GMM Estimation for Nonignorable Missing Data: Theory and Practice

by

Sanha Hemvanich

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Economics

University of Warwick, Department of Economics
December 2007

Contents

Acknowledgments

In completing this thesis, I am most indebted to my supervisors, Richard J. Smith and Mark B. Stewart. I would like to thank Richard for taking me under his supervision. His suggestions and insights are invaluable to the development of my thesis. I would like to thank Mark for taking a major role in supervising my thesis after Richard left Warwick. He was very generous with his time and provided incredible help and support. I am also indebted to Jeremy Smith for introducing me to econometrics. Without his excellent teaching skills, I would have chosen another research area.

I am also grateful to the members of the econometrics workshop for their useful comments and suggestions on numerous occasions; especially Valentina Corradi, Michael Pitt and Paulo M.D.C Parente.

I would also like to thank my family for their encouragement and support; my fiancé, Pornpinun Chantapacdepong, for her love, care and understanding; and my PhD colleagues for making my study an enjoyable one. Help from several support staffs in the Department of Economics is also appreciated. I dedicate this dissertation to my parents.

Declaration

I hereby declare that the work in this dissertation is my own work and no part of the dissertation has been submitted for any other academic award. Any views expressed in the dissertation are those of the author.

Abstract

This thesis develops several Generalized Method of Moments (GMM) estimators for analysing Not Missing at Random (NMAR) data, which is commonly referred to as the self-selection problem in an economic context. We extend the semiparametric estimation procedures of Ramalho and Smith (2003) to include the case where the missing data mechanism (MDM) depends on both a continuous response variable and covariates. Within this framework, it is possible to avoid imposing any assumptions on the missing data mechanism. We also discuss the asymptotic properties of the proposed GMM estimators and establish the connections of them to the GMM estimators of Ramalho and Smith and to the pseudolikelihood estimators of Tang, Little and Raghunathan (2003). All of the aforementioned estimators are then compared to other standard estimators for missing data such as the inverse probability weighted and sample selection model estimators in a number of Monte Carlo experiments. As an empirical application, these estimators are also applied to analyse the UK wage distribution. We found that, in many circumstances, our proposed estimators perform better than the other estimators described; especially when there is no exclusion restriction or other additional information available. Finally, we summarise that, since the true MDM is unlikely to be known, several estimators which impose different assumptions on the MDM should be used together to examine the sensitivity of the outcomes of interest with respect to the assumptions made and the estimation procedures adopted.

A List of Abbreviations

Abbreviations	Definitions
CGMM	Two-step INRY-GMM estimator
CGMM1	One-step INRY-GMM estimator
DGMM	RS GMM estimator
GMM	Generalized method of moments
INR	Item nonresponse
INRY	Item nonresponse when the MDM depends on the dependent variable only
INTREG	Interval regression estimator
IPWLS	Inverse probability weighted Least Squares estimator
LFS	Labour force survey
MAR	Missing at random
MCAR	Missing completely at random
MDM	Missing data mechanism
MICE	Multiple Imputation by Chained Equations
NMAR	Not missing at random
NMW	National minimum wage
ONS	Office of national statistics
RMSE	Root mean square error
RS	Ramalho and Smith (2003)
SSML	Maximum likelihood sample selection model estimator
SSTS	Two-step sample selection model estimator
TLR	Tang, Little and Raghunathan (2003)
ULS	Unweighted Lease Squares estimator
UNR	Unit nonresponse

Chapter 1
Introduction

In a microeconometric context, random samples are rarely encountered in practice. In particular, it is unusual for each member of the population of interest to have the same probability of being included in the sample. As a result, most empirical modelling must cope with some form of selection in the available sample. A common source of selection in sampled data is when some data may be missing, i.e., although the initial sample is drawn randomly from the target population, values of some variables are unavailable to the investigator. Attrition in longitudinal studies and nonresponse in survey research are two cases in point.

Intuitively, whenever the available sample suffers from missing data, the investigator has at least two options. *Either* to discard the original sample and again attempt to draw a well-defined random sample from the same population *or* to analyse the incomplete sample using a technique that takes into account the bias occasioned by the missing data.

In practice, the literature on missing data appears to have evolved roughly in accordance with this intuition. On the one hand, the availability of a random refreshment sample, which is not subject to censoring or truncation, is assumed and the estimation of the parameters of interest from combining it with the initial sample is investigated; see, e.g., Dolton (2002), Hirano, Imbens, Ridder and Rubin (2001) and Tripathi (2003). On the other hand, knowledge of the missing data mechanism (MDM) is assumed and estimation and inference procedures are developed based only on the incomplete sample.

Accordingly, the missing data literature may be categorised into two broad groups, *viz*. approaches whose validity depends on the presence of a complete random refreshment sample and approaches that do not require such a sample. This second group is of primary interest in our study. Since knowledge of the MDM is assumed, this group may be further sub-categorised on the basis of type of MDMs.

First, consider a situation where the MDM consists of an initial sample being drawn randomly with a constant probability that a sample unit is selected into the observed sample. In this instance, the fully observed units in the incomplete sample still constitute a random subsample of all chosen units. This implies that conventional estimation and inference procedures are still valid in this context. Such MDMs are referred to as *Missing Completely At Random* (MCAR). In practice, however, it is unusual for an incomplete sample to satisfy MCAR.

Whenever the MDM depends on random variables which constitute the econometric or statistical model of interest, standard estimation and inference techniques typically need amending because of the bias caused by the MDM. A MDM is referred to as *Missing At Random* (MAR) if, given the fully observed variables, it does not depend on any variables with missing values.[1] The MDM is called *Not Missing At Random* (NMAR) if it is influenced by outcomes of both missing and fully observed variables.

---

[1] MAR is also known as *selection on observables* or *ignorablility of selection*. The second label comes from the fact that the MAR MDM can be generally ignored in likelihood inference; see Little and Rubin (2002, p.119).

Hence, the second group of robust approaches for missing data may be sub-classified into three sub-groups according to type of MDMs: (i) MCAR, (ii) MAR and (iii) NMAR; see Figure (1.1). Clearly, NMAR is more general than the other two types of MDMs.

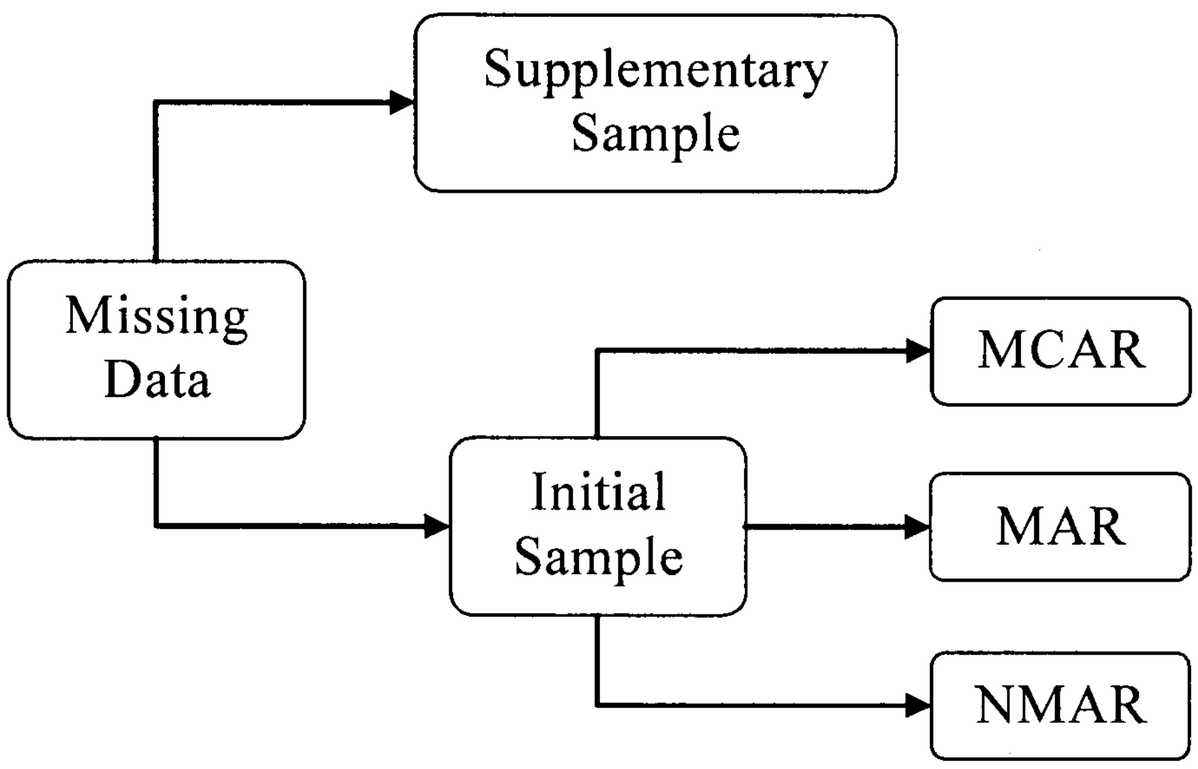

Fig. 1.1. Literature on Missing data

In this study, we are interested in settings where only the outcomes of the dependent variables are missing. In such a setting, maintaining MAR implies that MDM can be explained exclusively by observed exogenous covariates. On the other hand, NMAR means that MDM is a function of both response variable and covariates.

