobit Model

In this section, we compare the performance of the usual (homoskedastic) MLE with the heteroskedastic MLE of section 3 and the two step estimator of section 4. In light of the

identification conditions needed for the two step estimator, the simplest specification of the model is given by

$$y_t^* = a + bx_t + cz_t + u_t \tag{5.95}$$

where u_t are independent normal random variables with mean zero and variance given by some function of x_t, $\sigma(x_t)^2$. The x_t and z_t are independent, uniform random variables on $(0.1, 6.1)$, and the parameters will be fixed at $(-6, 1, 1)$ so that approximately one half of the observations will be censored, with the observed dependent variable given by the rule $y_t = \max\{0, y_t^*\}$. Writing in shorthand the mean of the latent variable as μ_t we can write down the contribution to the logarithm of the likelihood function when the square root of the variance function is approximated by the function $f(x_t)$ as

$$l_t = (1 - I(y_t^* > 0))\log \Phi\left(\frac{\mu_t}{f(x_t)}\right) + I(y_t^* > 0)\log\left(\frac{1}{f(x_t)}\phi(\frac{y_t - \mu_t}{f(x_t)})\right) \tag{5.96}$$

In the homoskedastic MLE case, the function will be a constant, and in the heteroskedastic case we will use the approximation based on

$$f(x_t) = \exp\left(\sum_{j=1}^{K} \theta_j \psi_j(x_t)\right) \tag{5.97}$$

where the ψ_j functions are as in the Probit case. As in the Probit case, we distinguish different MLEs by the number of terms used to approximate the variance function. For the experiments we use the shorthand HMLEK where HMLE denotes the fact that it is the heteroskedastic MLE and the K denotes the number of terms in the approximation. We perform the experiments for $K = 3$ and $K = 7$ due to the cost of computing the MLE in a Monte Carlo experiment. The homoskedastic MLE corresponds to $K = 1$, but is denoted MLE.

To compute the two-step estimator we need to define the qualitative variable $v_t = I(y_t^* > 0)$. The first step in the computation of the two step estimator regresses v_t on

the FFF constructed from the variables x and z as in

$$v_t = \sum_{i=1}^{K_1} \theta_i^1 \psi_i(x_t, z_t) + error \tag{5.98}$$

where the ψ functions are constructed using the algorithm described in Gallant (1981):the first five consist of $1, x, z, x^2, z^2, xz$ and correspond to the quadratic part of the FFF; next there are the trigonometric functions of the form

$$\sin(k_x x + k_z z), \cos(k_x x + k_z z)$$

where the k_x, k_z are constructed using Gallant (1981), the first few being

$$(1,0), (0,1), (2,0), (0,2), (1,-1)....$$

The terms are ordered by the index formed by $|k_x| + |k_z|$, and since these will give groups of increasing size we ensure that the estimators are computed using all the terms up to a specific group.

Denote the predicted values from this regression by $\hat{v}_t$. As noted in the Chapter 4, we must censor these values to ensure that they are between 0 and 1. This is done using the rule in Chapter 4 where the value of δ is 0.01. As noted in Chapter 4, this is somewhat arbitrary. As this value gave estimators that performed fairly well, no experimentation with different values was done. Denote the censored values by $\tilde{v}_t$. In the second stage, we transform these values to create the analogue of the inverse Mills ratio by

$$\hat{\lambda}_t = \frac{\phi(\Phi^{-1}(\tilde{v}_t))}{\tilde{v}_t} \tag{5.99}$$

In the second stage the following regression is performed using the observations for which $y_t^* > 0$,

$$\frac{y_t}{\hat{\lambda}_t} = a\frac{1}{\hat{\lambda}_t} + b\frac{x_t}{\hat{\lambda}_t} + c\frac{z_t}{\hat{\lambda}_t} + \sum_{k=1}^{K_2} \theta_k^2 \psi_k(x_t) + error \tag{5.100}$$

where the ψ_k functions are as above. We denote the different estimators in this case by $TK1, K2$ where T signifies that it is a two step estimator and $K1$ and $K2$ give the number of terms used in the approximating functions.

In all experiments, we use 200 observations and perform 250 replications due to the cost of computation. As before, the variance functions used in the five experiments are given by:

$$\begin{aligned}
(1)\ \sigma^2(x) &= 10 \\
(2)\ \sigma^2(x) &= c_2 x \\
(3)\ \sigma^2(x) &= c_3 \exp(0.3x)\exp(\exp(0.3x)) \\
(4)\ \sigma^2(x) &= c_4 \exp(-x)\exp(\exp(-x)) \\
(5)\ \sigma^2(x) &= c_5(5(x-3)^4+1)
\end{aligned}$$

The constants are chosen so that the average variance in all the experiments is about 10. Again, to gain some idea of the specification errors involved and the ability of the FFF to approximate these functions, we compute measures of fit analogous to those in Table 5.1. We do this for two normalisations. The first, which is used for the two-step estimators, recognises that σ appears in the regression function in the second step. The second, relevant for the MLE based estimators recognises that we are implicitly trying to approximate $\log \sigma$.

No. of Terms	Measure	Exp2	Exp3	Exp4	Exp5
1	L_2	9.9	48.7	47.2	37.9
	L_∞	7.7	7.9	22.6	15.3
3	L_2	0.05	0.33	1.7	0.05
	L_∞	0.95	2.5	5.1	0.56
7	L_2	0.0013	0.003	0.020	0.007
	L_∞	0.17	0.21	0.48	0.21
11	L_2	0.0001	0.0001	0.009	0.0004
	L_∞	0.07	0.044	0.12	0.045

Table 5.7: Approximate Specification Errors - Tobit, σ

No. of Terms	Measure	Exp2	Exp3	Exp4	Exp5
1	L_2	0.17	0.93	0.88	1.02
	L_∞	1.5	2.01	1.9	1.5
3	L_2	0.0083	0.00056	0.00067	0.109
	L_∞	0.47	0.076	0.091	0.79
7	L_2	0.00047	0.0	0.0	0.00083
	L_∞	0.12	0.003	0.006	0.08

Table 5.8: Approximate Specification Errors - Tobit, $\log \sigma$

All estimators were computed using SHAZAM which uses numeric derivatives and the algorithm of Davidon-Fletcher-Powell to compute the MLEs. The main stumbling block in computing the two-step estimator was in obtaining the inverse of the Normal cdf. SHAZAM was used because it allowed easy calculation of the inverse while TSP did not. Convergence in the MLE computations generally took two to three times the number of iterations usually required for the homoskedastic Tobit MLE. Note that we also computed the Heckman two-step estimator to enable some sort of comparison between the MLE and this estimator. It will be denoted Heck in the tables. The results are contained in Tables 5.9 to 5.13.

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	5.772	0.773	-0.771	-0.155	0.332
	x	-0.524	0.142	0.368	0.471	0.577
	z	-0.519	0.153	0.379	0.489	0.591
MLE	one	-0.088	0.887	-6.707	-6.05	-5.455
	x	0.017	0.156	0.904	1.018	1.130
	z	0.008	0.179	0.889	1.005	1.131
HMLE3	one	-0.592	1.343	-7.464	-6.590	-5.585
	x	-0.113	0.204	0.768	0.903	1.030
	z	0.233	0.337	1.035	1.288	1.453
HMLE7	one	0.088	1.026	-6.631	-5.795	-5.135
	x	-0.019	0.186	0.854	0.985	1.094
	z	0.001	0.187	0.875	0.990	1.137
Heck	one	0.282	7.787	-10.228	-5.341	-0.816
	x	-0.026	0.745	0.506	0.902	1.423
	z	-0.032	0.721	0.524	0.939	1.329
T6,3	one	-1.216	43.565	-25.447	-6.374	11.453
	x	0.236	4.488	-0.424	1.021	2.810
	z	0.031	4.014	-0.731	1.033	3.255
T10,7	one	1.851	31.565	-19.878	-2.434	12.150
	x	-0.023	3.667	-0.755	0.831	2.771
	z	-0.232	2.706	-0.533	0.579	1.998
T18,11	one	3.027	19.174	-11.216	-1.765	6.974
	x	-0.102	2.442	-0.337	0.779	1.917
	z	-0.356	1.750	-0.298	0.577	1.560

Table 5.9: Tobit experiment 1

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	4.564	0.937	-2.074	-1.401	-0.766
	x	-0.152	0.166	0.738	0.843	0.973
	z	-0.544	0.169	0.348	0.450	0.565
MLE	one	-2.186	1.117	-8.847	-8.115	-7.432
	x	0.409	0.193	1.283	1.411	1.544
	z	0.095	0.200	0.950	1.095	1.236
HMLE3	one	-1.668	1.473	-8.559	-7.609	-6.792
	x	0.066	0.217	0.923	1.084	1.215
	z	0.314	0.327	1.155	1.316	1.519
HMLE7	one	-0.099	1.096	-6.555	-5.838	-5.160
	x	-0.041	0.192	0.819	0.948	1.094
	z	0.007	0.195	0.881	0.990	1.134
Heck	one	1.772	9.238	-11.279	-4.942	0.765
	x	0.197	0.871	0.635	1.190	1.755
	z	-0.213	0.849	0.218	0.753	1.293
T6,3	one	2.052	48.138	-22.590	-2.595	14.500
	x	-0.163	2.748	-0.252	1.027	2.530
	z	-0.269	5.639	-1.617	0.514	2.555
T10,7	one	2.248	33.259	-16.602	-3.742	11.184
	x	0.093	3.092	-0.039	0.908	2.456
	z	-0.321	3.979	-0.819	0.469	1.933
T18,11	one	4.023	21.732	-9.286	-1.614	8.479
	x	-0.163	2.748	-0.150	0.805	1.899
	z	-0.408	2.328	-0.430	0.552	1.421

Table 5.10: Tobit experiment 2

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	0.238	1.373	-6.722	-5.819	-4.882
	x	0.663	0.255	1.491	1.651	1.832
	z	-0.396	0.198	0.465	0.600	0.731
MLE	one	-5.312	1.599	-12.238	-11.195	-10.169
	x	0.876	0.283	1.672	1.847	2.069
	z	0.396	0.205	1.256	1.385	1.521
HMLE3	one	-0.327	0.768	-6.769	-6.291	-5.823
	x	0.023	0.114	0.936	1.026	1.108
	z	0.055	0.128	0.977	1.042	1.127
HMLE7	one	-0.071	0.843	-6.627	-5.974	-5.507
	x	0.002	0.150	0.899	1.003	1.107
	z	0.016	0.130	0.932	0.998	1.097
Heck	one	1.555	8.363	-9.779	-4.756	1.035
	x	0.557	0.700	1.148	1.548	1.988
	z	-0.541	0.966	-0.179	0.500	1.081
T6,3	one	-1.987	23.413	-19.509	-5.909	4.858
	x	0.678	1.453	0.860	1.513	2.503
	z	-0.132	3.627	-1.000	0.367	2.806
T10,7	one	-0.295	17.690	-16.315	-6.073	5.300
	x	0.856	1.408	0.855	1.771	2.717
	z	-0.526	2.850	-1.259	0.373	2.160
T18,11	one	0.529	7.140	-9.777	-4.952	-1.229
	x	0.346	0.802	0.774	1.295	1.794
	z	-0.282	1.100	0.116	0.667	1.380

Table 5.11: Tobit experiment 3

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	6.776	1.009	0.099	0.836	1.482
	x	-1.107	0.177	-0.212	-0.107	0.012
	z	-0.396	0.103	0.536	0.612	0.677
MLE	one	1.938	0.615	-4.454	-4.020	-3.636
	x	-0.493	0.124	0.415	0.507	0.589
	z	0.017	0.115	0.932	1.009	1.089
HMLE3	one	-0.174	0.435	-6.441	-6.154	-5.889
	x	0.008	0.078	0.944	1.009	1.060
	z	0.029	0.043	0.997	1.030	1.056
HMLE7	one	0.004	0.373	-6.228	-5.998	-5.729
	x	0.004	0.067	0.954	0.994	1.040
	z	0.002	0.041	0.971	1.003	1.030
Heck	one	-3.423	2.609	-11.150	-9.253	-7.509
	x	0.005	0.328	0.789	1.000	1.216
	z	0.623	0.239	1.437	1.614	1.758
T6,3	one	0.111	2.022	-7.472	-5.960	-4.476
	x	-0.023	0.291	0.772	0.989	1.216
	z	-0.002	0.123	0.908	0.994	1.092
T10,7	one	-0.013	1.913	-7.389	-6.106	-4.752
	x	0.000	0.295	0.817	1.024	1.223
	z	0.002	0.121	0.922	0.993	1.076
T18,11	one	1.540	1.760	-6.500	-5.643	-4.621
	x	-0.110	0.321	0.752	0.931	1.065
	z	-0.005	0.079	0.943	0.994	1.043

Table 5.12: Tobit experiment 4

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	4.850	1.27	-1.916	-1.020	-0.204
	x	-0.260	0.216	0.532	0.709	0.867
	z	-0.630	0.178	0.250	0.366	0.467
MLE	one	-1.250	1.187	-8.111	-7.178	-6.340
	x	0.012	0.227	0.854	1.022	1.170
	z	0.274	0.183	1.149	1.277	1.385
HMLE3	one	-0.05	1.008	-6.648	-5.918	-5.373
	x	-0.466	0.228	0.355	0.524	0.700
	z	0.354	0.186	1.230	1.343	1.470
HMLE7	one	0.037	0.478	-6.244	-5.916	-5.632
	x	-0.019	0.112	0.893	0.988	1.052
	z	0.006	0.063	0.962	1.004	1.056
Heck	one	-6.79	6.296	-16.378	-11.939	-8.450
	x	0.489	0.489	1.166	1.472	1.764
	z	0.848	0.802	1.296	1.721	2.335
T6,3	one	-1.162	28.173	-19.244	-5.313	6.319
	x	0.676	1.845	0.674	1.387	2.428
	z	-0.320	3.874	-0.993	0.438	2.536
T10,7	one	-1.870	10.36	-12.306	-7.488	-2.325
	x	0.680	1.269	0.938	1.536	2.320
	z	-0.185	1.429	0.016	0.825	1.630
T18,11	one	-0.140	3.516	-8.307	-5.826	-3.866
	x	0.170	0.732	0.709	1.117	1.604
	z	-0.082	0.518	0.597	0.889	1.224

Table 5.13: Tobit experiment 5

The results indicate that the heteroskedastic MLEs perform quite well under all forms of heteroskedasticity. In particular the MLE based on $K = 7$ is generally significantly (at the 0.01 level) less biased and more efficient than the usual MLE under heteroskedasticity and is almost unbiased in most of the cases. The MLE based on a quadratic approximation generally performs well, but is usually inferior to the other heteroskedastic MLE. In the homoskedastic case, there is some cost in terms of bias and efficiency loss in using HMLE3 and, to a lesser extent, HMLE7.

With regard to the two-step estimators the efficiency loss is quite marked. Moreover, in the cases where there is not substantial heteroskedasticity the heteroskedastic two-step estimators perform quite poorly, as in experiments 1 and 2. The estimates of the intercepts are particularly biased and variable, although T6,3 and T10,7 are not significantly biased. The performance also varies quite a lot depending on the number of terms used in the approximation. However, it seems that increasing the number of terms reduces the standard deviation, at least for the estimators considered here. One would hope that the performance in such situations would improve in larger samples, and with more terms added to the approximations. In the experiments with considerable heteroskedasticity (3, 4 and 5), the heteroskedastic estimators generally improve the bias significantly over the Heck estimator (especially in experiments 4 and 5). Curiously, however, the bias reduction is more pronounced for the constant and the coefficient of z_t, both of which are poorly estimated by Heck. In experiments 3, 4 and 5, T18,11 estimator is just as efficient as Heck. The T18,11 estimator is not always the least biased of the two-step estimators so that in practice some criteria for choosing the number of terms may be useful. The performance of both the heteroskedastic MLE and the two-step estimators is encouraging. With regard to the nature of the bias in the MLE for the Tobit model, it appears that as in the Probit case an increasing variance function causes a downward bias of the intercept and an upward bias in the slope coefficients with the bias being

larger for the coefficient of x (which is the variable in the variance function). A decreasing variance function has the opposite effect on the intercept and coefficient of x and little effect on the coefficient of z, while the nonmonotonic variance function seems to influence the intercept and coefficient of z most severely. The bias in using Heck is not always the same sign as that in the Tobit MLE, especially with respect to the effects on the intercept and coefficient of z. Unlike the Probit case, the degree of bias did appear to be related to the degree of mis-specification as measured in Tables 5.7 and 5.8.

5.4 The Sample Selection Model

The simplest model that we can consider in which all of the identification conditions of section 4 are satisfied is given by

$$y_{1t}^{*} = a + bx_t + u_{1t} \tag{5.101}$$

$$y_{2t}^{*} = c + dx_t + ez_t + u_{2t} \tag{5.102}$$

where the dependent variables that are observed are given by the relations

$$y_{1t} = I(y_{2t}^{*} > 0)y_{1t}^{*} \tag{5.103}$$

$$y_{2t} = I(y_{2t}^{*} > 0) \tag{5.104}$$

and where the latent errors u_{1t} and u_{2t} are bivariate normal random variables with covariance matrix that depends on some function of x_t. The coefficients in the model are set to $\{0, 1, -6, 1, 1\}$ so that approximately one half of the observations will be censored and hence one half "selected". The correlation between the latent residuals will be set at 0.75 for all the experiments. As in the previous models, the x_t and z_t variables are uniformly distributed on $(0.1, 6.1)$, and are held fixed throughout the different experiments.

We compare the performance of the two-step estimator with that of the homoskedastic two step estimator proposed by Heckman (1979). The homoskedastic two-step estimator

is computed in the usual way, by performing Probit estimation in the first stage using y_{2t}, x_t and z_t. Then the inverse Mills ratio is constructed and used as a regressor in the second step regression, where for observations with $y_{2t} = 1$, we regress y_{1t} on a constant, x_t and the inverse mills ratio. We also compute a weighted version of this estimator, where the weights are the inverse of the inverse mills ratio. This is because the heteroskedastic two step estimator is based on a similar weighting. These estimators will be denoted by $H1$ and $H2$ respectively. The heteroskedastic two step estimators are computed in a way that is very similar to the Tobit case except that in the second step regression the term corresponding to the z_t regressor is omitted. We compute three different heteroskedastic two step estimators. The first, denoted by $T6,3$, is based on only the use of quadratic approximations The next two denoted by $T10,7$ and $T18,11$ add on the denoted number of Fourier terms (*i.e.*, $T10,7$ has four Fourier terms in addition to the quadratic terms in each of the two approximations). In addition we compute an estimator which uses the second step of the heteroskedastic estimator, but uses the homoskedastic inverse Mills ratio. This estimator will be denoted by $TP,11$, and is computed to see to what extent the bias is due to the bias in the estimation of the first step.

No. of Terms	Measure	Exp2	Exp3	Exp4	Exp5
1	L_2	9.9	49.6	38.7	37.8
	L_∞	7.7	18.4	16.8	15.3
3	L_2	0.045	0.362	0.270	0.049
	L_∞	0.948	2.07	1.56	0.56
7	L_2	0.0013	0.0008	0.0006	0.007
	L_∞	0.17	0.06	0.073	0.207
11	L_2	0.0001	0.00002	0.00002	0.0004
	L_∞	0.069	0.017	0.015	0.045

Table 5.14: Approximate Specification Errors - Sample Selection, h

The number of observations is set at 200 and there are 500 replications. The errors are generated by taking i.i.d. Normal random pairs with correlation 0.75, and multiplying them by a function of x_t, denoted by $h(x_t)$, which is constructed so that the average variance is roughly constant (as close to 100 as possible) across the different experiments. The five h functions used are given by

$$\begin{aligned}
(1)\ h(x) &= 10 \\
(2)\ h(x) &= c_1\sqrt{x} \\
(3)\ h(x) &= c_2 \exp(\frac{1}{2}x) \\
(4)\ h(x) &= c_3 \exp(-\frac{1}{2}x) \\
(5)\ h(x) &= c_4(5(x-3)^4+1)^{1/2}.
\end{aligned}$$

The constants are chosen so that the mean of the function (over x) is approximately 100. Again, we provide measures of fit for the FFF approximations, by fitting the various FFF approximations to the data on h generated using these functions.

The estimators are computed using SHAZAM. Various summary statistics for the

estimators (in addition to the least squares estimator) are contained in Tables 5.15 to 5.19.

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	7.478	1.842	6.351	7.416	8.669
	x	-0.489	0.481	0.207	0.516	0.849
H1	one	-1.234	25.021	-6.629	1.087	6.434
	x	0.096	1.789	0.451	0.915	1.547
H2	one	-1.349	25.587	-6.832	0.764	5.783
	x	0.101	1.831	0.411	0.977	1.596
TP,11	one	0.653	63.613	-11.487	2.301	13.650
	x	-0.104	12.844	-2.570	0.417	3.503
T6,3	one	1.707	20.276	-7.943	3.421	12.911
	x	-0.153	4.870	-1.932	0.609	3.621
T10,7	one	3.508	18.782	-4.413	4.125	14.060
	x	-0.163	4.999	-1.698	0.775	3.141
T18,11	one	5.736	22.911	-3.797	6.304	14.164
	x	-0.369	4.982	-1.437	0.668	2.861

Table 5.15: Sample selection experiment 1

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	4.016	1.610	3.013	3.972	5.072
	x	0.556	0.510	1.189	1.586	1.897
H1	one	-5.086	17.550	-9.967	-2.538	2.668
	x	1.287	1.596	1.518	2.095	2.796
H2	one	-5.48	18.324	-10.844	-2.877	-3.534
	x	1.281	1.626	1.363	2.062	2.762
TP,11	one	1.234	66.014	-14.453	0.582	11.509
	x	-0.376	16.529	-2.006	1.184	4.784
T6,3	one	-0.374	19.123	-8.876	0.767	9.748
	x	0.575	5.250	-1.355	1.351	4.495
T10,7	one	0.661	18.602	-7.345	0.837	9.467
	x	0.699	5.031	-1.025	1.672	4.491
T18,11	one	2.475	18.394	-4.943	2.818	10.365
	x	0.518	4.629	-0.630	1.269	3.689

Table 5.16: Sample selection experiment 2

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	-2.500	2.330	-3.979	-2.378	-0.938
	x	2.310	0.843	2.753	3.308	3.876
H1	one	-7.903	10.475	-12.656	-6.668	-2.289
	x	2.831	1.353	2.991	3.797	4.562
H2	one	-11.036	14.687	-17.978	10.015	-2.188
	x	3.377	1.939	3.189	4.181	5.509
TP,11	one	-8.993	31.024	-23.343	-7.758	6.870
	x	2.941	8.074	-0.539	3.791	8.044
T6,3	one	-2.506	17.541	-10.896	-2.275	5.972
	x	1.138	5.810	-0.800	2.235	5.543
T10,7	one	-4.700	18.557	-12.877	-3.367	4.823
	x	2.087	5.921	-0.483	3.124	6.354
T18,11	one	-4.913	28.367	-11.486	-2.938	4.261
	x	2.359	6.729	0.097	3.019	6.104

Table 5.17: Sample selection experiment 3

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	11.967	2.788	10.06	11.904	13.839
	x	-2.325	0.603	-1.706	-1.327	-0.931
H1	one	6.564	6.451	2.794	7.217	10.725
	x	-1.625	0.949	-1.247	-0.720	-0.071
H2	one	4.230	4.095	1.869	4.374	6.980
	x	-1.081	0.630	-0.512	-0.112	0.269
TP,11	one	4.163	8.109	-0.988	4.365	9.504
	x	-0.935	1.604	-1.007	0.062	1.512
T6,3	one	0.358	7.168	-4.300	0.529	5.251
	x	-0.068	1.267	0.086	0.913	1.773
T10,7	one	0.853	8.431	-4.170	0.851	5.961
	x	-0.157	1.532	-0.066	0.804	1.741
T18,11	one	3.279	8.975	-2.913	2.943	8.249
	x	-0.607	1.654	-0.517	0.446	1.512

Table 5.18: Sample selection experiment 4

Estimator	Variable	Bias	SD	LQ	Median	UQ
OLS	one	4.490	2.538	2.783	4.515	6.265
	x	-0.094	0.709	0.444	0.912	1.419
H1	one	-0.842	4.061	-3.151	-0.475	1.821
	x	0.250	0.798	0.667	1.215	1.786
H2	one	-3.744	4.762	-6.410	-3.419	-0.100
	x	0.937	1.035	1.208	1.904	2.619
TP,11	one	-0.262	8.72	-5.523	0.493	4.954
	x	0.529	2.476	0.020	1.372	3.055
T6,3	one	-2.376	24.7	-5.992	1.076	7.382
	x	0.891	5.700	-0.582	1.164	3.277
T10,7	one	-1.434	20.824	-7.345	1.145	7.38
	x	0.645	4.889	-0.846	1.165	3.513
T18,11	one	-2.044	15.383	-7.303	-0.216	5.682
	x	0.738	4.050	-0.451	1.258	3.405

Table 5.19: Sample selection experiment 5

The results indicate that the heteroskedastic estimators perform fairly well. The bias reduction is significant (at the 0.01 level) in experiments 3 and 4, with respect to both coefficients, in experiment 5 for the constant, experiment 2 for the slope. As in the Tobit experiments, the heteroskedastic two-step estimators do not estimate the intercept very well when there is little in the way of heteroskedasticity, as in experiments 1 and 2, although the estimates of the slope are not badly biased in these cases. Although there is substantial bias reduction in using the heteroskedastic two-step estimators, the estimators, are in most cases (with the quadratic approximation estimator being slightly better), still significantly biased. The quadratic approximation based estimator appears to be somewhat better than the other heteroskedastic estimators in estimating the intercept.

It is noteworthy that the usual homoskedastic estimators, H1 and H2, are biased significantly (at the 0.01 level) compared to Heck in the Tobit case. This appears to be due to the fact that average variance is higher in this case. The estimator based on standard Probit in the first step, performs well in some cases, but is very inefficient in the first three experiments. Even in the first experiment where the Probit estimator is valid in the first step, the TP,11 estimator is very inefficient relative to the two step estimators that take account of heteroskedasticity (when they need not) in both steps. The conjecture, regarding possible efficiency gains from using a mis-specified first step estimator, made in Chapter 4, appears to be untrue, at least in the experiments conducted. This seems to indicate that both sources of mis-specification may have serious consequences for the properties of the usual estimators.

Generally, the heteroskedastic estimators are less efficient than the homoskedastic ones, although in the estimation of b in experiment 4 the difference is not that great. Among the different heteroskedastic estimators the relative variances depend on the experiment, once again suggesting the need for some criteria for choosing an "optimal"

estimator. Although the quadratic approximation does quite well it is evident that when the variance function is not monotonic some of the Fourier terms may improve the estimator as evidenced in Table 5.19 where the quadratic approximation does worst among the three heteroskedastic estimators.

The nature of the biases is as one might anticipate with the increasing variance function biasing the coefficients in a way that indicates a rotation of the regression line in an anti-clockwise direction. That is, the slope increases and the intercept decreases. This is quite marked in experiment 3 where only the quadratic estimator seems to succeed in reducing the slope appreciably. In experiment (4), precisely the opposite occurs, and in this case the heteroskedastic estimators are quite successful at reducing the bias without sacrificing much in the way of efficiency. In experiments 2 and 3 the bias in the use of the homoskedastic estimators is so bad that even the OLS estimator appears to be preferred. As in the Tobit case, there did seem to be some relationship between the degree of mis-specification measures and the biases.

5.5 Conclusions

In this chapter, we have considered the performance of the estimators developed in Chapters 3 and 4. The findings may be summarized as follows. Both the MLE based estimators of Chapter 3 and the two step estimators of Chapter 4 seem to be successful in reducing the bias when there is moderate to extreme heteroskedasticity. In the case of the Probit and Tobit experiments, the bias reduction in using the MLE based estimators may even be accompanied by improved efficiency, although this only occurs in one case considered for the Probit model. The two-step estimators are generally much less efficient for the Tobit and sample selection models, but relative to their homoskedastic counterparts they are also successful in reducing bias, with in some cases little if any loss in efficiency. It

was also noted that when there was little or no heteroskedasticity, the two step estimators performed relatively poorly, so there may be some cost in using them when they are not needed. This preliminary examination of these estimators on the whole gives quite encouraging results with regard to the performance of the estimators.

It was quite apparent that substantial biases may occur in the usual estimators and a distinct pattern emerged (at least in the experiments performed in this chapter). If the variance is related positively to one of the regressors then this generally results in an upward bias in the regression coefficient of that variable and a downward bias in the intercept. The opposite seemed to hold for the case where the variance is negatively related to the regressor.

One thing that was noticable was that the performance among various versions (corresponding to different numbers of approximating terms) of the estimators proposed in this thesis differed so there appears to be some need for criteria for choosing an optimal estimator in a given instance may be desirable. The work of Andrews (1989e) and references therein, relating to cross validation may be of some use here, and it would be interesting to see how the chosen optimal estimator performs relative to the above.

Chapter 6

Conclusions

This thesis has considered the problem of estimating limited dependent variable models when there is heteroskedasticity of unknown form in the latent residuals, under a maintained assumption of normality. As noted, this situation differs from the usual regression model since the commonly applied estimators (obtained under a homoskedastic assumption) are inconsistent as well as inefficient. The aim of the thesis was to propose estimation methods which yield estimators that are consistent and two such methods were considered.