Since standard estimation and inference procedures for random sampling are valid for MCAR data, it receives no attention in the literature. In contrast, many estimation approaches such as multiple imputation and inverse probability weighting have been developed for MAR data; see *inter alia* Little and Rubin (2002), Robins, Rotnitzky and Zhoa (1995), Schafer (1997), Tsiatis (2006), van der Laan and Robins (2003) and Wooldridge (2002a, 2003).

Nevertheless, it is more usual to encounter NMAR data in many economic contexts being commonly referred to as the *self-selection problem*. There is an extensive literature on sample selection which considers essentially the situation where missing data is NMAR. Most of the early literature pioneered by Heckman (1976) focuses on the fully parametric approach. However, this method is strongly criticised for the sensitivity of its conclusion to a mild change in the functional form restrictions. The later literature therefore attempts to weaken these assumptions in certain respects; see, e.g., Manski (1989) and Vella (1998). Nonetheless, these relaxations are generally possible only at the cost of imposing some form of exclusion restriction.

Consequently, the MDM is usually restricted and specified whenever data is NMAR. To our knowledge, Manski's (2003) bounds approach is the only method which analyses NMAR data that imposes no restrictions on the MDM. The cost of this degree of robustness is that the method can only place the parameters of interest within a set-valued identification region.

Ramalho and Smith (2003), RS henceforth, has proposed a Generalized Method of Moments (GMM) estimator for a special case of NMAR where the MDM depends only on values of the discrete missing response variable. This approach is based on an efficient GMM estimator for choice-based samples which is developed in Imbens (1992). An advantage of RS's approach over the estimation procedures previously proposed is that it avoids an explicit specification of the form of MDM. Tang, Little and Raghunathan (2003), henceforth TLR, proposed a set of estimators for the same special case of NMAR allowing the response variable to be continuously distributed. Their estimators are also based on

the tools originally developed for response-biased sampling by Chen (2001) and Liang and Qin (2000).

A main contribution of this thesis, which is presented in Chapter 2, is the extension of the semiparametric estimation procedures developed in RS to include the case where the response variable is continuous and where the MDM depends on both the response variable and covariates. Within this framework, it is possible to avoid imposing any assumptions on the MDM. However, the approach is not as general as Manski's because a correct specification of conditional population distribution of the dependent variable given covariates must be assumed. Asymptotic properties of the proposed GMM estimation are also discussed in Chapter 2. In addition, the connections of the estimators to both RS and TLR are established and the identification of them in both theory and practice is considered.

Chapter 3 provides Monte Carlo evidence on the finite sample performance of a subset of the proposed GMM estimators in comparison to other estimators for missing data. These estimators are inverse probability weighted estimators, unweighted estimators, sample selection model estimators, the pseudolikelihood estimators of TLR and the GMM estimators of RS. The Monte Carlo experiments conducted are designed to demonstrate the contrasting performance of all estimators in a variety of circumstances. The differing assumptions required of the underlying model for the estimators to display satisfactory finite sample performance are also stressed.

Chapter 4 focuses on comparing all estimators considered in Chapter 3 using real data. The estimators are applied to analyse the UK wage distribution. The approach of this empirical chapter is based on Skinner et al (2002) which develops a method of estimating

the distribution of the Labour Force Survey (LFS) wage rate variable for the Office for National Statistics (ONS). In addition to the aforementioned estimators, an imputation technique called Multiple Imputation by Chained Equations (MICE) is described and then employed in the empirical investigation. Finally, the conclusions and some suggestions for future research are given in Chapter 5.

Chapter 2
Two-Step GMM Estimators for Nonignorable Missing Data

2.1 Introduction

An advantage of both the RS GMM and the TLR's pseudolikelihood estimators is that they do not require an explicit specification of the MDM. However, their major weakness is that the MDM is restricted to depend only on values of the missing response variable. In this chapter, the RS's estimation procedure is extended to include the case where MDM depends on both a response variable and covariates. As a result, it is possible to avoid imposing any assumptions on the MDM within this framework. The asymptotic properties of the GMM estimators proposed are also given in Section 2.5. The connections of the estimators to both RS and TLR are then established in Sections 2.6 and 2.7. Section 2.8 explains why the proposed GMM estimators are theoretically identified. It also suggests conditions for identification in practice and considers some related identification issues. Section 2.9 provides the summaries of this chapter.

2.2 Model Specification and Sampling Process

Let Y and X denote a scalar random variable and a p-vector of weakly exogenous covariates with respective sample spaces $\mathcal{Y}$ and $\mathcal{X}$. Suppose the population conditional density

function of Y given X is

$$f(y|x;\theta_0),$$

where $f(\cdot|\cdot;\theta)$ is known up to the parameter vector θ of dimension p, $\theta \in \Theta \subset \Re^p$, and θ_0 denotes the true value. The problem of interest is consistent estimation of and inference on the parameter vector θ_0.

If a random sample of realised values on Y and X were available, a simple solution would be to use standard conditional ML estimation to estimate θ_0. However, whenever values of Y are not observed from some sample units, this solution is not feasible for reliable inference. To address this issue, let R be an indicator variable taking value 1 if no data is missing and 0 otherwise and let P_{yx} denote the MDM, i.e.,

$$P_{yx} = \mathcal{P}\{R = 1|Y = y, X = x\}, \tag{2.1}$$

which is referred to as the conditional probability of response.

As in RS, we assume that a sample with missing data is generated as follows. First, a sampling unit is randomly drawn from the population. Secondly, the unit is *either* completely observed, in which case $R = 1$, with probability P_{yx} *or* incompletely observed when only the value of X is recorded and $R = 0$. A random sample of size N is collected on the triple (Y, X, R). The resultant sample will thus include sampling units with missing Y values. In principle, all values in the sample spaces $\mathcal{Y}$ and $\mathcal{X}$ are observable unlike in the censored regression or Tobit model case where values of the latent regression model regressand below the known censoring point can never be observed.

The MDM of interest, (2.1), is NMAR since P_{yx} can vary over both $\mathcal{Y}$ and $\mathcal{X}$. Another MDM, considered below, which is a special case of (2.1), allows $\mathcal{P}\{R = 1|Y = y, X = x\}$ to vary only with values of Y, i.e.,

$$\begin{aligned} P_y &= \mathcal{P}\{R = 1|Y = y, X = x\} \\ &= \mathcal{P}\{R = 1|Y = y\}. \end{aligned} \tag{2.2}$$

We develop our approach for both MDMs, (2.1) and (2.2). This enables us to compare our approach to those of RS and TLR. For completeness, we also consider an intermediate MDM. Let $X = (X_1, X_2)$. Then,

$$\begin{aligned} P_{yx_1} &= \mathcal{P}\{R = 1|Y = y, X_1 = x_1, X_2 = x_2\} \\ &= \mathcal{P}\{R = 1|Y = y, X_1 = x_1\}. \end{aligned} \tag{2.3}$$

This case has interest because it allows a subset of covariates X_2 to affect the conditional distribution of Y, but not the MDM.

Since only values of Y are missing in all cases we consider, they can be referred to collectively as item nonresponse (INR), as opposed to unit nonresponse (UNR) where values of all variables are jointly missing for nonrespondent units. In what follows, we refer to INR settings associated with (2.1), (2.2) and (2.3) as INRYX, INRY and INRYX1 respectively.