The first, proposed in Chapter 3, is based on maximum likelihood estimation, and deals with the heteroskedasticity by approximating the variance with a Fourier series approximation. Estimation proceeds by finding the parameters of the approximation and the parameters of interest that maximise the likelihood function. Given various regularity conditions (including smoothness of the variance function) and an assumption that the number of terms in the approximation increases with the sample size, such estimators were shown to be consistent in the three most commonly used LDVMs, the Probit, Tobit, and sample selection models.

The second type of estimation technique, proposed in Chapter 4, is based on a two-step estimation strategy and was developed for the Tobit and sample selection models. The method takes into account the possibility of heteroskedasticity in both stages using Fourier series type regression estimators. In the first stage, a nonparametric regression is performed to estimate the discrete part of the model and then in the second stage these

nonparametric estimates are used to construct a correction term. This correction term is used to transform the second stage regression so that in the second stage the model resembles a semi-parametric regression. The estimators were shown to be consistent and asymptotically normal at the usual $N^{1/2}$ rate for the two models under fairly general conditions. One of the interesting features of the estimator is that although the first stage estimates are consistent at a slower than $N^{1/2}$ rate, and appear as pre-estimation error in the second stage, the (weighted) averaging of these over the observations restores the $N^{1/2}$ for the estimates from the second stage. As a by-product of the results, we were able to construct specification tests for heteroskedasticity and selectivity bias based on the work of Hausman (1978) and White (1980b).

Since the arguments used to justify the estimators are largely asymptotic, a small sampling experiment was conducted in Chapter 5 to see whether the estimators could perform well in small samples, and to examine the feasibility of computing the maximum likelihood estimators. The results were encouraging. Both estimators seemed to be successful in reducing the bias in the usual estimators, and in some situations this came at a small efficiency loss. In fact, in some instances the heteroskedasticity robust estimators appeared to be more efficient than the usual estimators. This was particularly the case with the likelihood based Tobit estimators. In addition, fairly small order approximations seemed to perform quite well. As far as computation goes the estimators were relatively simple to compute. The two step estimators needed only regressions and a nonlinear transformation, while the maximum likelihood estimators could be computed with packages which possessed nonlinear optimisation routines.

One of the obvious avenues where further research is needed is in obtaining distributional results for the likelihood based methods of Chapter 3. This appears to be in sight as methods for regressions with increasing dimension are now available (as in Andrews

(1988, 1989d)). Also Portnoy (1988) has considered the distribution of parameter estimates in the maximum likelihood context when the number of parameters increases. He does, however, assume that the data generating process changes with the sample size so that the model is correctly specified at each stage. The situation we have is one where the data generation process is fixed and the number of parameters increases so that in the limit the model is correctly specified. It would seem that some headway could be made by adapting the results of Portnoy to the mis-specification situation, as done for the standard case by White (1982), and to then show that mis-specification does in fact disappear at a sufficiently fast rate so that the asymptotic distributions are appropriately centered.

One of the drawbacks of the results obtained in this thesis is that they are obtained under a maintained assumption of normality. It would be nice if one could relax this. The work of Powell (1984, 1986) and Manski (1975, 1985) is one approach, but as mentioned earlier this approach appears to be limited to univariate models. If this is indeed the case, then one would be required to generalise the estimators developed in this thesis to a heteroskedastic non-normal situation. One obstacle to such a development is a possible identification problem. This seems to be particularly true for the two-step estimation methods where the structure possessed when the model is either non-normal (in which case it possesses a multiple index structure) normal and heteroskedastic will be lost and one may end up with an almost pure semi-parametric regression model. As the work of Robinson (1988) shows certain exclusion restrictions are neeeded to identify the usual parameters of interest, and these are likely to be much stronger than those needed in any of Newey (1988), Powell (1988) or the work of Chapter 4. Nevertheless, further work is needed on the identification and estimation of LDVMs under general data generating processes.

Appendix A

Proofs for Chapter 3

Proof of Theorem 1:

This is essentially the same as the proof of Theorem 0 in Gallant and Nychka (1987) but is included for completeness. We consider those points ω on the abstract probability space $(\Omega, \Im, P)$ for which we have $\lim_{n\to\infty} K_n = \infty$ and for which condition (c) of the theorem holds. Almost all ω satisfy this. Now since $(\hat{\beta}_n, \hat{\sigma}_n) \in B \times \Sigma$ and $B \times \Sigma$ is compact in $(\|\beta\|_E^2 + \|\sigma\|^2)^{1/2}$, we must have a subsequence $(\hat{\beta}_{n_j}, \hat{\sigma}_{n_j})$ converging to some $(\beta^0, \sigma^0) \in B \times \Sigma$ in this metric. By condition (b), however, there is a sequence $\{\sigma^0_{n_j}\}_{j=1}^{\infty}$ with $\sigma^0_{n_j} \in \Sigma_{K_{n_j}}$ such that $\lim_{j\to\infty} \|\sigma^0_{n_j} - \sigma^*\| = 0$. (Here we have corrected an error in the proof of Theorem 0 in Gallant and Nychka (1987)). Since the sequence $(\hat{\beta}_{n_j}, \hat{\sigma}_{n_j})$ minimises the objective function, whereas $(\beta^*, \sigma^0_{n_j})$ need not, then

$$s_{n_j}(\hat{\beta}_{n_j}, \hat{\sigma}_{n_j}) \leq s_{n_j}(\beta^*, \sigma^0_{n_j})$$

for all j. By uniform convergence (condition (c)) we have

$$\lim_{j\to\infty} s_{n_j}(\hat{\beta}_{n_j}, \hat{\sigma}_{n_j}) = \bar{s}(\beta^0, \sigma^0, \beta^*, \sigma^*)$$

and

$$\lim_{j\to\infty} s_{n_j}(\beta^*, \sigma^0_{n_j}) = \bar{s}(\beta^*, \sigma^*, \beta^*, \sigma^*).$$

Since the inequality must be preserved in the limit we get

$$\bar{s}(\beta^0, \sigma^0, \beta^*, \sigma^*) \leq \bar{s}(\beta^*, \sigma^*, \beta^*, \sigma^*)$$

which implies, using condition (d) that

$$\|\beta^0 - \beta^*\|_E = 0$$

and

$$\|\sigma^0 - \sigma^*\| = 0.$$

Since every subsequence of $(\hat{\beta}_n, \hat{\sigma}_n)$ has a further subsequence satisfying this, then the sequence itself must have only one limit point and it must be (β^*, σ^*) **Q.E.D.**

Proof of Lemma 2:

Let

$$s_n(\beta, \sigma) = (1/n)\sum_{t=1}^{n} s(u_t, x_t, \beta, \sigma, \beta^*)$$

where

$$s(u, x, \beta, \sigma, \beta^*) = -I(x\beta + u \geq 0)g^1(x, \beta, \sigma) - I(x\beta + u > 0)g^2(x, \beta, \sigma)$$

with

$$g^1(x, \beta, \sigma) = g^1 = \log \Phi(\frac{x\beta}{\sigma})$$

and

$$g^2(x, \beta, \sigma) = g^2 = \log(1 - \Phi(\frac{x\beta}{\sigma}))$$

Following GN we will argue that the required $\bar{s}$ function is given by

$$\bar{s}(\beta, \sigma, \beta^*, \sigma^*) = \int_X \int_R s(u, x, \beta, \sigma, \beta^*)h^*(u)du d\mu(x)$$

where $h^*(u)$ is the true (normal) density of u, and show that it is continuous in (β, σ). First, we will show continuity. (In referring to the space of σ functions we will use Σ.) Define

$$A_1 = \sup_{X \times B \times \Sigma} |\frac{x\beta}{\sigma}| < \infty$$

First, consider

$$\bar{s}^1(\beta,\sigma,\beta^*,\sigma^*) = \int_X \int_R -I(x\beta^* + u \geq 0)g^1 h^*(u)dud\mu(x)$$

Now note that g^1 is continuous in (x,β,σ) in $\|(x,\beta,\sigma)\| = (\|x\|_E^2 + \|\beta\|_E^2 + \|\sigma\|^2)^{1/2}$ since convergence for σ implies uniform and hence pointwise convergence, so that the argument of $\log\Phi$ is continuous. Since $\log\Phi$ is a continuous function it is continuous in (x,β,σ). Now suppose we have a sequence (β_n,σ_n) converging to some point (β_0,σ_0) in the sense of the metric above. Then clearly

$$I(x\beta^* + u \geq 0)g^1(x,\beta_n,\sigma_n) \to I(x\beta^* + u \geq 0)g^1(x,\beta_0,\sigma_0)$$

pointwise. Since

$$|I(x\beta^* + u \geq 0)g^1| \leq |\log\Phi(-A_1)| < \infty$$

for any admissable sequence, and since

$$\int_X \int_R |\log\Phi(-A_1)|h^*(u)dud\mu = |\log\Phi(-A_1)| < \infty$$

then we have by the dominated convergence theorem (Royden (1987)), that $\bar{s}^1$ is continuous in (β,σ). A similar argument establishes continuity for the term corresponding to g^2. Next we verify the uniform convergence part. First, note that the continuity result holds when we replace I with any function χ which is continuous and bounded between 0 and 1. Hence, define the function $\chi(z)$ by $\chi(z) = 1$ if $z > \gamma$, $\chi(z) = 0$ if $z < 0$, and χ continuous on $[0,\gamma]$. Define the function s^γ to be the same as s but with I replaced by χ. This function will also be continuous and dominated by the function

$$-\chi(x\beta^* + u)\log\Phi(-A_1) - \chi(x\beta^* + u)\log(1 - \Phi(A_1))$$

Also, define the functions s_n^γ and $\bar{s}^\gamma$ in a similar way — *i.e.*, replace I with χ in the definitions. Now

$$\sup_{B\times\Sigma} |s_n(\beta,\sigma) - \bar{s}(\beta,\sigma,\beta^*,\sigma^*)| \leq \sup_{B\times\Sigma} |s_n^\gamma(\beta,\sigma) - \bar{s}^\gamma(\beta,\sigma,\beta^*,\sigma^*)|$$

$$+\sup_{B\times\Sigma}|s_n^\gamma(\beta,\gamma)-s_n(\beta,\sigma)|$$
$$+\sup_{B\times\Sigma}|\bar{s}^\gamma(\beta,\sigma,\beta^*,\sigma^*)-\bar{s}(\beta,\sigma,\beta^*,\sigma^*)|$$

Now consider each of these terms in turn By Theorem 1 of Burguette, Gallant, and Souza (1980) (hereafter BGS) the first converges to 0 as $n\to\infty$, since the s^γ functions are continuous. For the next, as in GN define the function α^γ by $0\le\alpha^\gamma\le 1$, $\alpha^\gamma(z)=0$ for $|z|>2\gamma$, $\alpha^\gamma(z)=1$ for $|z|<\gamma$ and it is continuous. Then

$$\begin{aligned}|s_n^\gamma(\beta,\sigma)-s_n(\beta,\sigma)| &\le \sum_{t=1}^{n}|\chi(x\beta^*+u)-I(x\beta^*+u\ge 0)||g^1+g^2|\\ &\le \sum_{t=1}^{n}\alpha^\gamma(x\beta^*+u)|\xi|\end{aligned}$$

where $\xi=\log\Phi(-A_1)+\log(1-\Phi(A_1))$. Again applying Theorem 1 of BGS we obtain that

$$\lim_{n\to\infty}\sup_{B\times\Sigma}|s_n^\gamma(\beta,\sigma)-s_n(\beta,\sigma)|\le\int_X\int_R\alpha^\gamma(x\beta^*+u)|\xi|h^*(u)dud\mu$$

A similar bound can be obtained for the third term directly, since it does not depend on n. But both bounds can be made arbitrarily small by picking small γ, since the function α^γ converges to zero except at the set of points $\{(x,u):x\beta^*+u=0\}$ which has probability zero with respect to $h^*(u)dud\mu$. Thus all three terms are zero in the limit. **Q.E.D.**

Proof of Lemma 3:

The proof follows along the same lines as Lemma 2 with minor modifications. In this case, we have

$$\begin{aligned}g^1(u,x,\beta,\sigma,\beta^*) &= \log\frac{1}{\sigma(x)}\phi\left(\frac{x(\beta^*-\beta)+u}{\sigma(x)}\right)\\ g^2(x,\beta,\sigma) &= \log\left(1-\Phi\left(\frac{x\beta}{\sigma(x)}\right)\right)\end{aligned}$$

Clearly, both g^1 and g^2 are continuous in the arguments. To obtain the desired continuity of the s function, all that is required is that we find dominating functions for the g

functions and show that they are integrable (with respect to $h^*(u)dud\mu$). Clearly

$$|g^2(x,\beta,\sigma)| \leq |\log(1-\Phi(A_1))|$$

as before. For g^1 we have

$$g^1 = -\log\sigma(x) - \frac{1}{2\sigma(x)^2}((x(\beta^* - \beta))^2 + u^2 + 2x(\beta^* - \beta)u)$$

Let

$$A_2 = \sup_{B\times X\times\Sigma} |\frac{x(\beta^* - \beta)}{\sigma(x)^2}| < \infty$$

and

$$c = \max(|\log\delta|, \log b) < \infty$$

Then we have

$$|g^1| \leq c + \frac{1}{2}(A_2^2 + \frac{u^2}{\delta^2} + 2A_2|u|)$$

which is integrable since u is normally distributed. The proof then follows along similar lines to that for Lemma 2. **Q.E.D.**

Proof of Lemma 4:

As before we require the bounding function which enables the method used in the proof of Lemma 2 to be used. Here

$$\begin{aligned} g^1 &= \log\Phi(\frac{x\beta_1 + \rho\frac{\sigma_1(x)}{\sigma_2(x)}(x(\beta_2^* - \beta_2) + u_2)}{\sigma_1(x)(1-\rho^2)^{1/2}}) \\ &+ \log\frac{1}{\sigma_2(x)}\phi(\frac{x(\beta_2^* - \beta_2) + u_2}{\sigma_2(x)}) \\ g^2 &= \log(1-\Phi(\frac{x\beta_1}{\sigma_1(x)}) \end{aligned}$$

These are clearly continuous in any argument. The bounding function for g^2 is as before. The second part of the g^1 function can be bounded as in the Tobit case. Therefore we

are left with only the log $\Phi(.)$ component. The Φ part is just the following integral. (The arguments of the σ functions are suppressed for convenience.)