2.3 Moment Indicators for a Discrete Setting

Consider the above sampling process in a simple setting where Y and X are two discrete random variables with $\mathcal{Y} = \{y^1, ..., y^I\}$ and $\mathcal{X} = \{x^1, ..., x^J\}$. The joint probability

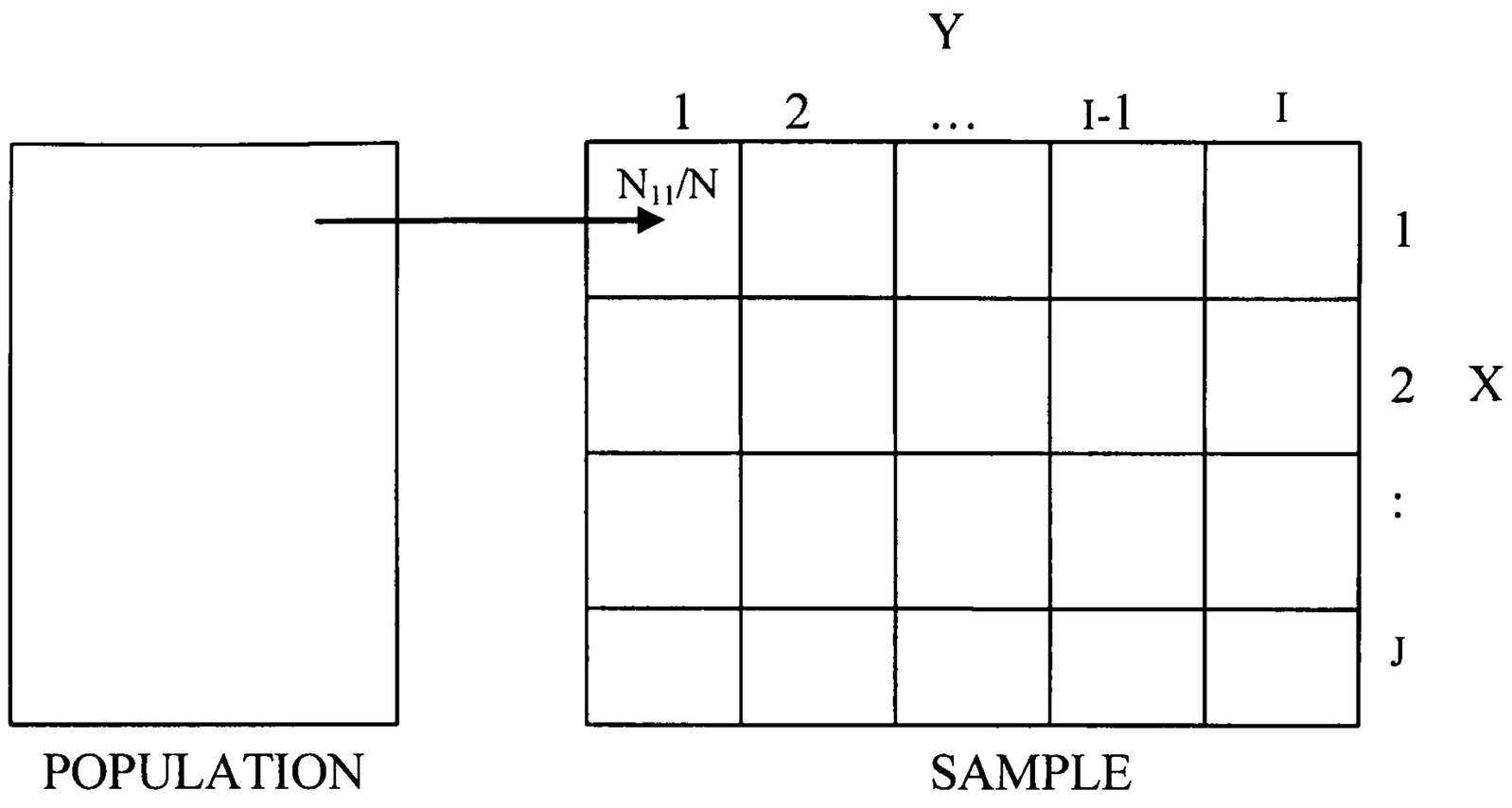

Fig. 2.2. The Sampling Process

$\mathcal{P}\{Y = y, X = x, R = 1\}$ may always be estimated for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$ using only the information from the sample. For example, let N_{ij} be the number of respondent units for which $Y = y_i$ and $X = x_j$. Then, $\mathcal{P}\{Y = y_i, X = x_j, R = 1\}$ can be simply estimated by the sample proportion N_{ij}/N. Such information, obtained directly from the sample, will prove crucial in the estimation procedure to be introduced below.

Within this discrete setting, we consider the likelihood functions of the observed data from INRYX, INRY and INRX1 and their first-order conditions. The first-order conditions are rewritten to concentrate out nuisance parameters and are used as moment conditions for GMM estimation. In term of notation, we use $\mathcal{P}\{\cdot|\cdot;\theta\}$, rather than $f(\cdot|\cdot;\theta)$, to stress that the setting under consideration is discrete.

2.3.1 INRYX

Observed Discrete Data Likelihoods

For a respondent unit, $R = 1$, the observed data likelihood is

$$\mathcal{P}\{Y = y, X = x, R = 1\} = P_{yx}\mathcal{P}\{y|x;\theta\}\mathcal{P}_X\{x\},$$

where $\mathcal{P}_X\{x\}$ denotes the true unknown marginal density function of X. Similarly, the observed data likelihood for a non-respondent unit, $R = 0$, is

$$\mathcal{P}\{X = x, R = 0\} = (1 - \sum_{y\in\mathcal{Y}} P_{yx}\mathcal{P}\{y|x;\theta\})\mathcal{P}_X\{x\}.$$

The joint observed data likelihood of a sample unit is thus given by

$$[P_{yx}\mathcal{P}\{y|x;\theta\}\mathcal{P}_X\{x\}]^r\,[(1 - \sum_{y\in\mathcal{Y}} P_{yx}\mathcal{P}\{y|x;\theta\})\mathcal{P}_X\{x\}]^{1-r}.$$

Even though $\mathcal{P}_X\{x\}$ is unknown, because X is discrete, we may replace $\mathcal{P}_X\{x\}$ by the (unknown) probability π_x associated with each mass point x, $x \in \mathcal{X}$. Both P_{yx} and the probability masses π_x will be jointly estimated with θ_0 in the proposed estimation procedures.

The objective function based on the above likelihood function for the random sample $\{y_n, x_n, r_n\}_{n=1}^N$ is

$$\begin{aligned}\mathcal{L} \;=\; & \sum_{n=1}^{N}\{r_n[\log P_{y_n x_n} + \log\mathcal{P}\{y_n|x_n;\theta\} + \log\pi_{x_n}] \\ & + (1 - r_n)\,[\log(1 - \sum_{y\in\mathcal{Y}} P_{yx_n}\mathcal{P}\{y|x_n;\theta\}) + \log\pi_{x_n}]\} \\ & - \mu(\sum_{x\in\mathcal{X}}\pi_x - 1).\end{aligned}$$

The first-order conditions with respect to the unknowns π_x, θ, P_{yx} and μ are

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \pi_x} &= \sum_{n=1}^{N}\left\{r_n\frac{1[x_n=x]}{\pi_x}-(1-r_n)\frac{1[x_n=x]}{\pi_x}\right\}-\mu \\ &= \sum_{n=1}^{N}\left\{\frac{1[x_n=x]}{\pi_x}\right\}-\mu, x\in\mathcal{X};\end{aligned} \tag{2.4}$$

$$\frac{\partial \mathcal{L}}{\partial \theta}=\sum_{n=1}^{N}\left\{r_n\frac{\partial \log \mathcal{P}\{y_n|x_n;\theta\}}{\partial\theta}-(1-r_n)\frac{1}{1-\sum_{y\in\mathcal{Y}}P_{yx_n}\mathcal{P}\{y|x_n;\theta\}}\sum_{y\in\mathcal{Y}}P_{yx_n}\frac{\partial\mathcal{P}\{y|x_n;\theta\}}{\partial\theta}\right\}; \tag{2.5}$$

$$\frac{\partial \mathcal{L}}{\partial P_{yx}}=\sum_{n=1}^{N}\left\{r_n\frac{1[y_n=y]\cdot 1[x_n=x]}{P_{yx}}-(1-r_n)\frac{1[x_n=x]\cdot\mathcal{P}\{y|x;\theta\}}{1-\sum_{y\in\mathcal{Y}}P_{yx}\mathcal{P}\{y|x;\theta\}}\right\}, y\in\mathcal{Y}, x\in\mathcal{X}; \tag{2.6}$$

$$\frac{\partial \mathcal{L}}{\partial \mu}=\sum_{x\in\mathcal{X}}\pi_x-1.$$

Moment Indicators

Equating (2.6) to zero, we can derive the following relationship

$$\frac{N_{yx}^r}{N_x^{nr}}=\frac{\widehat{P}_{yx}\mathcal{P}\{y|x;\widehat{\theta}\}}{1-\sum_{y\in\mathcal{Y}}\widehat{P}_{yx}\mathcal{P}\{y|x;\widehat{\theta}\}}, \tag{2.7}$$

where N_{yx}^r is the number of respondent units with $Y=y$ and $X=x$, and N_x^{nr} is the number of non-respondent units with $X=x$. Hence, re-arranging (2.7),

$$\widehat{P}_{yx}=\frac{N_{yx}^r}{N_x^{nr}}\left(\frac{\mathcal{P}\{y|x;\widehat{\theta}\}}{1-\sum_{y\in\mathcal{Y}}\widehat{P}_{yx}\mathcal{P}\{y|x;\widehat{\theta}\}}\right)^{-1}. \tag{2.8}$$