$$\begin{aligned}
&\int_{-x\beta_1}^{\infty} \frac{1}{(2\pi)^{1/2}\sigma_1(1-\rho^2)^{1/2}} \times \\
&\qquad \exp(-\tfrac{1}{2\sigma_1^2(1-\rho^2)}(u_1 - \rho\tfrac{\sigma_1}{\sigma_2}(y_2 - x\beta_2))^2)du_1 \\
= &\qquad \exp(\tfrac{-\rho^2\frac{\sigma_1^2}{\sigma_2^2}(y_2-x\beta_2)^2}{2\sigma_1^2(1-\rho^2)}) \times \\
&\sigma_1(1-\rho^2)^{1/2}\textstyle\int_{\frac{-x\beta_1}{\sigma_1(1-\rho^2)^{1/2}}}^{\infty} \exp(\tfrac{2\rho\frac{\sigma_1}{\sigma_2}(y_2-x\beta_2)u_1^*}{2\sigma_1(1-\rho^2)^{1/2}})\phi(u_1^*)du_1^* \\
= &\quad \exp(\tfrac{-j^2}{2})\Phi(\tfrac{x\beta_1}{\sigma_1(1-\rho^2)^{1/2}})E(\exp(ju_1^*)|u_1^* \geq \tfrac{-x\beta_1}{\sigma_1(1-\rho^2)^{1/2}}) \\
\geq &\quad \exp(\tfrac{-j^2}{2})\Phi(\tfrac{x\beta_1}{\sigma_1(1-\rho^2)^{1/2}})\exp(jE(u_1^*|u_1^* \geq \tfrac{-x\beta_1}{\sigma_1(1-\rho^2)^{1/2}}))
\end{aligned}$$

where the inequality follows from the use of Jensen's inequality for the convex function $\exp(x)$, and j is defined by,

$$j = \frac{\rho\frac{\sigma_1}{\sigma_2}(y_2 - x\beta_2)}{\sigma_1(1-\rho^2)^{1/2}}$$

and

$$u_1^* = \frac{u_1}{\sigma_1(1-\rho^2)^{1/2}}$$

Taking logarithms and evaluating the expectation we get

$$0 \geq g^1 \geq -\frac{j^2}{2} + \log\Phi(\frac{x\beta_1}{\sigma_1(1-\rho^2)^{1/2}}) + j\frac{\phi(k)}{\Phi(k)}$$

where

$$k = \frac{x\beta_1}{\sigma_1(1-\rho^2)^{1/2}}$$

Observing the usual bounds on the parameters and x's then we can bound each of these functions from below. In particular,

$$\frac{-j^2}{2} \geq -\frac{(1-\epsilon^2)\frac{b^2}{\delta^2}}{\delta^2\epsilon^2}(A_3^2 + u_2^2 + 2A_3|u_2|)$$

where

$$A_3 = \sup_{X\times B}|x(\beta_2^* - \beta_2)| < \infty$$

This function is integrable since u is normally distributed. Next,

$$\log \Phi(k) \geq \log(\frac{-A_1}{\epsilon\delta}) > -\infty$$

Similarly,

$$j\frac{\phi(k)}{\Phi(k)} \geq -\frac{(1-\epsilon)b\delta}{\delta\epsilon}\frac{\phi(0)}{\Phi(\frac{A_1}{\epsilon})}$$

which is integrable since each part is finite and $|u_2|$ is integrable since u_2 is normal. **Q.E.D.**

Proof of Lemma 5:

Step 1.

First we show that $\frac{x\beta^*}{\sigma^*(x)}$ is identifiable. Suppose $\frac{x\beta}{\sigma(x)} \neq \frac{x\beta^*}{\sigma^*(x)}$ at some point x_0. Then there is an open ball $B(x_0, \vartheta)$ such that (assume without loss of generality that $\frac{x\beta}{\sigma} > \frac{x\beta^*}{\sigma}$)

$$\frac{x\beta^*}{\sigma^*(x)} \geq \frac{x\beta^*}{\sigma^*(x)} + \epsilon$$

for all x contained in the ball. This implies that

$$Pr(y=0; x, \beta, \sigma) \neq Pr(y=0; x, \beta^*, \sigma^*)$$

for all such x. Since $\mu(B) > 0$ by Assumption 5 we have that $\frac{x\beta^*}{\sigma^*}$ is identifiably unique.

Step 2.

Given the result of step 1 and the fact that $\sigma^*(x) > 0$ it is clear that $\text{sign}(x\beta^*)$ is identified. Suppose that

$$\text{sign}(xb) = \text{sign}(x\beta^*) a.e.[\mu]$$

and that $\|b\|_E = \|\beta^*\|_E = 1$ but that $b \neq \beta^*$. Then it must also be the case on the set $\{x_k \in (e,f), x_l \in C\}$. Assume without loss of generality that $\beta_k^* > 0$. Now consider the set L which is a subset of C, containing all x_l for which the sign equation is violated for some x_k. We want to show that for the sign condition to hold a.e. then L must have

zero measure. It will then follow that $b = \beta^{\star}$ which will provide us with a contradiction. Now for each $x_l \in L$ there is an $\hat{x}_k$ such that the sign equation is violated. Since xb is continuous in x_k for any x_l, this must also be true on $B(\hat{x}_k, \hat{\vartheta})$ for some $\vartheta > 0$. Denote the everywhere positive (on (e, f)) conditional density of x_k given x_l by $g(x_k|x_l)$. We have then that the probability that the sign condition is violated is given by

$$\int_L \int_{B(\hat{x}_k, \hat{\vartheta})} g(x_k|x_l) dx_k dP(x_l)$$

But

$$\int_{B(\hat{x}^k, \hat{\vartheta})} g(x_k|x_l) dx_l > 0$$

for all $x_l \in L$ by the above argument and Assumption 7. Thus this probability is only zero if $P(L) = 0$. Hence almost every $x_l \in C$ has the sign condition satisfied for all $x_k \in (e, f)$. Restricting attention to such x_l we get that

$$\text{sign}(x_k b_k + x_l b_l) = \text{sign}(x_k \beta_k^{\star} + x_l \beta_l^{\star})$$

for all $x_k \in (e, f)$. Clearly $b_k > 0$ must be true. Since this is true there must be an $\hat{x}_k(x_l) \in (e, f)$ such that

$$\hat{x}_k(x_l) b_k + x_l b_l = 0 \Rightarrow \hat{x}_k(x_l) = -\frac{x_l b_l}{b_k}$$

$$\hat{x}_k(x_l) \beta_k^{\star} + x_l \beta_l^{\star} = 0 \Rightarrow \hat{x}_k(x_l) = -\frac{x_l \beta_l^{\star}}{\beta_k^{\star}}$$

This implies that

$$\frac{x_l b_l}{b_k} = \frac{x_l \beta_l^{\star}}{\beta_k^{\star}}$$

for almost all $x_l \in C$. By the full rank condition we must have that

$$\frac{b_l}{b_k} = \frac{\beta_l^{\star}}{\beta_k^{\star}}$$

The condition $\|b\|_E^2 = \|\beta^*\|_E^2 = 1$ implies that

$$b_k^2 + \frac{b_k^2}{\beta_k^{*2}}\|\beta_l\|_E^2 = 1$$

which implies that $b_k = \beta_k^*$ and hence that $b = \beta^*$ which contradicts the assumption that they differ. **Q.E.D.**

Proof of Lemma 6:

We can use the same argument as in Step 1 of Lemma 5 to show that $\frac{x\beta^*}{\sigma^*(x)}$ is identified. Now suppose that there is a point x_0 such that $\sigma(x_0) \neq \sigma^*(x_0)$ then it must also be true on $B(x_0, \vartheta)$ for some $\vartheta > 0$. Assume it is greater (without loss of generality), then for all $x \in B$

$$\sigma(x) \geq \sigma^*(x) + \varepsilon$$

for some $\varepsilon > 0$. Note that $\mu(B) > 0$. This implies that

$$\frac{1}{\sigma(x)}\phi(\frac{y - x\beta}{\sigma(x)}) = \frac{1}{\sigma^*(x)}\phi(\frac{y - x\beta^*}{\sigma^*(x)})$$

at at most two points (using properties of the normal densities). Thus they differ almost everywhere for $y > 0$ and hence there are clearly events regarding y which have different probabilites under σ compared to that under σ^*. By a similar argument we obtain $xb = x\beta^*$ for almost all x. By the full rank condition we must have that $b = \beta^*$. **Q.E.D.**

Proof of Lemma 7:

Using a similar argument to that in Lemma 5 we can show that $\sigma^*(x)$ is identified and β^* is identified to scale. Thus clearly we have $\frac{x\beta^*}{\sigma^*(x)}$ identified also. (This is using only the observations for which $y_1 = 0$.) Next, consider the other part of the density function given by

$$\frac{1}{\sigma_2^*(x)}\phi(\frac{y_2 - x\beta_2}{\sigma_2(x)})\Phi(\frac{\frac{x\beta_1}{\sigma_1(x)} + \rho(\frac{y_2 - x\beta_2}{\sigma_2(x)})}{(1-\rho^2)^{1/2}})$$

Hence both parts are clearly continuous in y_2, however the first part is not monotone in y_2 unlike the second part (again using properties of the normal density function). Now

suppose that $\sigma_2(x_0) \neq \sigma_2^*(x_0)$, at some point $x_0 \in X$, then it must also be the case that they differ on an open ball B around the point and that $\mu(B) > 0$. By a similar argument to that used in Lemma 6, the normal density parts under these different σ_2 functions will be equal at two different points (note that y_2 is not restricted to be positive — the addition of this restriction would turn the model into Cragg's (1971) double hurdle model). However the Φ part is monotone in y_2 so evaluated at different σ_2 functions will only cross once. Together these two facts ensure that the densities evaluated at the different σ_2 fuctions will differ on a nontrivial set of y_2. Hence σ_2 is identified. A similar argument can be used to ensure that the other parameters , ρ^* and β_2^* are identified. **Q.E.D.**

Proof of Lemma 8

This follows from the use of Jensen's inequality and Lemmata 5-7. In particular suppose we have

$$\bar{s}(\beta^0, \sigma^0, \beta^*, \beta^*) \leq \bar{s}(\beta^*, \sigma^*, \beta^*, \sigma^*)$$

Now by Jensen's inequality we have

$$\begin{aligned}
&- \int_X \int_{R^2} s(u, x, \beta^*, \sigma^*, \beta^*) - s(u, x, \beta^0, \sigma^0, \beta^*) h^*(u) du d\mu(x) \\
&\geq - \log \int_X \int_{R^2} \exp(s(u, x, \beta^*, \sigma^*, \beta^*) - s(u, x, \beta^0, \sigma^0, \beta^*)) h^*(u) du d\mu(x) \\
&= - \log \int_X \int_Z \exp(- \log p(z|x, \beta^*, \sigma^*) + \log p(z|x, \beta^0, \sigma^0)) \\
&\qquad p(z|x, \beta^*, \sigma^*) d\nu(z) d\mu(x) \\
&= - \log \int_X \int_Z p(z|x, \beta^0, \sigma^0) d\nu(z) d\mu(x) \\
&= - \log 1 = 0
\end{aligned}$$

Hence we must have that

$$\bar{s}(\beta^0, \sigma^0, \beta^*, \sigma^*) \geq \bar{s}(\beta^*, \sigma^*, \beta^*, \sigma^*)$$

with the equality holding only if we have

$$s(u, x, \beta^0, \sigma^0, \beta^*) - s(u, x, \beta^*, \sigma^*, \beta^*)$$

is constant a.e. $h^*(u)dud\mu(x)$. Together we have then that

$$s(u, x, \beta^0, \sigma^0, \beta^*) = s(u, x, \beta^*, \sigma^*, \beta^*)$$

a.e. $h^*(u)dud\mu(x)$. But by Lemmata 5-7 this is only true if

$$\|\beta^0 - \beta^*\|_E = 0$$

and

$$\|\sigma^0 - \sigma^*\| = 0$$

Q.E.D.

Appendix B

Proofs for Chapter 4

Proof of Theorem 1:

By Assumption 2.1 and Assumption 2.2 we have

$$\frac{1}{N_1}(Z - T_3)'(Z - T_3) = \bar{A}_1 + o_p(1)$$

and using Assumption 2.3 we can apply the Liapunov Central Limit Theorem for independent non-identically distributed (i.n.i.d) random variables so that

$$\frac{1}{\sqrt{N_1}}(Z - T_3)'D_3^{-1}\xi \rightarrow N(0, \text{plim}(\frac{1}{\sqrt{N_1}}(Z - T_3)'D_3\Gamma D_3(Z - T_3)))$$

and the result follows.**Q.E.D.**

Proof of Lemma 1:

The proof is similar to that of Andrews and Whang (1989). We condition on x in this so all expectations are conditional. Write

$$MSE(\hat{g}, g) = \frac{1}{N}E\|\hat{g} - g\|^2 = \frac{1}{N}E\|\hat{g} - P_x y\|^2$$

$$= \frac{1}{N}\|\hat{g} - P_x g\|^2 + \frac{1}{N}E\|P_x u\|^2$$

using the parallelogram law in a Hilbert space. Now,

$$\frac{1}{N}E\|P_x u\|^2 = \frac{1}{N}E(\text{tr}P_x uu'P_x)^2 = \frac{1}{N}E(\text{tr}P_x uu')$$

$$= \frac{1}{N}\text{tr}P_x\Psi_u \leq \sup_t \sigma_t^2 \frac{K(N)}{N}$$

by Assumption 3.1 and since P_x is a projection matrix with $K(N)$ regressors. Next denoting the remainder term of the difference between the approximation in Assumption 3.2, denoted g_K, and the true function, g, by g_K^r, and noting that

$$P_x g_K = g_K$$

then we have that we have that,

$$\frac{1}{N}\|\hat{g} - P_x g\|^2 = \frac{1}{N}\|g - P_x g_K^r + P_x g_K\|^2$$

$$= \frac{1}{N}\|(I - P_x)g_K^r\|^2 = \frac{1}{N}\|g_K^r\|^2 \leq \|g_K^r\|_{0,\infty,X}^2$$

using facts about projections on Hilbert spaces. Therefore, we have that

$$MSE(g, \hat{g}) \leq \sup_t \sigma_t^2 \frac{K(N)}{N} + \|g_K^r\|_{0,\infty,X}^2$$

Then if the conditions of (i) hold then

$$MSE(g, \hat{g}) \to 0$$

by Assumption 3.2. If $K(N) = O(N^r)$ as in (ii) then

$$\begin{aligned} MSE(g, \hat{g}) &= O(N^{r-1}) + O(N^{-2\alpha r} K(N)^{2\alpha} \|g_K^r\|_{0,\infty,X}^2 \\ &= O(N^{r-1}) + O(N^{-2\alpha r}) \end{aligned}$$

by Assumption 3.2, so the result holds. For (iii) first note that by Assumption 3.2 we can choose an $\alpha < \frac{S(g)}{d}$ such that

$$K(N)^\alpha \|g_K^r\|_{0,\infty,X} \to 0$$

and since $K(N) = O(N^{q-\gamma})$ then

$$N^{\alpha(q-\gamma)} \|g_K^r\|_{0,\infty,X} \to 0$$

and the exponent can be made bigger than 1/2 provided that $\frac{S(g)}{d} > \frac{1}{q}$ and $0 < \gamma < q - \frac{d}{2S(g)}$. Finally, we must also have that if $q < 1/2$ then

$$1 - r = 1 - (q - \gamma) > 1/2$$

and $2\alpha r$ can be made bigger than 1 using the above mentioned α and γ. **Q.E.D.**

Proof of Theorem 2:

We can write

$$\sqrt{N}(\frac{1}{N}c'P_x y - \frac{1}{N}c'g) = (\frac{1}{\sqrt{N}}c'P_x g + \frac{1}{\sqrt{N}}c'P_x u - \frac{1}{\sqrt{N}}c'g)$$

$$= \frac{1}{\sqrt{N}}c'P_x u - \frac{1}{\sqrt{N}}c'(I - P_x)g$$

Consider the second term first and condition on x so that the results will hold for every possible sequence of x

$$\frac{1}{\sqrt{N}}|c'(I - P_x)g| \leq \sqrt{N}(\frac{c'c}{N})^{1/2}(\frac{1}{N}\|(I - P_x)g\|^2)^{1/2}$$

$$\leq O(1)\sqrt{N}\|g_K^r\|_{0,\infty,X} \to 0$$

using the fact that c_t is a bounded sequence and (i) and (ii) of the Theorem which allow the application of Lemma 1(iii).

Next, consider

$$\frac{1}{\sqrt{N}}c'u - \frac{1}{\sqrt{N}}c'P_x u = \frac{1}{\sqrt{N}}c'(I - P_x)u$$

Suppose we condition on x again (and hence on c) and find the variance of this, then

$$E((\frac{1}{\sqrt{N}}c'(I - P_x)u)^2) = \frac{1}{N}c'(I - P_x)\Psi_u(I - P_x)c$$

$$\leq \sup_t \sigma_t^2 \frac{1}{N}\|(I - P_x)c\|^2 \leq \sup_t \sigma_t^2 \|c_K^r\|_{0,\infty,X}^2 \to 0$$

using (iii) and (ii) by Lemma 1, so by Chebyshev's Inequality

$$\frac{1}{\sqrt{N}}c'(I - P_x)u = o_p(1).$$

Finally,

$$\frac{1}{\sqrt{N}}c'u \rightarrow N(0, \text{plim}(\frac{c'\Psi_u c}{N}))$$

using Liapunov's Central Limit Theorem, which is applicable since Assumption 3.1 implies that

$$E(|u_t|^{2+\delta}) \leq \Delta < \infty$$

and c_t are bounded. **Q.E.D.**

Proof of Corollary 1:

Follows simply from the proof of Theorem 2 and the fact that $P_x\iota = \iota$. **Q.E.D.**

Proof of Theorem 3:

The result will follow from the following two results

$$A_1^N = \bar{A}_1^N + o_p(1)$$

where

$$A_1^N = \frac{1}{N_1}(Z - P_{x_3}Z)'(Z - P_{x_3}Z)$$

and

$$\bar{A}_1^N = \frac{1}{N_1}(Z - T_3)'(Z - T_3)$$

and

$$\frac{1}{\sqrt{N_1}}(Z - P_{x_3}Z)'(g + \eta) = \frac{1}{\sqrt{N_1}}(Z - T_3)'\eta + o_p(1)$$

with $\eta = \frac{\xi}{\lambda} = D_3^{-1}\xi$. For the first write

$$Z - P_{x_3}Z = Z - T_3 + T_3 - P_{x_3}Z$$

so that

$$A_1^N = \bar{A}_1^N + \frac{1}{N_1}(Z - T_3)'(T_3 - P_{x_3}) + \frac{1}{N_1}(T_3 - P_{x_3}Z)'(Z - T_3)$$
$$+\frac{1}{N_1}(T_3 - P_{x_3}Z)'(T_3 - P_{x_3}Z)$$

Using the Cauchy Inequality, it suffices to show that for each i

$$\frac{1}{N_1}\|P_{x_3}z_i - \tau_i\|^2 = o_p(1)$$

and

$$\frac{1}{N_1}\|z_i - \tau_i\|^2 = O_p(1)$$

which follow using (ii) (iii), by Lemma 1, and Assumption 2.3.

Next, consider

$$\frac{1}{\sqrt{N_1}}(Z - P_{x_3}Z)'g = \frac{1}{\sqrt{N_1}}Z'(g - P_{x_3}g).$$

By the Cauchy Inequality

$$\frac{1}{\sqrt{N_1}}|z_i(g - P_{x_3}g)| \leq \sqrt{N_1}(\frac{z_i'z_i}{N_1})^{1/2}(\frac{1}{N_1}\|g - P_{x_3}g\|^2)^{1/2}$$

$$\leq (\frac{z_i'z_i}{N_1})^{1/2}\sqrt{N_1}\|g_K^r\|_{0,\infty,X} \to 0$$

by Assumption 2.3, (ii) and (iii) and Lemma 1(iii). Finally, consider

$$\frac{1}{\sqrt{N_1}}(Z - P_{x_3}Z)'\eta = \frac{1}{\sqrt{N_1}}(Z - T_3)'\eta + \frac{1}{\sqrt{N_1}}(T_3 - P_{x_3}Z)'\eta$$

Thus, the result will follow provided for each i,

$$\frac{1}{\sqrt{N_1}}(\tau_i - P_{x_3}z_i)'\eta = o_p(1).$$

To show this it will suffice that

$$E((\frac{1}{\sqrt{N_1}}(\tau_i - P_{x_3}z_i)'\eta)^2) \to 0$$

by Chebyshev's Inequality. To compute this expectation first condition on x so that

$$E((\frac{1}{\sqrt{N_1}}(\tau_i - P_{x_3}z_i)'\eta)^2)$$

$$= E(\frac{1}{N_1}(\tau_i - P_{x_3}z_i)'E(\eta\eta'|x)(\tau_i - P_{x_3}z_i))$$

$$= E(\frac{1}{N_1}(\tau_i - P_{x_3}z_i)'D_3^{-1}\Gamma D_3^{-1}(\tau_i - P_{x_3}z_i))$$

$$\leq \sup_t V(\xi_t|x_t) \sup_t |\frac{1}{\lambda_t^2}| E\frac{1}{N_1}\|\tau_i - P_{x_3}z_i\|^2 \to 0$$

given Assumption 2.3, (ii) and (iii) (which allow the application of Lemma 1). Therefore, by Theorem 1 the result holds. **Q.E.D.**

Proof of Lemma 2:

Using a second order mean value expansion of $\hat{\lambda}_t$ about $\lambda(l(x_{1t}))$, which is valid for observations with $\hat{l} \in [\delta/2, 1-\delta/2]$ and defining

$$I_t = I(\hat{l}(x_{1t}) < \delta/2, (\hat{l}(x_{1t}) > 1 - \delta/2)$$

we have

$$\hat{\lambda}_t - \lambda_t = d_1(l(x_{1t}))(\hat{l}(x_{1t}) - l(x_{1t})) + (1/2)d_2(l^*)(\hat{l}(x_{1t}) - l(x_{1t}))^2$$

for some l^* lying between l and $\hat{l}$. Using this we then have

$$\frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})(\hat{\lambda}_t - \lambda_t) = \frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})d_1(l(x_{1t}))(\hat{l}(x_{1t}) - l(x_{1t}))$$

$$+\frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})I_t(\hat{\lambda}_t - \lambda_t)$$

$$-\frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})d_1(l(x_{1t}))I_t(\hat{l}(x_{1t}) - l(x_{1t}))$$

$$+\frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})d_2(l^*(x_{1t}))I_t(\hat{l}(x_{1t}) - l(x_{1t}))^2$$

Since $c(x_{1t})$, $\hat{\lambda}_t$ are both bounded sequences, the second term on the right of this satisfies,

$$E|\frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})I_t|$$

$$\leq C\frac{1}{\sqrt{N}}\sum_{t=1}^{N} P(\hat{l}(x_{1t}) < \delta/2, \hat{l}(x_{1t}) > 1 - \delta/2)$$

$$\leq C\frac{1}{\sqrt{N}}\sum_{t=1}^{N} P(|\hat{l}(x_{1t}) - l(x_{1t})| > \delta/2)$$

$$\leq \frac{2C}{\delta}\frac{1}{\sqrt{N}}\sum_{t=1}^{N} E((\hat{l}(x_{1t}) - l(x_{1t}))^2)$$

$$\leq \frac{2C}{\delta}\sqrt{N}\frac{1}{N}E\|\hat{l} - l\|^2 \to 0$$

where convergence follows from the application of Lemma 1(iii), given (i) and (ii), which imply that

$$\sqrt{N}E\frac{1}{N}\|\hat{l} - l\|^2 \to 0.$$

Next, using the Cauchy and Chebyshev Inequalities, and the boundedness of c and d_1

$$E|\frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})d_1(l(x_{1t}))I_t(\hat{l}(x_{1t}) - l(x_{1t}))|$$

$$\leq C\sqrt{N}(E\frac{\sum_{t=1}^{N} I_t^2}{N})^{1/2}(E\frac{1}{N}\|\hat{l} - l\|^2)^{1/2}$$

$$\leq C\sqrt{N}(\frac{\sum_{t=1}^{N} P(\hat{l}(x_{1t}) < \delta/2, \hat{l}(x_{1t}) > 1 - \delta/2)}{N})^{1/2}(E\frac{1}{N}\|\hat{l} - l\|^2)^{1/2}$$

$$\leq \frac{2C}{\delta}\sqrt{N}(E\frac{1}{N}\|\hat{l} - l\|^2)^{1/2}(E\frac{1}{N}\|\hat{l} - l\|^2)^{1/2}$$

$$= \frac{2C}{\delta}\sqrt{N}E\frac{1}{N}\|\hat{l} - l\|^2 \to 0$$

as above. Similarly since d_2 are bounded we have for the final term

$$E|\frac{1}{\sqrt{N}}\sum_{t=1}^{N} c(x_{1t})d_2(l^{*}(x_{1t}))I_t(\hat{l}(x_{1t}) - l(x_{1t}))^2|$$

$$\leq C\sqrt{N}E\frac{1}{N}\|\hat{l} - l\|^2 \to 0$$

and the result follows by applying Markovs Inequality to each of these terms to show that all but the first term are $o_p(1)$. **Q.E.D.**

Proof of Lemma 3:

This result follows from Lemma 2 and Theorem 2, since d_1 is a smooth function and $c(x_{1t})$ satisfies $S(c) > 0$ so the product will also have at least one derivative bounded and also be bounded. **Q.E.D.**

Proof of Theorem 4:

Consider first

$$\tilde{A}_N^1 = \frac{1}{N_1}(\hat{Z} - P_{x_3}\hat{Z})'(\hat{Z} - P_{x_3}\hat{Z})$$

and show that

$$\tilde{A}_1^N = \bar{A}_1^N + o_p(1) \tag{B.105}$$

Since we can write

$$\hat{Z} - P_{x_3}\hat{Z} = Z - T_3 + T_3 - P_{x_3}T_3 + P_{x_3}T_3 - P_{x_3}Z \tag{B.106}$$

$$+P_{x_3}Z - P_{x_3}\hat{Z} + \hat{Z} - Z$$

then by the Cauchy Inequality applied element by element to the product it suffices to show that

$$\frac{1}{N_1}\|\tau_i - P_{x_3}\tau_i\|^2 = o_p(1) \tag{B.107}$$

$$\frac{1}{N_1}\|P_{x_3}(\tau_i - z_i)\|^2 = o_p(1) \tag{B.108}$$

$$\frac{1}{N_1}\|z_i - \tau_i\|^2 = O_p(1) \tag{B.109}$$

$$\frac{1}{N_1}\|P_{x_3}(\hat{z}_i - z_i)\|^2 = o_p(1) \tag{B.110}$$

$$\frac{1}{N_1}\|\hat{z}_i - z_i\|^2 = o_p(1). \tag{B.111}$$

For (B.107) for any sequence of x,

$$\frac{1}{N_1}\|\tau_i - P_{x_3}\tau_i\|^2 \leq \|\tau_{iK_1}^{\tau}\|_{0,\infty,X}^2 \to 0$$

given (ii) and (iii), by Lemma 1. For (B.108), taking expectations first conditional on x we get

$$E\frac{1}{N_1}\|P_{x_3}(\tau_i - z_i)\|^2 = E\frac{1}{N_1}(\mathrm{tr}P_{x_3}u_iu_i')$$
$$= E\frac{1}{N_1}(\mathrm{tr}P_{x_3}E(u_iu_i'|x))$$
$$\leq C\frac{K^3(N_1)}{N_1} \to 0$$

using Assumption 4.2, and the fact that P_{x_3} is a projection matrix with $K^3(N_1)$ regressors, so (B.108) follows from Markov's Inequality. Next, given Assumption 3.2 and (iii) (B.109) is true. For (B.110) we have

$$\frac{1}{N_1}\|P_{x_3}(\hat{z}_i - z_i)\|^2 = \frac{1}{N_1}(\hat{z}_i - z_i)'P_{x_3}(\hat{z}_i - z_i)$$
$$\leq \frac{1}{N_1}(\hat{z}_i - z_i)'(\hat{z}_i - z_i)$$

due to the fact that P_{x_3} is a projection matrix, so that (B.110) will follow from (B.111). For (B.111),

$$\frac{1}{N_1}(\hat{z}_i - z_i)'(\hat{z}_i - z_i) = \frac{1}{N_1}\sum_{t=1}^{N_1} x_{it}^2(\frac{1}{\hat{\lambda}_t^2} - \frac{1}{\lambda_t^2})$$
$$= \frac{1}{N_1}\sum_{t=1}^{N_1} x_{it}^2\frac{(\lambda_t - \hat{\lambda}_t)(\lambda_t + \hat{\lambda}_t)}{\hat{\lambda}_t^2\lambda_t^2}$$
$$\leq C\frac{1}{N_1}\|x_i\|\|\hat{\lambda} - \lambda\|$$

since, λ_t and $\hat{\lambda}_t$ are bounded variables for all t. By the mean value theorem applied in Theorem 2,

$$\lambda_t - \hat{\lambda}_t = -d_1(l^*(x_1t))(l(x_{1t}) - \hat{l}(x_{1t}))$$

for observations where $I_t = 0$ (where this is defined in the proof of Theorem 2), so that

$$\frac{1}{N_1}\|\hat{\lambda} - \lambda\|^2 \leq \frac{N}{N_1}\frac{1}{N}\sum_{t=1}^{N} d_1(l^*(x_{1t}))^2(1 - I_t)(\hat{l}(x_{1t}) - l(x_{1t}))^2$$

$$+C\frac{N}{N_1}\frac{1}{N}\sum_{t=1}^{N} I_t$$

and since d_1 are bounded, and by (iv) $\frac{N}{N_1} \to \frac{1}{p}$, the first term on the right converges to 0 in probability provided that $\dot{l}$ converges to l in MSE, as does the second given the proof in Lemma 2 (using Chebyshev's Inequality). But this is guaranteed by (i). So (B.105) follows since $\frac{1}{\sqrt{N_1}}\|x_i\| = O_p(1)$.