Summing (2.7) over $y\in\mathcal{Y}$ yields

$$\frac{N_x^r}{N_x^{nr}}=\frac{\sum_{y\in\mathcal{Y}}\widehat{P}_{yx}\mathcal{P}\{y|x;\widehat{\theta}\}}{1-\sum_{y\in\mathcal{Y}}\widehat{P}_{yx}\mathcal{P}\{y|x;\widehat{\theta}\}},$$

where $N_x^r = \sum_{y\in\mathcal{Y}} N_{yx}^r$ is the number of respondent units with $X = x$. This relationship implies that

$$\sum_{y\in\mathcal{Y}} \widehat{P}_{yx}\mathcal{P}\{y|x;\widehat{\theta}\} = \frac{N_x^r}{N_x},$$

where N_x is the total number of sampling units with $X = x$ in the sample, i.e., $N_x = N_x^r + N_x^{nr}$. Note that the right hand side is the sample proportion of respondents with covariate value x which is a natural estimator for the probability $\mathcal{P}\{R = 1|X = x\} = \sum_{y\in\mathcal{Y}} P_{yx}\mathcal{P}\{y|x,\theta_0\}$. Substituting into the expression for $\widehat{P}_{yx}$ in (2.8) we get

$$\begin{aligned}\widehat{P}_{yx} &= \frac{N_{yx}^r}{N_x\mathcal{P}\{y|x;\widehat{\theta}\}} \\ &= \frac{N_{yx}^r/N}{(N_x/N)\mathcal{P}\{y|x;\widehat{\theta}\}},\end{aligned} \tag{2.9}$$

an intuitive expression for an estimator of P_{yx}, being the sample analogue of

$$\begin{aligned}P_{yx} &= \mathcal{P}\{R=1|Y=y,X=x\} \\ &= \frac{\mathcal{P}\{R=1,Y=y,X=x\}}{\mathcal{P}_X\{x\}\mathcal{P}\{y|x;\theta_0\}}.\end{aligned}$$

The relationship in (2.9) will be crucial in the derivation of two-step GMM estimator for INRYX in section 2.4.1.

Equating (2.4) to zero, multiplying by $\widehat{\pi}_x$ and summing over $x \in \mathcal{X}$ yields $\widehat{\mu} = N$. Substituting back into (2.4)

$$\widehat{\pi}_x = \frac{N_x}{N},$$

reflecting the ancillarity of X for θ_0 and confirming the above claim in (2.9) that the estimator for $\mathcal{P}_X\{x\} = \pi_x$ should be N_x/N.

For the moment indicators corresponding to P_{yx}, from (2.9),

$$0 = \sum_{n=1}^{N} \left\{ \widehat{P}_{yx} - r_n \frac{1[y_n = y] \cdot 1[x_n = x]}{\mathcal{P}\{y|x;\widehat{\theta}\}\hat{\pi}_x} \right\}, y \in \mathcal{Y}, x \in \mathcal{X}. \tag{2.10}$$

This set of moment conditions depend on the estimator for marginal distribution of X, $\hat{\pi}_x$, $x \in \mathcal{X}$, which can be estimated by N_x/N as shown above.

Moreover, the moment indicator for $\widehat{\theta}$ is (2.5) that depends only implicitly on the marginal distribution for X through $\widehat{P}_{yx}$ in the second term. Thus, the resultant set of GMM moment indicators are

$$\begin{aligned} \theta \quad &: \quad r \cdot \frac{\partial \log \mathcal{P}\{y|x;\theta\}}{\partial \theta} - (1-r) \cdot \frac{1}{1 - \sum_{y \in \mathcal{Y}} P_{yx} \mathcal{P}\{y|x;\theta\}} \sum_{y \in \mathcal{Y}} P_{yx} \frac{\partial \mathcal{P}\{y|x;\theta\}}{\partial \theta}, \\ P_{st} \quad &: \quad P_{st} - r \cdot \frac{1[y = s] \cdot 1[x = t]}{\mathcal{P}\{s|t;\theta\}(N_t/N)}, s \in \mathcal{Y}, t \in \mathcal{X}. \end{aligned}$$

2.3.2 INRY

Observed Discrete Data Likelihoods

Given (2.2), the observed discrete data likelihood for a respondent unit is

$$\begin{aligned} \mathcal{P}\{Y = y, X = x, R = 1\} &= P_y \mathcal{P}\{y|x;\theta\} \mathcal{P}_X\{x\} \\ &= \frac{H_y}{Q_y} \mathcal{P}\{y|x;\theta\} \mathcal{P}_X\{x\}, y \in \mathcal{Y}, x \in \mathcal{X}, \end{aligned}$$

where $H_y = \mathcal{P}\{Y = y, R = 1\}$ and $Q_y = \sum_{x \in \mathcal{X}} \mathcal{P}\{y|x;\theta\}\mathcal{P}_X\{x\}$. Define $H_y^{nr} = \mathcal{P}\{Y = y, R = 0\}$. Hence, the observed discrete data likelihood for a non-respondent unit is

$$\begin{aligned}\mathcal{P}\{X = x, R = 0\} &= \sum_{y \in \mathcal{Y}} \frac{H_y^{nr}}{Q_y}\mathcal{P}\{y|x;\theta\}\mathcal{P}_X\{x\} \\ &= (1 - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x;\theta\})\mathcal{P}_X\{x\}, x \in \mathcal{X}.\end{aligned}$$

Thus, the joint observed data likelihood of a sample unit is then given by

$$\left[\frac{H_y}{Q_y}\mathcal{P}\{y|x;\theta\}\mathcal{P}_X\{x\}\right]^r [(1 - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x;\theta\})\mathcal{P}_X\{x\}]^{1-r}.$$

As above, because X is discrete, the unknown $\mathcal{P}_X\{\cdot\}$ may be replaced by the (unknown) probability π_x associated with each mass point x, $x \in \mathcal{X}$. The objective function based on the above likelihood function for the random sample $\{y_n, x_n, r_n\}_{n=1}^N$ is

$$\begin{aligned}\mathcal{L} &= \sum_{n=1}^{N}\{r_n[\log H_y + \log \mathcal{P}\{y_n|x_n;\theta\} + \log \pi_{x_n} - \log Q_{y_n}] \\ &\quad + (1 - r_n)\,[\log(1 - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x_n;\theta\}) + \log \pi_{x_n}]\} \\ &\quad -\mu(\sum_{x \in \mathcal{X}} \pi_x - 1),\end{aligned}$$

where $Q_y = \sum_{x \in \mathcal{X}} \mathcal{P}\{y|x;\theta\}\pi_x$.