Next, consider

$$\frac{1}{\sqrt{N_1}}(\hat{Z} - P_{x_3}\hat{Z})'g = \frac{1}{\sqrt{N_1}}\hat{Z}'(g - P_{x_3}g) \tag{B.112}$$

To show that this is $o_p(1)$, it suffices that for each i

$$\frac{1}{\sqrt{N_1}}\hat{z}_i(g - P_{x_3}g) = o_p(1)$$

By the Cauchy inequality

$$\frac{1}{\sqrt{N_1}}|\hat{z}_i(g - P_{x_3}g)| \leq (\frac{1}{N_1}\|\hat{z}_i\|^2)^{1/2}\sqrt{N_1}(\frac{1}{N}\|(I - P_{x_3})g\|^2)^{1/2}$$

$$\leq O_p(1)\sqrt{N_1}\|g^r_{K^3(N_1)}\| \to 0$$

using Assumption 2.3 and Lemma 1, which applies given (iii). So that (B.112) is $o_p(1)$.

Next consider,

$$\frac{1}{\sqrt{N_1}}(\hat{Z} - P_{x_3}\hat{Z})'\hat{D}_3\xi.$$

Using (B.106) we show the following hold for each i

$$\frac{1}{\sqrt{N_1}}(z_i - \tau_i)'\hat{D}_3^{-1}\xi = \frac{1}{\sqrt{N_1}}(z_i - \tau_i)'D_3^{-1}\xi + o_p(1) \tag{B.113}$$

$$\frac{1}{\sqrt{N_1}}(\tau_i - P_{x_3}\tau_i)'\hat{D}_3^{-1}\xi = o_p(1) \tag{B.114}$$

$$\frac{1}{\sqrt{N_1}}(P_{x_3}(\tau_i - z_i))'\hat{D}_3^{-1}\xi = o_p(1) \tag{B.115}$$

$$\frac{1}{\sqrt{N_1}}(P_{x_3}(z_i - \hat{z}_i))'\hat{D}_3^{-1}\xi = o_p(1) \tag{B.116}$$

$$\frac{1}{\sqrt{N_1}}(\hat{z}_i - z_i)'\hat{D}_3^{-1}\xi = o_p(1) \tag{B.117}$$

To show all but (B.113) we consider the expectation of the term squared, conditional on y_1 and x and show that each of these converges to zero, so then by Chebyshev's Inequality the results follow. First, consider (B.114) taking expectation first conditional on y_1 and x and then over these we get,

$$E((\frac{1}{\sqrt{N_1}}(\tau_i - P_{x_3}\tau_i)'\hat{D}_3^{-1}\xi)^2) =$$

$$= E(\frac{1}{N_1}(\tau_i - P_{x_3}\tau_i)'\hat{D}_3^{-1}E(\xi\xi'|x, y_1)\hat{D}_3^{-1}(\tau_i - P_{x_3}\tau_i))$$

$$= E(\frac{1}{N_1}(\tau_i - P_{x_3}\tau_i)'\hat{D}_3^{-1}\Gamma\hat{D}_3^{-1}(\tau_i - P_{x_3}\tau_i))$$

$$\leq E(\frac{1}{N_1}\|\tau_i - P_{x_3}\tau_i\|^2 \to 0$$

using Assumption 2.3, (ii), (iii) and the result of Lemma 1. Similarly, for (B.115) we have

$$E((\frac{1}{\sqrt{N_1}}(P_{x_3}(\tau_i - z_i))'\hat{D}_3^{-1}\xi)^2)$$

$$\leq CE(\frac{1}{N_1}(z_i - \tau_i)'P_{x_3}(z_i - \tau_i))$$

$$\leq C\frac{K^3(N_1)}{N_1} \to 0$$

by Assumptions 2.3, 4.2 , (iii) and (ii). For (B.116)

$$E((\frac{1}{\sqrt{N_1}}(P_{x_3}(z_i - \hat{z}_i))'\hat{D}_3^{-1}\xi)^2)$$

$$\leq CE(\frac{1}{N_1}(z_i - \hat{z}_i)'P_{x_3}(z_i - \hat{z}_i))$$

$$\leq CE(\frac{1}{N_1}\|z_i - \hat{z}_i\|^2) \to 0$$

using Assumption 2.3, and a previous result. For (B.117),

$$E((\frac{1}{\sqrt{N_1}}(\hat{z}_i - z_i))'\hat{D}_3^{-1}\xi)^2)$$

$$\leq CE(\frac{1}{N_1}\|\hat{z}_i - z_i\|^2) \to 0$$

as before. Finally, for (B.113),

$$E((\frac{1}{\sqrt{N_1}}(z_i - \tau_i)'(\hat{D}_3 - D_3^{-1})\xi)^2)$$

$$\leq CE(\frac{1}{N_1}\sum_{t=1}^{N_1}(z_{it} - \tau_i(x_{3t}))^2(\frac{1}{\hat{\lambda}_t} - \frac{1}{\lambda_t})^2$$

$$\leq CE(\frac{1}{N_1}\sum_{t=1}^{N_1}(\lambda_t - \hat{\lambda}_t)^2) \to 0$$

using the boundedness of $z_{it} - \tau_i(x_{3t})$ and the MSE convergence of $\hat{\lambda}$ to λ which is shown above. Therefore, (B.113) holds and we have from Theorem 1 that since

$$\frac{1}{\sqrt{N_1}}B_1^N\xi \to N(0, \text{plim}(B_1^{N'}\Gamma B_1^N))$$

where we let $B_1^N = \frac{1}{\sqrt{N_1}}(Z - T_3)'D_3^{-1}$, then since

$$\frac{1}{\sqrt{N_1}}(\hat{Z} - P_{x_3}\hat{Z})'\hat{D}_3\xi = B_1^N\xi + o_p(1)$$

it is also true that

$$\frac{1}{\sqrt{N_1}}(\hat{Z} - P_{x_3}\hat{Z})'\hat{D}_3\xi \to N(0, \text{plim}(B_1^{N'}\Gamma B_1^N)).$$

Finally, we show that

$$\frac{1}{\sqrt{N_1}}(\hat{Z} - P_{x_3}\hat{Z})'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} \to N(0, S_2).$$

We do this by breaking up the term using (B.106) and show the following,

$$\frac{1}{\sqrt{N_1}}(z_i - \tau_i)'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} = \frac{1}{\sqrt{N_1}}(z_i - \tau_i)'g\frac{(\lambda - \hat{\lambda})}{\lambda} + o_p(1) \tag{B.118}$$

$$\frac{1}{\sqrt{N_1}}(\tau_i - P_{x_3}\tau_i)'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} = o_p(1) \tag{B.119}$$

$$\frac{1}{\sqrt{N_1}}(P_{x_3}(\tau_i - z_i))'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} = o_p(1) \tag{B.120}$$

$$\frac{1}{\sqrt{N_1}}(P_{x_3}(z_i - \hat{z}_i))'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} = o_p(1) \tag{B.121}$$

$$\frac{1}{\sqrt{N_1}}(\hat{z}_i - z_i)'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}} = o_p(1). \tag{B.122}$$

Consider first (B.119),

$$E|\frac{1}{\sqrt{N_1}}(\tau_i - P_{x_3}\tau_i)'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}}|$$

$$\leq (\sqrt{N_1}E\frac{1}{N_1}\|\tau_i - P_{x_3}\tau_i\|^2)^{1/2}(\sqrt{N_1}E\frac{1}{N_1}\sum_{t=1}^{N_1}\frac{g(x_{3t})^2}{\hat{\lambda}_t^2}(\lambda_t - \hat{\lambda}_t)^2)^{1/2}$$

$$\leq C(\sqrt{N_1}E\frac{1}{N_1}\|\tau_i - P_{x_3}\tau_i\|^2)^{1/2}(\sqrt{N_1}E\frac{1}{N}\|\lambda - \hat{\lambda}\|^2)^{1/2}$$

$$\to 0$$

using (i), (ii), (iii), (iv) by Lemma 1, and boundedness of $\hat{\lambda}_t$. For (B.120),

$$E|\frac{1}{\sqrt{N_1}}(P_{x_3}(\tau_i - z_i))'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}}|$$

$$\leq (\sqrt{N_1}E(\frac{1}{N_1}u_i'P_{x_3}u_i))^{1/2}(\sqrt{N_1}E\frac{1}{N_1}\sum_{t=1}^{N_1}\frac{g(x_{3t})^2}{\hat{\lambda}_t^2}(\lambda_t - \hat{\lambda}_t)^2)^{1/2}$$

$$\leq C(\sqrt{N_1}\frac{K^3(N_1)}{N_1})^{1/2}(\sqrt{N_1}E\frac{1}{N}\|\lambda - \hat{\lambda}\|^2)^{1/2}$$

$$\to 0$$

by (ii), (i) and from above. Next for (B.121),

$$E|\frac{1}{\sqrt{N_1}}(P_{x_3}(z_i - \hat{z}_i))'g\frac{(\lambda - \hat{\lambda})}{\hat{\lambda}}|$$

$$\leq C(\sqrt{N_1}E\frac{1}{N_1}((z_i - \hat{z}_i)'P_{x_3}(z_i - \hat{z}_i)))^{1/2}(\sqrt{N_1}E\frac{1}{N}\|\lambda - \hat{\lambda}\|^2)^{1/2}$$

$$\leq C(\sqrt{N_1}E\frac{1}{N_1}((z_i - \hat{z}_i)'(z_i - \hat{z}_i)))^{1/2}(\sqrt{N_1}E\frac{1}{N}\|\lambda - \hat{\lambda}\|^2)^{1/2}$$

$$\leq C(\sqrt{N_1}E\frac{1}{N_1}\sum_{t=1}^{N_1}x_{it}^2\frac{(\lambda_t-\hat{\lambda}_t)^2}{\hat{\lambda}_t^2\lambda_t^2}(\sqrt{N_1}E\frac{1}{N}\|\lambda-\hat{\lambda}\|^2)^{1/2}$$

$$\leq C\sqrt{N_1}E\frac{1}{N}\|\lambda-\hat{\lambda}\|^2\to 0$$

using boundedness of x_{it}, λ_t and the above result. Similarly (B.122) holds. Finally for (B.118) we have

$$E|\frac{1}{\sqrt{N_1}}(z_i-\tau_i)'(g\frac{(\lambda-\hat{\lambda})}{\hat{\lambda}}-g\frac{(\lambda-\hat{\lambda})}{\lambda}|$$

$$=\frac{1}{\sqrt{N_1}}E|\sum_{t=1}^{N_1}(z_{it}-\tau_i(x_{3t}))\frac{g(x_{3t})}{\lambda\hat{\lambda}_t^2}(\lambda_t-\hat{\lambda}_t)^2|$$

$$\leq C\frac{1}{\sqrt{N_1}}E\|\lambda-\hat{\lambda}\|^2\to 0$$

as before. Therefore,

$$\frac{1}{\sqrt{N_1}}(\hat{Z}-P_{x_3}\hat{Z})'g\frac{(\lambda-\hat{\lambda})}{\hat{\lambda}}=\frac{1}{\sqrt{N_1}}(Z-T_3)'g\frac{(\lambda-\hat{\lambda})}{\lambda}+o_p(1)$$

Using the definition of y_1 we can write

$$\frac{1}{\sqrt{N_1}}(z_i-\tau_i)'g\frac{(\lambda-\hat{\lambda})}{\hat{\lambda}}$$

$$=\frac{1}{\sqrt{N_1}}\sum_{t=1}^{N}y_{1t}(z_{it}-\tau_i(x_{3t}))\frac{g(x_{3t})}{\lambda_t}(\lambda_t-\hat{\lambda}_t)$$

$$=\frac{\sqrt{N}}{\sqrt{N_1}}\frac{1}{\sqrt{N}}\sum_{t=1}^{N}y_{1t}(z_{it}-\tau_i(x_{3t}))\frac{g(x_{3t})}{\lambda_t}d_1(l(x_{1t}))(l(x_{1t})-\hat{l}(x_{1t}))+o_p(1)$$

using the result of Lemma 2. Letting $Y_1=\text{diag}\{y_{1t}\}$ this can be written as

$$-\frac{\sqrt{N}}{\sqrt{N_1}}\frac{1}{\sqrt{N}}(z_i-\tau_i)'GD_3^{-1}D_1Y_1(P_{x_1}y_1-l)$$

$$=-\frac{\sqrt{N}}{\sqrt{N_1}}\frac{1}{\sqrt{N}}(z_i-\tau_i)'GD_3^{-1}D_1Y_1P_{x_1}\epsilon$$

$$-\frac{\sqrt{N}}{\sqrt{N_1}}\frac{1}{\sqrt{N}}(z_i-\tau_i)'GD_3^{-1}D_1Y_1(P_{x_1}l-l)$$

Since (iv) holds we ignore the factor $\frac{\sqrt{N}}{\sqrt{N_1}}$ for the moment to simplify the notation. Consider now the second term on the right,

$$|\frac{\sqrt{N}}{\sqrt{N_1}}\frac{1}{\sqrt{N}}(z_i-\tau_i)'GD_3^{-1}D_1Y_1(P_{x_1}l-l)|$$
$$\leq C\sqrt{N}\|l'_{K^1(N)}\|_{0,\infty,X}\to 0$$

using boundedness assumptions, and (i) which allows application of Lemma 1(iii). Next, define the following vector $(z_i-\tau_i)y_1$ by

$$(z_i-\tau_i)y_1 = Y_1(z_i-\tau_i)$$

and noting that

$$(z_i-\tau_i)y_1 = \nu_i(x_1)+v_i$$

then we can write the first term as

$$\frac{1}{\sqrt{N}}(z_i-\tau_i)'GD_3^{-1}D_1Y_1P_{x_1}\epsilon$$
$$=\frac{1}{\sqrt{N}}\nu_i(x_1)'GD_3^{-1}D_1P_{x_1}\epsilon$$
$$+\frac{1}{\sqrt{N}}v_i'GD_3^{-1}D_1P_{x_1}\epsilon.$$

Defining ζ_{it} by

$$\zeta_{it} = v_{it}\frac{g(x_{3t})}{\lambda_t}d_1(l(x_{1t}))$$

we have that $E(\zeta_{it}|x_1)=0$ and $V(\zeta_{it}|x_1)$ is bounded above given Assumption 6.2 and boundedness of the other functions. Hence,

$$E|\frac{1}{\sqrt{N}}v_i'GD_3^{-1}D_1P_{x_1}\epsilon|$$
$$\leq\frac{1}{\sqrt{N}}E|\zeta_i'P_{x_1}\epsilon|\to 0$$

by Lemma A2.1 since ϵ has conditional mean zero and bounded conditional variance and since (i) holds.

Thus, we need only consider

$$\frac{1}{\sqrt{N}}\vartheta_i(x_1)'GD_3^{-1}D_1P_{x_1}\epsilon = \frac{1}{\sqrt{N}}B_{12}^{N'}P_{x_1}\epsilon$$

and show that for any vector b such that $\|b\| = 1$,

$$\frac{1}{\sqrt{N}}b'B_{12}^{N'}P_{x_1}\epsilon$$

is normally distributed. But by Theorem 2 we have that since $b'B_{12}^{N'}$ is a function of x_1 which has positive Sobolev smoothness index then

$$\frac{1}{\sqrt{N}}b'B_{12}^{N'}P_{x_1}\epsilon = \frac{1}{\sqrt{N}}b'B_{12}^{N'}\epsilon + o_p(1)$$

and this is asymptotically normal using the Liapunov Central Limit Theorem so that

$$\frac{1}{\sqrt{N}}b'B_{12}^{N'}P_{x_1}\epsilon \to N(0, S_2)$$

Finally, we have that

$$\sqrt{N_1}(\tilde{\beta}_2 - \beta_2) = (A_1^N)^{-1}(B_{11}^{N'}\xi + B_{12}^{N'}\epsilon) + o_p(1)$$

and since ξ and ϵ have zero correlation, then the result holds. **Q.E.D.**

Lemma A2.1 : *Let ζ_1 and ζ_2 be vectors of dimension N and P_x be a $N \times N$ projection matrix formed by K functions of x and suppose that the observations are independent and*

$$E(\zeta_{it}|x_t) = 0$$

$$0 < \sup_t E(\zeta_{it}^2|x_t) < \infty$$

then

$$\frac{1}{\sqrt{N}}E(\zeta_1'P_x\zeta_2) \leq C\frac{K}{N}$$

for some finite C. **Proof:**

$$E|\frac{1}{\sqrt{N}}\zeta_1' P_x \zeta_2| \leq \frac{1}{\sqrt{N}}(E(\zeta_1' P_x \zeta_1))^{1/2}(E(\zeta_2' P_x \zeta_2))^{1/2}$$
$$= \frac{1}{\sqrt{N}}(E(trP_x E(\zeta_1\zeta_1'|x))^{1/2}(E(trP_x E(\zeta_2\zeta_2'|x))^{1/2}$$
$$\leq C\frac{1}{\sqrt{N}}E(\mathrm{tr}P_x) \leq C\frac{K}{\sqrt{N}}$$

since the variances are bounded by some finite constant C and since the $trP_x = K$ because it is a projection matrix. **Q.E.D.**

Proof of Theorem 5:

It is easy to show that under the conditions of this result, the conclusion of Theorem 4 hold. In fact instead of having MSE of the various functions being $o_p(N^{-1/2})$ it is instead $o_p(N^{-3/4})$, although this fact is only used to prove part of this result. The convergence at rate $o_p(N^{-1/2})$ suffices for most of the proof. Clearly, we have

$$\tilde{A}_1^N = A_1^N + o_p(1)$$

as in the proof of Theorem 3 so it remains to consider the remaining two matrices. This will be done by showing convergence element by element. In the case of S_1 the typical element will be written in shorthand notation as

$$\frac{1}{N_1}\sum_{t=1}^{N_1}\hat{a}_{it}\hat{a}_{jt}\frac{\hat{\xi}_t^2}{\hat{\lambda}_t^2}$$

where

$$\hat{a}_{it} = \hat{z}_{it} - P_{x_3}^t \hat{z}_i$$

with $P_{x_3}^t$ denoting the t row of P_{x_3}. To show that

$$\tilde{S}_1 = S_1 + o_p(1) \tag{B.123}$$

we show that for all i and j

$$\frac{1}{N_1}\sum \hat{a}_{it}\hat{a}_{jt}\frac{\hat{\xi}_t^2}{\hat{\lambda}_t^2} - \frac{1}{N_1}\sum a_{it}a_{jt}\frac{\xi^2}{\lambda_t^2} = o_p(1)$$

After some rearranging it is easy to show that it will be sufficient to show the following results,

$$\frac{1}{N_1}\sum \hat{a}_{it}\hat{a}_{jt}\frac{1}{\hat{\lambda}_t^2}(\hat{\xi}_t^2 - \xi_t^2) = o_p(1) \tag{B.124}$$

$$\frac{1}{N_1}\sum \hat{a}_{it}\hat{a}_{jt}\frac{(\lambda_t + \hat{\lambda}_t)}{\hat{\lambda}_t^2\lambda_t^2}(\lambda_t - \hat{\lambda}_t)\xi_t^2 = o_p(1) \tag{B.125}$$

$$\frac{1}{N_1}\sum \hat{a}_{it}(\hat{a}_{jt} - a_{jt})\frac{1}{\lambda_t^2}\xi_t^2 = o_p(1) \tag{B.126}$$

$$\frac{1}{N_1}\sum a_{jt}(\hat{a}_{it} - a_{it})\frac{1}{\lambda_t^2}\xi_t^2 = o_p(1) \tag{B.127}$$

The following inequalities will prove useful in verifying these four results. For all i we have

$$\max_t |\hat{a}_{it}| \leq \Delta + \max_t |\psi_t(\Psi'\Psi)^{-1}\Psi'\hat{z}_i|$$

Now we have,

$$\begin{aligned}
\max_t |\psi_t(\Psi'\Psi)^{-1}\Psi'\hat{z}_i| &\leq \sup_{x_3} |\psi(x_3)'\theta_{K^3}| \\
&+ \max_t(\psi_t(\Psi'\Psi)^{-1}\psi_t)^{1/2}((\hat{z}_i - z_i)'(\hat{z}_i - z_i))^{1/2} \\
&+ \max_t(\psi_t(\Psi'\Psi)^{-1}\psi_t)^{1/2}(u_i'P_{x_3}u_i)^{1/2} \\
&+ \max_t(\psi_t(\Psi'\Psi)^{-1}\psi_t)^{1/2}\sqrt{N_1}\|\tau_i^r\|_{0,\infty,\mathcal{X}} \\
&\leq \Delta + o_p(N^{1/4}) + o_p(N^{1/8}) + o_p(1)
\end{aligned}$$

using Assumption 6.2 and respectively, (ii) and Lemma 1, Assumption 2.3 and MSE convergence for $\hat{\lambda}_t$ of $o_p(N^{-1/2})$, Lemma A2.1 and (ii) and finally (ii) ,(iii) and Lemma 1(iii). Thus we have that

$$\max_t |\hat{a}_{it}| \leq \Delta o_p(N^{1/4}) \tag{B.128}$$

for some finite Δ. Next consider the MSE convergence of the estimated g for the observations with $y_1 = 1$. We have that

$$\frac{1}{\sqrt{N_1}}\|\hat{g}_1 - g_1\| \leq \frac{1}{\sqrt{N_1}}(\|P_{x_3}\hat{Z}(\beta_2 - \tilde{\beta}_2)\|$$

$$+\|(I-P_{x_3})g_1\| + \|P_{x_3}g\frac{(\lambda-\hat{\lambda})}{\hat{\lambda}}\| + \|P_{x_3}\frac{\xi}{\hat{\lambda}}\|)$$

The following four results give the desired MSE convergence result for $\hat{g}_1$.

$$\frac{1}{N_1}\|P_{x_3}\hat{Z}(\beta_2-\tilde{\beta}_2)\|^2 \leq \Delta O_p(N^{-1})$$

$$(\frac{1}{N_1}\|(I-P_{x_3})g_1\|^2)^{1/2} \leq \|g^r_{K^3}\|_{0,\infty,\mathcal{X}} = o_p(N^{-1/2})$$

$$(\frac{1}{N_1}\|P_{x_3}g\frac{(\lambda-\hat{\lambda})}{\hat{\lambda}}\|^2)^{1/2} \leq \Delta o_p(N^{-1/4})$$

$$(\frac{1}{N_1}\|P_{x_3}\frac{\xi}{\hat{\lambda}}\|^2)^{1/2} \leq \Delta o_p(N^{-3/8})$$

using respectively, root N convergence of $\tilde{\beta}_2$, (ii) and Lemma 1, MSE convergence for $\hat{\lambda}_t$, and (i). Therefore we have that

$$\frac{1}{\sqrt{N_1}}\|\hat{g}_1 - g_1\| \leq \Delta o_p(N^{-1/4}) \tag{B.129}$$

Finally, noting that

$$\xi_t = x_{2t}(\beta_2-\tilde{\beta}_2) + g(x_{3t})(\lambda_t-\hat{\lambda}_t) + \hat{\lambda}_t(\hat{g}(x_{3t})-g(x_{3t})) + \xi_t$$

it is easy to show that

$$\frac{1}{N_1}\sum(\hat{\xi}_t^2-\xi_t^2) \leq \Delta o_p(N^{-1/2}) \tag{B.130}$$

using the boundedness of g, $\hat{\lambda}$ and the aforementioned results regarding $\tilde{\beta}_2$, $\hat{\lambda}_t$ and $\hat{g}$. Using this and (B.128) then we can show that (B.124) holds. For (B.125) write it as,

$$\frac{1}{N_1}\sum \hat{a}_{it}(\hat{a}_{jt}-a_{jt})\frac{(\lambda_t+\hat{\lambda}_t)}{\hat{\lambda}_t^2\lambda_t^2}(\lambda_t-\hat{\lambda}_t)\xi_t^2$$

$$+\frac{1}{N_1}\sum a_{it}\frac{(\lambda_t+\hat{\lambda}_t)}{\hat{\lambda}_t^2\lambda_t^2}(\lambda_t-\hat{\lambda}_t)\xi_t^2$$

then by (B.128) boundedness of the a_{it}, the convergence in MSE at $o_p(N^{-1/2})$ for $\hat{a}_j$ and $\hat{\lambda}$, the boundedness of the fourth moments of ξ_t we can show that each of these terms

converges to zero by the use of Cauchy Inequality and Markovs Inequality. Note that we have that

$$\frac{1}{N_1}\|\hat{a}_i - a_i\|^2 = o_p(N^{-3/4}) \tag{B.131}$$

although this is faster than is needed in this part of the proof. Finally the same sort of arguments verify that (B.126) and (B.127) hold and so we have verified that (B.123) holds.

We next must show that

$$\hat{S}_2^N = S_2^N + o_p(1) \tag{B.132}$$

and this will be done in a similar fashion to the previous result. First, denote the ith element of $\hat{\nu}$ by

$$\hat{\nu}_i = P_{x_1}((\hat{z}_i - P_{x_3}\hat{z}_i)'0')$$

and denote the population values for each t by ν_{it}. Then we must show that for all i and j that

$$\frac{1}{N}\sum_{t=1}^{N}\hat{\nu}_{it}\hat{\nu}_{jt}\hat{g}(x_{3t})^2\hat{c}_t - \frac{1}{N}\sum_{t=1}^{N}\nu_{it}\nu_{jt}g(x_{3t})^2c_t = o_p(1)$$

where

$$c_t = g(x_{3t})^2\frac{1}{\lambda_t^2}d_1(l_t)^2l_t(1-l_t)$$

and $\hat{c}_t$ denotes the estimate of this. Since we have that l_t d_1, and λ_t and the estimates of these are all bounded, and the estimators converge in MSE to the true values at $o_p(N^{-3/4})$, by Lemma 1 the main concern will be with $\hat{g}$ and the $\hat{\nu}_{it}$. Since in this case we must estimate g for all the observations, we must verify that a similar result to that in (B.130) holds. Also similar inequalities to that in (B.128) will be required. First, we deal with $\hat{g}$, which is given by,

$$\hat{g} = ((P_{x_3}(\hat{w} - \hat{Z}\tilde{\beta}_2))'(\Psi_{30}(\Psi_3'\Psi_3)^{-1}\Psi_3'((\hat{w} - \hat{Z}\tilde{\beta}_2))')'$$

where Ψ_{30} is the matrix of basis functions evaluated at the x_{3t} for observations for which $y_1 = 0$. For any t we can write the estimate as

$$\hat{g}(x_{3t}) = \psi_t'(\Psi_3'\Psi_3)^{-1}\Psi_3'((\dot{w} - \hat{Z}\tilde{\beta}_2)$$

so that as we did above to yield inequality (B.128) we have that

$$|\hat{g}(x_{3t})| \leq (\psi_t'(\Psi_3'\Psi_3)^{-1}\psi_t)^{1/2}((\tilde{\beta}_2 - \beta_2)'\dot{Z}'\hat{Z}(\tilde{\beta}_2 - \beta_2))^{1/2}$$

$$+|\psi_t'\theta_{K^3}| + (\psi_t'(\Psi_3'\Psi_3)^{-1}\psi_t)^{1/2}\sqrt{N}\|g^r_{K^3}\|_{0,\infty,X}$$

$$+(\psi_t'(\Psi_3'\Psi_3)^{-1}\psi_t)^{1/2}\Delta\|\lambda - \hat{\lambda}\| + (\psi_t'(\Psi_3'\Psi_3)^{-1}\psi_t)^{1/2}\Delta\|\xi P_{x_3}\|$$

$$\leq \Delta o_p(N^{1/8})$$

using argument similar to the above and the fact that convergence of $\hat{\lambda}$ is $o_p(N^{-3/4})$, using (i) and (ii), and Assumption 6.2 which by Andrews (1989) shows that

$$(\psi_t'(\Psi_3'\Psi_3)^{-1}\psi_t)^{1/2} = o_p(1) \tag{B.133}$$

It is clear that since (B.129) holds to show MSE convergence for $\hat{g}$ which is based on all the observations we need only consider the convergence for the observations with $y_1 = 0$. Therefore, we note that we can write the appropriate expression as,

$$\hat{g_0} - g_0 = \Psi_{30}(\Psi_3'\Psi_3)^{-1}\Psi_3'\hat{Z}(\tilde{\beta}_2 - \beta_2) + \Psi_{30}(\Psi_3'\Psi_3)^{-1}\Psi_3'\Psi_3\theta_{K^3}$$

$$-g_0 + \Psi_{30}(\Psi_3'\Psi_3)^{-1}\Psi_3'g^r_{K^3} + \Psi_{30}(\Psi_3'\Psi_3)^{-1}\Psi_3'\frac{\xi}{\hat{\lambda}}$$