The first-order conditions with respect to H_y, π_x, θ and μ are

$$\frac{\partial \mathcal{L}}{\partial H_y} = \sum_{n=1}^{N}\left\{r_n \frac{1[y_n = y]}{H_y} - \frac{(1 - r_n)}{1 - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x_n;\theta\}} \frac{\mathcal{P}\{y|x_n;\theta\}}{Q_y}\right\}, y \in \mathcal{Y}; \quad (2.11)$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \pi_x} &= \sum_{n=1}^{N}\left\{r_n\left[\frac{1[x_n = x]}{\pi_x} - \frac{1}{Q_{y_n}}\mathcal{P}\{y_n|x;\theta\}\right] + \right. \quad (2.12) \\ &\quad \left. (1 - r_n)\left[\frac{\sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y^2}\mathcal{P}\{y|x_n;\theta\}\mathcal{P}\{y|x;\theta\}}{1 - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x_n;\theta\}} + \frac{1[x_n = x]}{\pi_x}\right]\right\} - \mu, x \in \mathcal{X};\end{aligned}$$

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{n=1}^{N} \left\{ r_n \left[\frac{\partial \log \mathcal{P}\{y_n|x_n;\theta\}}{\partial \theta} - \frac{1}{Q_{y_n}} \sum_{x \in \mathcal{X}} \frac{\partial \mathcal{P}\{y_n|x;\theta\}}{\partial \theta} \pi_x \right] \right. \tag{2.13}$$
$$\left. - \frac{(1-r_n)}{1 - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \mathcal{P}\{y|x_n;\theta\}} \right.$$
$$\left. \times \left[\sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y} \frac{\partial \mathcal{P}\{y|x_n;\theta\}}{\partial \theta} - \sum_{y \in \mathcal{Y}} \frac{H_y}{Q_y^2} \mathcal{P}\{y|x_n;\theta\} \sum_{x \in \mathcal{X}} \frac{\partial \mathcal{P}\{y|x;\theta\}}{\partial \theta} \pi_x \right] \right\};$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = \sum_{x \in \mathcal{X}} \pi_x - 1.$$

Moment Indicators

From (2.11), the ML estimator for H_y is given by

$$\widehat{H}_y = N_y^r \widehat{Q}_y \left[\sum_{n=1}^{N} \frac{(1-r_n)\,\mathcal{P}\{y|x_n;\widehat{\theta}\}}{1 - \sum_{y \in \mathcal{Y}} \frac{\widehat{H}_y}{\widehat{Q}_y} \mathcal{P}\{y|x_n;\widehat{\theta}\}} \right]^{-1}, \tag{2.14}$$

where N_y^r is the number of respondent units with $Y = y$. Notice that, from (2.11), the second and the third terms in (2.12) sum to zero. Then, multiplying the resultant expression from (2.12) by $\widehat{\pi}_x$ and summing over $x \in \mathcal{X}$ yields

$$\widehat{\mu} = \sum_{n=1}^{N} \{r_n + (1-r_n)\} \sum_{x \in \mathcal{X}} 1[x_n = x] = N.$$

Consequently, again from (2.12), the ML estimator $\widehat{\pi}_x = N^{-1} \sum_{n=1}^{N} 1[x_n = x]$. Moreover, the ML estimator for Q_y is then given by

$$\begin{aligned} \widehat{Q}_y &= \sum_{x \in \mathcal{X}} \mathcal{P}\{y|x;\widehat{\theta}\}\widehat{\pi}_x \\ &= \frac{1}{N} \sum_{n=1}^{N} \mathcal{P}\{y|x_n;\widehat{\theta}\}, y \in \mathcal{Y}. \end{aligned}$$

By a similar argument, the second and the forth terms of (2.13) sum to zero. Hence, (2.13) can be re-written as

$$0 = \sum_{n=1}^{N} \left\{ r_n \frac{\partial \log \mathcal{P}\{y_n|x_n;\widehat{\theta}\}}{\partial \theta} - \frac{(1-r_n)}{1-\sum_{y\in\mathcal{Y}} \frac{\widehat{H}_y}{\widehat{Q}_y}\mathcal{P}\{y|x_n;\widehat{\theta}\}} \left[\sum_{y\in\mathcal{Y}} \frac{\widehat{H}_y}{\widehat{Q}_y} \frac{\partial \mathcal{P}\{y|x_n;\widehat{\theta}\}}{\partial \theta} \right] \right\}. \quad (2.15)$$

Therefore, the GMM moment indicators are given by

$$\begin{aligned} H_s &: r1[y=s] - \frac{H_s}{Q_s} \frac{(1-r)\mathcal{P}\{s|x;\theta\}}{1-\sum_{y\in\mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x;\theta\}}, s \in \mathcal{Y}; \\ \theta &: r\frac{\partial \log \mathcal{P}\{y|x;\theta\}}{\partial \theta} - \frac{(1-r)}{1-\sum_{y\in\mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x;\theta\}} \sum_{y\in\mathcal{Y}} \frac{H_y}{Q_y} \frac{\partial \mathcal{P}\{y|x;\theta\}}{\partial \theta}; \quad (2.16) \\ Q_y &: Q_y - \mathcal{P}\{y|x;\theta\}, y \in \mathcal{Y}. \end{aligned}$$

2.3.3 INRYX1

In this subsection, $X = (X_1, X_2)'$, where X_1 and X_2 are discrete random vector with respective sample spaces $\mathcal{X}_1 = \{x_1^1, ..., x_1^{J_1}\}$ and$\mathcal{X}_2 = \{x_2^1, ..., x_2^{J_2}\}$. Define $J = J_1 J_2$. Therefore, we can treat X as in the previous subsections since it takes values from a set of J outcomes.

Observed Discrete Data Likelihoods

Under (2.3), the MDM depends on values of Y and X_1 while the conditional distribution of Y is a function of both X_1 and X_2. Define $H_{yx_1} = \mathcal{P}\{Y = y, X_1 = x_1, R = 1\}$ and $Q_{yx_1} = \sum_{x_2\in\mathcal{X}_2} \mathcal{P}\{y|x_1, x_2;\theta\}\mathcal{P}_{X_1|X_2}\{x_1|x_2\}\mathcal{P}_{X_2}\{x_2\}$.

For a respondent unit, the observed data likelihood is

$$\mathcal{P}\{Y = y, X = x, R = 1\} = \frac{H_{yx_1}}{Q_{yx_1}}\mathcal{P}\{y|x_1, x_2;\theta\}\mathcal{P}_{X_1|X_2}\{x_1|x_2\}\mathcal{P}_{X_2}\{x_2\},$$

where $\mathcal{P}_{X_1|X_2}\{\cdot|\cdot\}$ and $\mathcal{P}_{X_2}\{\cdot\}$ are unknown. The observed data likelihood for a non-respondent unit is

$$\mathcal{P}\{X=x, R=0\} = (1-\sum_{y\in\mathcal{Y}}\frac{H_{yx_1}}{Q_{yx_1}}\mathcal{P}\{y|x_1,x_2;\theta\})\mathcal{P}_{X_1|X_2}\{x_1|x_2\}\mathcal{P}_{X_2}\{x_2\}$$

The joint observed data likelihood of a sample unit is thus given by

$$\left[\frac{H_{yx_1}}{Q_{yx_1}}\mathcal{P}\{y|x_1,x_2;\theta\}\mathcal{P}_{X_1|X_2}\{x_1|x_2\}\mathcal{P}_{X_2}\{x_2\}\right]^r [(1-\sum_{y\in\mathcal{Y}}\frac{H_{yx_1}}{Q_{yx_1}}\mathcal{P}\{y|x_1,x_2;\theta\})\mathcal{P}_{X_1|X_2}\{x_1|x_2\}\mathcal{P}_{X_2}\{x_2\}]^{1-r}$$

Because X_2 is discrete, we may replace $\mathcal{P}_{X_2}\{x_2\}$ by the (unknown) probability π_{x_2} associated with each mass point x_2, $x_2 \in \mathcal{X}_2$. Similarly, we replace $\mathcal{P}_{X_1|X_2}\{x_1|x_2\}$ by the (unknown) probability $\omega_{x_1|x_2}$ associated with the mass point (x_1, x_2), $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$.

The parameters π_{x_2} and $\omega_{x_1|x_2}$ will be jointly estimated with θ_0 in the proposed estimation procedure.

The objective function based on the above likelihood function for the random sample $\{y_n, x_n, r_n\}_{n=1}^N$ is

$$\begin{aligned}\mathcal{L} = & \sum_{n=1}^{N}\{r_n[\log H_{y_n x_{1n}} + \log\mathcal{P}\{y_n|x_{1n},x_{2n};\theta\} + \log\omega_{x_{1n}|x_{2n}} + \log\pi_{x_{2n}} - \log Q_{y_n x_{1n}}] \\ & + (1-r_n)\left[\log(1-\sum_{y\in\mathcal{Y}}\frac{H_{yx_{1n}}}{Q_{yx_{1n}}}\mathcal{P}\{y|x_{1n},x_{2n};\theta\}) + \log\omega_{x_{1n}|x_{2n}} + \log\pi_{x_{2n}}\right]\} \\ & -\mu(\sum_{x_2\in\mathcal{X}_2}\pi_{x_2}-1) - \sum_{x_2\in\mathcal{X}_2}\lambda_{x_2}\left(\sum_{x_1\in\mathcal{X}_1}\omega_{x_1|x_2}-1\right),\end{aligned}$$

where $Q_{yx_1} = \sum_{x_2\in\mathcal{X}_2}\mathcal{P}\{y|x_1,x_2;\theta\}\omega_{x_1|x_2}\pi_{x_2}$.