Since we have (B.133) then as before we can show that given (ii), (v) and the result proved in Lemma 1(iii), then

$$(\frac{1}{N}\|\hat{g}_0 - g_0\|^2)^{1/2} = o_p(N^{-1/4})$$

and so we have that

$$\frac{1}{N}\|\hat{g} - g\|^2 = o_p(N^{-1/2}) \tag{B.134}$$

Regarding $\hat{\nu}_{it}$ we have that

$$|\hat{\nu}_{it}| = |\psi_{1t}'(\Psi_1'\Psi_1)^{-1}\Psi_1(\hat{a}_i'0')|$$

$$\leq \psi_{1t}'(\Psi_1'\Psi_1)^{-1}\Psi_1'(a_i'0')'|$$

$$+\psi_{1t}'(\Psi_1'\Psi_1)^{-1}\Psi_1'((\hat{a}_i - a_i)'0')'|$$

$$\leq \Delta o_p(N^{1/8}) \tag{B.135}$$

using similar arguments to the above. Also, it is the case that

$$\frac{1}{N}\|\hat{\nu}_i - \nu_i\|^2 = o_p(N^{-1/2}). \tag{B.136}$$

In fact, this like the other rates is slower than the actual rates that obtain under the assumptions but is sufficient for the proof. Now to show (B.132) the following will suffice,

$$\frac{1}{N}\sum \hat{g}(x_{3t})^2\nu_{it}(\hat{\nu}_{jt} - \nu_{jt})$$

$$\leq \frac{1}{N}\sum \hat{g}(x_{3t})^2(\hat{\nu}_{it} - \nu_{it})(\hat{\nu}_{jt} - \nu_{jt}) + \frac{1}{N}\sum \hat{g}(x_{3t})^2\nu_{it}(\hat{\nu}_{jt} - \nu_{jt})$$

$$\leq \Delta o_p(N^{1/4})\frac{1}{N}\|\hat{\nu}_i - \nu_i\|\|\hat{\nu}_j - \nu_j\| + \Delta o_p(N^{1/4}\|\hat{\nu}_j - \nu_j\|$$

$$\leq \Delta o_p(1)$$

using (B.136), (B.135), and (B.134). Next,

$$\frac{1}{N}\sum \hat{g}(x_{3t})^2\nu_{jt}(\hat{\nu}_{it} - \nu_{it})$$

$$\leq o_p(N^{1/4})\Delta(\frac{1}{N}\|\hat{\nu}_i - \nu_i\|^2)^{1/2}$$

and

$$\frac{1}{N}\sum \nu_{it}\nu_{jt}(\hat{g}(x_{3t}) + g(x_{3t}))(\hat{g}(x_{3t}) - g(x_{3t}))$$

$$\leq \Delta o_p(1)(\frac{1}{N}\|\hat{g} - g\|^2)^{1/2} = o_p(1)$$

It remains then to show that

$$\frac{1}{N}\sum \nu_{it}\nu_{jt}g(x_{3t})^2(\hat{c}_t - c_t) = o_p(1)$$

and this follows due to the boundedness of all of the elements and the fact that each converges in MSE to the true value. Therefore, we have that (B.132) holds and so the result of the Theorem follows. **Q.E.D.**

Proof of Theorems 6 and 7:

The result follows from Lemma 3.3 of White (1980a). **Q.E.D.**

Bibliography

[1] Adams, R.A.: *Sobolev Spaces*, New York: Academic Press (1975).

[2] Amemiya, T.: "Regression Analysis When the Dependent Variable is Truncated Normal", *Econometrica*, 41 (1973), 997–1016.

[3] Amemiya, T.: *Advanced Econometrics*, Cambridge Massachusetts: Harvard University Press (1985).

[4] Andrews, D.W.K.: "Asymptotic Normality for Series Estimators for Various Nonparametric and Semiparametric Models", Cowles Discussion Paper 874, Yale University, (1988).

[5] Andrews, D.W.K.: "Asymptotics for Semiparametric Econometric Models I: Estimation", Cowles Discussion Paper 908, Yale University, (1989a).

[6] Andrews, D.W.K.: "Asymptotics for Semiparametric Econometric Models II: Stochastic Equicontinuity", Cowles Discussion Paper 909, Yale University, (1989b).

[7] Andrews, D.W.K.: "Asymptotics for Semiparametric Econometric Models III: Testing and Examples", Cowles Discussion Paper 910, Yale University, (1989c).

[8] Andrews, D.W.K.: "Asymptotic Normality for Series Estimators for Various Nonparametric and Semiparametric Models", Cowles Discussion Paper 874R, Yale University, (1989d).

[9] Andrews, D.W.K.: "Asymptotic Optimality of GC_L, Cross Validation, and GCV for Models with Heteroskedastic Errors", Cowles Discussion Paper 906, Yale University (1989e).

[10] Andrews, D.W.K. and Yoon-Jae Whang: "Additive Interactive Regression Models: Circumvention of the Curse of Dimensionality", Cowles Discussion Paper 925, Yale University, (1989).

[11] Arabmazar, A. and P. Schmidt: "Further Evidence on the Robustness of the Tobit Estimator to Heteroskedasticity", *Journal of Econometrics*, 17 (1981), 253–258.

[12] Arabmazar, A. and P.Schmidt: "An Investigation of the Robustness of the Tobit Estimator to Non-Normality", *Econometrica*, 50 (1982), 1055–1063.

[13] Blundell, R. (ed.): "Specification Testing in Limited and Discrete Dependent Variable Models", *Journal of Econometrics*, 34 (1987).

[14] Burguette, J.F., A.R. Gallant, and G. Souza: "On Unification of the Asymptotic Theory of Nonlinear Econometric Models", *Econometric Reviews*, 1 (1982), 151–190.

[15] Cosslett, S.R.: "Distribution Free Maximum Likelihood Estimation of the Binary Choice Model", *Econometrica*, 51 (1983), 765–782.

[16] Cosslett, S.R.: "Distribution-Free Estimator of Regression Model with Sample Selectivity", University of Florida, Unpublished manuscript, (1984).

[17] Cragg, J.G.: "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods", *Econometrica*, 39 (1971), 829–844.

[18] Cragg, J.G.: "Quasi-Aitken Estimation for Heteroskedasticity of Unknown Form", U.B.C. Discussion Paper 88–34, (1988).

[19] Dagenais, M.G.: "The Tobit Model with Serial Correlation", *Economics Letters*, 10 (1983), 263–267.

[20] Dhrymes, P.J. : "Limited Dependent Variables" in *Handbook of Econometrics*, Volume 3 edited by Z. Griliches and M.D. Intriligator, New York: North Holland, (1986).

[21] Donald, S.G.: "An Empirical Analysis of the Relationship Between Appliance Choice and Electricity Demand", unpublished B.Ec. thesis, University of Sydney, (1985).

[22] Duncan, G.M.(ed.): "Continuous / Discrete Econometric Models with Unspecified Error Distributions", *Journal of Econometrics*, 32 (1986a).

[23] Duncan, G.M.: "A Semi-Parametric Censored Regression Estimator", *Journal of Econometrics*, 32 (1986b), 5–34.

[24] Elbadawi, I.A., A.R. Gallant, and G. Souza: "An Elasticity Can be Estimated Consistently Without A Priori Knowledge of Functional Form", *Econometrica*, 51 (1983), 1731–1751.

[25] Fernandez, L.:"Nonparametric Maximum Likelihood Estimation of Censored Regression Models", *Journal of Econometrics*, 32 (1986), 32–57.

[26] Gallant, A.R.: "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form", *Journal of Econometrics*, 15 (1981), 211–245.

[27] Gallant A.R. and D Nychka: "Semi-nonparametric Maximum Likelihood Estimation", *Econometrica*, 55 (1987), 363–390.

[28] Gallant, A.R.: "Identification and Consistency in Seminonparametric Regression", in *Advances in Econometrics*, Fifth World Congress of the Econometric Society Vol. I., edited by T.F. Bewley, Cambridge: Cambridge University Press (1987).

[29] Geman, S. and C. Hwang: "Nonparametric Maximum Likelihood Estimation by the Method of Sieves", *Annals of Statistics*, 10 (1982), 401–414.

[30] Gourieroux, C., A. Monfort, and A. Trognon: "Estimation and Test in Probit Models with Serial Correlation", in *Alternative Approaches to Time Series Analysis*, Florens, J.P. and C. Simor (eds.), Brussels Publications des Facultes Universitaires Saint Louis (1984).

[31] Gronau, R.: "The Effect of Children on the Housewifes Value of Time", *Journal of Political Economy*, 81 (1973), 1119–1143.

[32] Hausman, J.A.: "Specification Tests in Econometrics", *Econometrica*, 46 (1978), 1251–1271.

[33] Heckman, J.: "Shadow Prices, Market Wages and Labor Supply", *Econometrica*, 42 (1974), 679–693.

[34] Heckman, J.: "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models", *Annals of Economic and Social Measurement*, 5 (1976), 475–492.

[35] Heckman, J.: "Sample Selection Bias as a Specification Error", *Econometrica*, 47 (1979), 153–161.

[36] Horowitz, J.L.: "A Distribution Free Least Squares Estimator for Censored Regression Models", *Journal of Econometrics*, 32 (1986), 59–84.

[37] Horowitz, J.L.: "Semiparametric M-Estimation of Censored Linear Regression Models" in *Advances in Econometrics: Robust and Nonparametric Inference* edited by G.F. Rhodes and T.B. Fomby (ed.), Greenwich: JAI Press, 7 (1988).

[38] Horowitz, J.L.: "A Smoothed Maximum Score Estimator for the Binary Response Model". Working Paper 89-30, Department of Economics, University of Iowa, (1989).

[39] Hurd, M.: "Estimation in Truncated Samples When There is Heteroskedasticity" *Journal of Econometrics*, 11 (1979), 247–258.

[40] Ichimura, H.: "Estimation of Single Index Models", unpublished Ph.D. Dissertation Dept of Economics MIT (1988).

[41] Klein, R.W. and R.H. Spady: "An Efficient Semiparametric Estimator for Discrete Choice Models", Economics Research Group, Bell Communications Research, Morristown New Jersey, (1987).

[42] McFadden, D.: "Conditional Logit Analysis of Qualitative Choice Behaviour", in *Frontiers in Econometrics*, edited by P. Zarembka, New York: Academic Press, (1974).

[43] McFadden, D.L.: "Econometric Analysis of Econometric Response Models", in, *Handbook of Econometrics*, Volume 2 edited by Z. Griliches and M.D. Intriligator, New York: North Holland, (1984).

[44] MacKinnon, J.G. and H. White: "Some Heteroskedasticity Consistent Covariance Matrix Estimators with Improved Finite Sample Properties", *Journal of Econometrics*, 29 (1985), 305–325.

[45] Maddala, G.S.: *Limited Dependent and Qualitative Variables in Econometrics*, Cambridge, England: Cambridge University Press, (1983).

[46] Maddala, G.S.: "Disequilibrium, Self Selection and Switching Models", in, *Handbook of Econometrics*, Volume 3 edited by Z. Griliches and M.D. Intriligator, New York: North Holland, (1987).

[47] Manski, C.: "Maximum Score Estimation of the Stochastic Model of Discrete Choice" *Journal of Econometrics*, 3 (1975), 205–228.

[48] Manski, C: "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator" *Journal of Econometrics*, 27 (1985), 313–333.

[49] Matzkin, R.L.: "Nonparametric and Distribution Free Estimation of Binary Choice and Threshold Crossing Models" Cowles Discussion Paper 889, Yale University, (1988).

[50] Mroz, T.A.: "The Sensitivity of an Empirical Model of Married Womens Hours of Work to Economic and Statistical Assumptions", *Econometrica*, 55 (1987), 765–799.

[51] Newey, W.K.: "Adaptive Estimation of Regression Models via Moment Conditions", *Journal of Econometrics*, 38 (1988), 301–339.

[52] Newey, W.K.: "Two Step Estimation of Sample Selection Models", unpublished manuscript, Dept. of Economics Princeton University (1988).

[53] Newey, W.K., J.L. Powell, and J.R. Walker: "Semiparametric Estimation of Selection Models: Some Empirical Results", *American Economic Review, Papers and Proceedings*, 80 (1990), 324–328.

[54] Paarsch, H.J.: "Monte-Carlo Comparison of Estimators for the Censored Regression Model", *Journal of Econometrics*, 24 (1984), 197–213.

[55] Pagan, A.R. and F. Vella: "Diagnostic Tests for Models Based on Individual Data: A Survey", University of Rochester Working Paper 162, (1988).

[56] Phillips, P.C.B.: "Reflections on Econometric Methodology", *The Economic Record*, (1988).

[57] Poirier, D.J. and P.A. Ruud: "Probit with Dependent Observations", *Review of Economic Studies*, LV (1988), 593–614.

[58] Portnoy, S.: "Asymptotic Behavior of Likelihood Methods for Exponential Families when the Number of Parameters Tends to Infinity", *Annals of Statistics*, 16 (1988), 356–366.

[59] Powell, J.L.: "Least Absolute Deviations Estimation for the Censored Regression Model", *Journal of Econometrics*, 25 (1984), 303–325.

[60] Powell, J.L.: "Symmetrically Trimmed Least Squares Estimation of Tobit Models" *Econometrica*, 54 (1986), 1435–1460.

[61] Powell, J.L.: "Semiparametric Estimation of Bivariate Latent Variable Models", SSRI Working Paper 8704 University of Wisconsin, Madison, (1987).

[62] Powell, J.L., J.H. Stock and T.M. Stoker: "Semiparametric Estimation of Index Coefficients", *Econometrica*, 57 (1989), 1403–1430.

[63] Rilstone, P.: "Nonparametric Hypothesis Testing for Realistically Sized Samples", Universite Laval Cahier 8823, (1988).

[64] Robinson, P.M.: "On the Asymptotic Properties of Estimators of Models Containing Limited Dependent Variables", *Econometrica*, 50 (1982), 27–41.

[65] Robinson, P.M.: "Root-N-Consistent Semiparametric Regression" *Econometrica*, 56 (1988), 931–954.

[66] Royden, H.L.: *Real Analysis*, Second Edition, New York: Macmillan, (1987).

[67] Ruud, P.: "Sufficient Conditions for Consistency of the Maximum Likelihood Estimator Despite Misspecification of Distribution", *Econometrica*, 51 (1983), 225–228.

[68] Theil, H.: "A Multinomial Extension of the Linear Logit Model", *International Economic Review*, 10 (1969), 251–259.

[69] Tobin, J.: "Estimation of Relationships for Limited Dependent Variables", *Econometrica*, 26 (1958), 24–36.

[70] Wald, A.: "A Note on the Consistency of the Maximum Likelihood Estimator" *Annals of Mathematical Statistics*, 60 (1949), 595–601.

[71] White, H: "Nonlinear Regression on Cross-Section Data", *Econometrica*, 48 (1980a), 721–726.

[72] White, H.: "A Heteroskedasticity Consistent Covariance Matrix and a Direct Test for Heteroskedasticity", *Econometrica*, 48 (1980b), 817–838.

[73] White, H.: *Asymptotic Theory for Econometricians*, Orlando: Academic Press, (1984).

[74] White, H.: "Maximum Likelihood Estimation of Misspecified Models", *Econometrica*, 50 (1982), 1–25.

[75] White, H. and M. Stinchcombe: "Adaptive Generalized Least Squares with Dependent Observations", University of California, San Diego Discussion Paper 89-45, (1989).