The first-order conditions are

$$\frac{\partial \mathcal{L}}{\partial H_{yx_1}} = \sum_{n=1}^{N} \left\{ r_n \frac{1[y_n = y]1[x_{1n} = x_1]}{H_{yx_1}} - \frac{(1 - r_n)\, 1[x_{1n} = x_1]}{1 - \sum_{y \in \mathcal{Y}} \frac{H_{yx_1}}{Q_{yx_1}} \mathcal{P}\{y|x_1, x_{2n}; \theta\}} \frac{\mathcal{P}\{y|x_1, x_{2n}; \theta\}}{Q_{yx_1}} \right\}, \tag{2.17}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \theta} = \sum_{n=1}^{N} \Bigg\{ & r_n \left[\frac{\partial \log \mathcal{P}\{y_n|x_{1n}, x_{2n}; \theta\}}{\partial \theta} - \frac{1}{Q_{y_n x_{1n}}} \sum_{x_2 \in \mathcal{X}_2} \omega_{x_{1n}|x_2} \pi_{x_2} \frac{\partial \mathcal{P}\{y_n|x_{1n}, x_2; \theta\}}{\partial \theta} \right] \\ & - \frac{(1 - r_n)}{1 - \sum_{y \in \mathcal{Y}} \frac{H_{yx_{1n}}}{Q_{yx_{1n}}} \mathcal{P}\{y|x_{1n}, x_{2n}; \theta\}} \left[\sum_{y \in \mathcal{Y}} \frac{H_{yx_{1n}}}{Q_{yx_{1n}}} \frac{\partial \mathcal{P}\{y|x_{1n}, x_{2n}; \theta\}}{\partial \theta} \right. \\ & \left. - \sum_{y \in \mathcal{Y}} \frac{H_{yx_{1n}}}{(Q_{yx_{1n}})^2} \mathcal{P}\{y|x_{1n}, x_{2n}; \theta\} \sum_{x_2 \in \mathcal{X}_2} \omega_{x_{1n}|x_2} \pi_{x_2} \frac{\partial \mathcal{P}\{y|x_{1n}, x_2; \theta\}}{\partial \theta} \right] \Bigg\}, \end{aligned} \tag{2.18}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \omega_{x_1|x_2}} = \sum_{n=1}^{N} \Bigg\{ & r_n \left[\frac{1[x_{1n} = x_1]1[x_{2n} = x_2]}{\omega_{x_1|x_2}} - \frac{1[x_{1n} = x_1]}{Q_{y_n x_1}} \mathcal{P}\{y_n|x_1, x_2; \theta\} \pi_{x_2} \right] \\ & + (1 - r_n) \left[\frac{1[x_{1n} = x_1]1[x_{2n} = x_2]}{\omega_{x_1|x_2}} + \right. \\ & \left. \frac{1[x_{1n} = x_1] \sum_{y \in \mathcal{Y}} \frac{H_{yx_1}}{(Q_{yx_1})^2} \mathcal{P}\{y|x_1, x_{2n}; \theta\} \mathcal{P}\{y|x_1, x_2; \theta\} \pi_{x_2}}{1 - \sum_{y \in \mathcal{Y}} \frac{H_{yx_1}}{Q_{yx_1}} \mathcal{P}\{y|x_1, x_{2n}; \theta\}} \right] \Bigg\} - \lambda_{x_2}. \end{aligned} \tag{2.19}$$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \pi_{x_2}} = \sum_{n=1}^{N} \Bigg\{ & r_n \left[\frac{1[x_{2n} = x_2]}{\pi_{x_2}} - \frac{1}{Q_{y_n x_{1n}}} \mathcal{P}\{y_n|x_{1n}, x_2; \theta\} \omega_{x_{1n}|x_2} \right] \\ & + (1 - r_n) \left[\frac{1[x_{2n} = x_2]}{\pi_{x_2}} + \right. \\ & \left. \frac{\sum_{y \in \mathcal{Y}} \frac{H_{yx_{1n}}}{(Q_{yx_{1n}})^2} \mathcal{P}\{y|x_{1n}, x_{2n}; \theta\} \mathcal{P}\{y|x_{1n}, x_2; \theta\} \omega_{x_{1n}|x_2}}{1 - \sum_{y \in \mathcal{Y}} \frac{H_{yx_{1n}}}{Q_{yx_{1n}}} \mathcal{P}\{y|x_{1n}, x_{2n}; \theta\}} \right] \Bigg\} - \mu, \end{aligned} \tag{2.20}$$

Moment Indicators

Rearrange (2.17) to obtain

$$\widehat{H}_{yx_1} = N^r_{yx_1}\widehat{Q}_{yx_1}\left[\sum_{n=1}^{N}\frac{(1-r_n)\,1[x_{1n}=x_1]\mathcal{P}\{y|x_1,x_{2n};\widehat{\theta}\}}{1-\sum_{y\in\mathcal{Y}}\frac{H_{yx_1}}{Q_{yx_1}}\mathcal{P}\{y|x_1,x_{2n};\widehat{\theta}\}}\right]^{-1} \tag{2.21}$$

Employing (2.17), the fourth term of (2.19) becomes

$$\sum_{y\in\mathcal{Y}}\frac{N^r_{yx_1}}{\widehat{Q}_{yx_1}}\mathcal{P}\{y|x_1,x_2;\widehat{\theta}\}\widehat{\pi}_{x_2} = \sum_{n=1}^{N} r_n 1[x_{1n}=x_1]\frac{1}{\widehat{Q}_{y_nx_1}}\mathcal{P}\{y_n|x_1,x_2;\widehat{\theta}\}\widehat{\pi}_{x_2}.$$

Thus, the second and the fourth terms of (2.19) cancel. Hence,

$$\widehat{\omega}_{x_1|x_2}\widehat{\lambda}_{x_2} = \sum_{n=1}^{N} 1[x_{1n}=x_1]1[x_{2n}=x_2]. \tag{2.22}$$

Multiplying (2.22) by $\widehat{\omega}_{x_1|x_2}$ and summing over $x_1 \in \mathcal{X}_1$ gives $\widehat{\lambda}_{x_2} = \sum_{n=1}^{N} I(x_{2n}=x_2) = N_{x_2}$. Substituting back into (2.22) yields

$$\widehat{\omega}_{x_1|x_2} = \frac{\sum_{n=1}^{N} 1[x_{1n}=x_1]1[x_{2n}=x_2]}{N_{x_2}}. \tag{2.23}$$

Moreover, one can write (2.21) as

$$\widehat{H}_{yx_{1n}} = N^r_{yx_{1n}}\widehat{Q}_{yx_{1n}}\left[\sum_{n'=1}^{N}\frac{(1-r_{n'})\,1[x_{1n'}=x_{1n}]\mathcal{P}\{y|x_{1n},x_{2n'};\widehat{\theta}\}}{1-\sum_{y\in\mathcal{Y}}\frac{H_{yx_{1n}}}{Q_{yx_{1n}}}\mathcal{P}\{y|x_{1n},x_{2n'};\widehat{\theta}\}}\right]^{-1}.$$

Also, from (2.23), $\widehat{\omega}_{x_{1n}|x_2}$ can be written as

$$\widehat{\omega}_{x_{1n}|x_2} = \frac{\sum_{n'=1}^{N} 1[x_{1n'}=x_{1n}]1[x_{2n'}=x_2]}{\sum_{n''=1}^{N} 1[x_{2n''}=x_2]}.$$

Substituting for $\widehat{\omega}_{x_{1n}|x_2}$ from (2.23) and comparing with (2.17) rewritten for $\widehat{H}_{yx_{1n}}$, the second and fourth terms of (2.20) are also zero. Hence, one can show that

$$\widehat{\pi}_{x_2} = N^{-1}\sum_{n=1}^{N} 1[x_{2n}=x_2].$$

Similarly, these expressions for $\widehat{H}_{yx_{1n}}$ and $\widehat{\omega}_{x_{1n}|x_2}$ demonstrate that the second and the fourth terms of (2.18) cancel. Thus (2.18) becomes

$$0 = \sum_{n=1}^{N}\{r_n \frac{\partial \log \mathcal{P}\{y|x_{1n}, x_{2n}; \widehat{\theta}\}}{\partial \theta} - \frac{(1-r_n)}{1 - \sum_{y\in\mathcal{Y}} \frac{\widehat{H}_{yx_{1n}}}{\widehat{Q}_{yx_{1n}}} \mathcal{P}\{y|x_{1n}, x_{2n}; \widehat{\theta}\}} \sum_{y\in\mathcal{Y}} \frac{\widehat{H}_{yx_{1n}}}{\widehat{Q}_{yx_{1n}}} \frac{\partial \mathcal{P}\{y|x_{1n}, x_{2n}; \widehat{\theta}\}}{\partial \theta}\}. \tag{2.24}$$

Finally, $\widehat{Q}_{yx_1}$ may be written as

$$\begin{aligned}
\widehat{Q}_{yx_1} &= \sum_{x_2\in\mathcal{X}_2} \mathcal{P}\{y|x_1, x_2; \widehat{\theta}\}\widehat{\omega}_{x_1|x_2}\widehat{\pi}_{x_2} \\
&= N^{-1}\sum_{n=1}^{N} 1[x_{1n} = x_1] \sum_{x_2\in\mathcal{X}_2} \mathcal{P}\{y|x_1, x_2; \widehat{\theta}\}1[x_{2n} = x_2] \\
&= N^{-1}\sum_{n=1}^{N} 1[x_{1n} = x_1]\mathcal{P}\{y|x_1, x_{2n}; \widehat{\theta}\}.
\end{aligned}$$

Thus, the GMM moment indicators are

$$\begin{aligned}
H_{st} &: r1[y = s]1[x_1 = t] - \frac{(1-r)\,1[x_1 = t]}{1 - \sum_{y\in\mathcal{Y}} \frac{H_{st}}{Q_{st}} \mathcal{P}\{s|t, x_2; \theta\}} \frac{H_{st}}{Q_{st}} \mathcal{P}\{s|t, x_2; \theta\}, s \in \mathcal{Y}, t \in \mathcal{X}_1; \\
\theta &: r\frac{\partial \log \mathcal{P}\{y|x_1, x_2; \theta\}}{\partial \theta} - \frac{(1-r)}{1 - \sum_{y\in\mathcal{Y}} \frac{H_{yx_1}}{Q_{yx_1}} \mathcal{P}\{y|x_1, x_2; \theta\}} \sum_{y\in\mathcal{Y}} \frac{H_{yx_1}}{Q_{yx_1}} \frac{\partial \mathcal{P}\{y|x_1, x_2; \theta\}}{\partial \theta}\}; \\
Q_{yt} &: Q_{yt} - 1[x_1 = t]\mathcal{P}\{y|t, x_2; \theta\}, y \in \mathcal{Y}.
\end{aligned} \tag{2.25}$$

2.4 Moment Indicators for Continuous Y and X

This section shows that the moment indicators obtained in the previous section can be modified in such a manner that they remain valid even when Y and X are continuous. The cost of the relaxation to allow inclusion of continuous Y and X is that some nuisance

functions must be estimated prior to the GMM estimation. Therefore, the approach taken here is based on two step GMM estimation.

2.4.1 INRYX

When Y and X are discrete, the sample proportion for $Y = y$, $X = x$ and $R = 1$, $\widehat{H}_{yx} = N^r_{yx}/N$, defines an empirical estimator for $\mathcal{P}\{Y = y, X = x, R = 1\}$. Hence, the sampling process enables us to estimate the quantities $\mathcal{P}\{Y = y, X = x, R = 1\}$ and also $\pi_x = \mathcal{P}\{X = x\}$ for all $y \in \mathcal{Y}$ and $x \in \mathcal{X}$. Note that we may re-express (2.9) as

$$\widehat{P}_{yx} = \widehat{H}_{yx}/(\mathcal{P}\{y|x;\widehat{\theta}\}\hat{\pi}_x). \tag{2.26}$$

Let $h(y, x, r = 1)$ denote the joint density function of Y, X and $R = 1$. Likewise, if Y and X are continuous, we can use either parametric or nonparametric methods to estimate $h(y, x, r = 1)$ from the sample. Similarly, since X is randomly drawn and is always observed, we can also estimate the marginal density function $f_X(x)$ for X.

For simplicity, we specify parametric models for $h(y, x, r = 1)$ and $f_X(x)$, rather than using a nonparametric approach. Let $h(y, x, r = 1; \widehat{\psi})$ and $f_X(x; \widehat{\alpha})$ denote the estimators for $h(y, x, r = 1)$ and $f_X(x)$. Thus, for the continuous case, adaption of the expression (2.26) for $\widehat{P}_{yx}$ yields an estimator for $\mathcal{P}\{R = 1|Y = y, X = x\}$ given by

$$\widehat{P}_{yx} = \frac{h(y, x, r = 1; \widehat{\psi})}{f(y|x; \widehat{\theta}) f_X(x; \widehat{\alpha})}. \tag{2.27}$$

Furthermore, we can re-write (2.5) for this setting as

$$0 = \sum_{n=1}^{N} \left\{ r_n \frac{\partial \log f(y_n|x_n; \widehat{\theta})}{\partial \theta} - (1 - r_n) \frac{1}{1 - \int_{\mathcal{Y}} \widehat{P}_{yx_n} f(y|x_n; \widehat{\theta}) dy} \int_{\mathcal{Y}} \widehat{P}_{yx_n} \frac{\partial f(y|x_n; \widehat{\theta})}{\partial \theta} dy \right\}. \tag{2.28}$$

Plugging (2.27) into (2.28), provides the moment indicator

$$\theta : r\frac{\partial \log f(y|x;\theta)}{\partial \theta} - (1-r)\frac{1}{1-\int_{\mathcal{Y}}(h(y,x,r=1;\psi)/f_X(x;\alpha))dy}\int_{\mathcal{Y}}\frac{h(y,x,r=1;\psi)}{f_X(x;\alpha)}\frac{\partial \log f(y|x;\theta)}{\partial \theta}dy. \tag{2.29}$$

Since $h(y,x,r=1)/f_X(x) = h(y,r=1|x)$, the conditional density of Y and R given X evaluated at $R=1$, $\mathcal{P}\{R=0|X=x\} = 1-\int_{\mathcal{Y}}(h(y,x,r=1)/f_X(x))dy$. Hence, it is straightforward to show that the population counterpart of (2.29) is

$$r\frac{\partial \log f(y|x;\theta_0)}{\partial \theta} - (1-r)\frac{1}{1-\int_{\mathcal{Y}}(h(y,x,r=1)/f_X(x))dy}\int_{\mathcal{Y}}\frac{h(y,x,r=1)}{f_X(x)}\frac{\partial \log f(y|x;\theta_0)}{\partial \theta}dy, \tag{2.30}$$

and has expectation with respect to the sample distribution equal to zero.[2] Therefore, the moment indicator (2.29), can be used for GMM estimation of θ_0 when Y and X are continuously distributed random vectors.

Because it requires correct parametric specifications for $f(y|x), h(y,x,r=1)$ and $f_X(x)$, a GMM estimator for θ_0 based on the moment indicator (2.29) may seem restrictive and may not be more attractive than the fully parametric approach. Nevertheless, the first-step parametric estimation is adopted here to simplify the discussion of asymptotic properties of the estimator provided in the next section. Theoretically, it is possible to estimate both $h(y,x,r=1)$ and $f_X(x)$ by nonparametric methods based only on sample information. Given such first-step nonparametric estimators, only correct specification of the population conditional density function $f(y|x;\theta)$ of Y given X is required for con-

[2] Note that $E[r\partial \log f(y|x;\theta_0)/\partial\theta] = \int_{\mathcal{Y}\times\mathcal{X}} h(y,x,r=1)\partial \log f(y|x;\theta_0)/\partial\theta dydx$.

sistent estimation of θ_0. This possibility represents a considerable advantage for GMM estimation over the fully parametric approach since it avoids parameterisation of both the MDM and the marginal density of the covariates, $f_X(x)$. A shortcoming of this GMM estimator, however, is that estimation of $h(y, x, r = 1)$ and $f_X(x)$ becomes more complex as the dimension of X increases.

2.4.2 INRY

Moment Indicators for Two-Step Estimation

Let $h(y, r = 1)$ denote the joint density function of Y and $R = 1$. Similarly to above, the INRY sampling process permits estimation of $h(y, r = 1)$ which is denoted here by $h(y, r = 1; \widehat{\psi})$ and is based on respondent units only. Rewrite (2.15) as[3]

$$0 = \sum_{n=1}^{N} \left\{ r_n \frac{\partial \log f(y_n|x_n;\widehat{\theta})}{\partial \theta} - \frac{(1-r_n)}{1 - \int_{\mathcal{Y}} \frac{h(y,r=1;\widehat{\psi})}{Q(y;\widehat{\theta})} f(y|x_n;\widehat{\theta}) dy} \left[\int_{\mathcal{Y}} \frac{h(y, r=1;\widehat{\psi})}{Q(y;\widehat{\theta})} \frac{\partial f(y|x_n;\widehat{\theta})}{\partial \theta} dy \right] \right\},$$

where $Q(y;\widehat{\theta}) = N^{-1} \sum_{n=1}^{N} f(y|x_n;\widehat{\theta})$. The corresponding moment indicator is

$$\theta : r \frac{\partial \log f(y|x;\theta)}{\partial \theta} - \frac{(1-r)}{1 - \int_{\mathcal{Y}} \frac{h(y,r=1;\psi)}{Q(y;\theta)} f(y|x;\theta) dy} \int_{\mathcal{Y}} \frac{h(y, r=1;\psi)}{Q(y;\theta)} \frac{\partial f(y|x;\theta)}{\partial \theta} dy. \tag{2.31}$$

It is straightforward to check that the expectation of (2.31) with respect to the sample distribution is equal to zero.

As noted above, the estimator $h(y, r = 1; \widehat{\psi})$ only uses information from respondent units which is quite different from the discrete case where information from non-respondent

---

[3] Cf. the first-order condition of the pseudolikelihood estimator in (2.45) given below. An advantage of this approach over (2.45) is that, by estimating $h(y, r = 1)$, the information on θ from non-respondent units is able to be incorporated into the estimation procedure.

units is also incorporated in the estimator $\widehat{H}_y$; see (2.16). Interestingly, we can show that the following relationship involving population quantities

$$r - (1-r)\frac{h(y, r=1)}{Q(y;\theta)} \frac{f(y|x;\theta)}{1 - \int_{\mathcal{Y}} \frac{h(y,r=1)}{Q(y;\theta)} f(y|x;\theta) dy}, \tag{2.32}$$

based on the moment indicator for H_y in (2.16), has expectation with respect to the sample distribution equal to zero. Hence, it is possible to combine the moment indicators (2.31) and (2.32) to estimate θ_0. This will move us from just-identified case to over-identified case.

Notice that, unlike INRYX, an increase in the dimension of X does not complicate first-step estimation, an advantage over the GMM estimator of section 2.4.1 that arises as a consequence of assuming the MDM (2.2) rather than (2.1).

Moment Indicators for One-Step Estimation

From (2.15), a possible consistent estimator for H_y is $N^{-1}\sum_{n=1}^{N} r \cdot 1[y_n = y]$. Substituting this estimator for H_y, expressions such as $\sum_{y\in\mathcal{Y}}[\widehat{H}_y/\widehat{Q}_y]\mathcal{P}\{y|x_n;\widehat{\theta}\}$ can be rewritten as

$$\begin{aligned}\sum_{y\in\mathcal{Y}} \frac{N^{-1}\sum_{n'=1}^{N} r_{n'} \cdot 1[y_{n'} = y]}{\widehat{Q}_y} \mathcal{P}\{y|x_n;\widehat{\theta}\} &= N^{-1}\sum_{n'=1}^{N} r_{n'} \left(\sum_{y\in\mathcal{Y}} 1[y_{n'} = y] \frac{\mathcal{P}\{y|x_n;\widehat{\theta}\}}{\widehat{Q}_y} \right) \\ &= N^{-1}\sum_{n'=1}^{N} r_{n'} \frac{\mathcal{P}\{y_{n'}|x_n;\widehat{\theta}\}}{\widehat{Q}_{y_{n'}}}.\end{aligned}$$

Thus, (2.15) can be re-expressed as

$$\begin{aligned}\sum_{n=1}^{N} \Bigg\{ & r_n \frac{\partial \log \mathcal{P}\{y_n|x_n;\widehat{\theta}\}}{\partial\theta} - \frac{(1-r_n)}{1 - \left(\frac{1}{N}\sum_{n'=1}^{N} r_{n'} \frac{\mathcal{P}\{y_{n'}|x_n;\widehat{\theta}\}}{\widehat{Q}_{y_{n'}}} \right)} \\ & \times \left[\frac{1}{N}\sum_{n''=1}^{N} r_{n''} \frac{1}{\widehat{Q}_{y_{n''}}} \frac{\partial \mathcal{P}\{y_{n''}|x_n;\widehat{\theta}\}}{\partial\theta} \right] \Bigg\}. \end{aligned} \tag{2.33}$$

Note that (2.33) no longer depends on H_y and remains valid even when Y is continuous; i.e., (2.33) may be written as

$$\sum_{n=1}^{N}\left\{r_n\frac{\partial \log f(y_n|x_n;\widehat{\theta})}{\partial\theta}-\frac{(1-r_n)}{1-(\frac{1}{N}\sum_{n'=1}^{N_r}\frac{f(y_{n'}|x_n;\widehat{\theta})}{Q(y_{n'};\widehat{\theta})})}(\frac{1}{N}\sum_{n''=1}^{N_r}\frac{1}{Q(y_{n''};\widehat{\theta})}\frac{\partial f(y_{n''}|x_n;\widehat{\theta})}{\partial\theta})\right\}, \tag{2.34}$$

where $Q(y;\widehat{\theta}) = N^{-1}\sum_{n=1}^{N} f(y|x_n;\widehat{\theta})$ as before. Hence, the corresponding moment indicator is

$$\theta : r\frac{\partial \log f(y|x;\theta)}{\partial\theta}-\frac{(1-r)}{1-(\frac{1}{N}\sum_{n'=1}^{N_r}\frac{f(y_{n'}|x;\theta)}{Q(y_{n'};\theta)})}(\frac{1}{N}\sum_{n''=1}^{N_r}\frac{1}{Q(y_{n''};\theta)}\frac{\partial f(y_{n''}|x;\theta)}{\partial\theta})$$

An crucial advantage of this moment indicator over (2.31) is that it does not require the first-step estimation of $h(y, r = 1)$.

2.4.3 INRYX1

Let $h(y, x_1, r = 1)$ denote the joint density function of Y, X_1 and $R = 1$ and let $h(y, x_1, r = 1;\widehat{\psi})$ denote a parametric estimator for $h(y, x_1, r = 1)$. Estimation of Q_{yx_1} is more complicated than that of Q_y in INRY because it involves an indicator function. If X_1 is discrete, we can use the estimator of Q_{yx_1} from the discrete setting, *viz.*

$$Q(y, x_1;\widehat{\theta}) = N^{-1}\sum_{n=1}^{N} 1[x_{1n} = x_1]f(y|x_1, x_{2n};\widehat{\theta}).$$

Whenever X_1 is continuous, one may estimate the sample density $h(y, x_1)$ using the following estimator

$$Q(y, x_1;\widehat{\theta}) = (Nh)^{-1}\sum_{n=1}^{N} I(x_1-\frac{h}{2}<x_{1n}<x_1+\frac{h}{2})f(y|x_1, x_{2n};\widehat{\theta}),$$

where h is a bandwidth parameter. Since X_1 and X_2 are fully observed, it is possible to specify a parametric model for X_1 conditional on X_2 and then to estimate the parameters

involved separately. Thus, $h(y, x_1)$ can be alternatively estimated using

$$Q(y, x_1; \widehat{\theta}, \widehat{\alpha}) = N^{-1} \sum_{n=1}^{N} f(y|x_1, x_{2n}; \widehat{\theta}) f(x_1|x_{2n}; \widehat{\alpha}),$$

where α parameterises the population density $f_{X_1|X_2}(x_1|x_2)$.

Thus, (2.24) can be rewritten for the continuous setting as

$$0 = \sum_{n=1}^{N} \{ r_n \frac{\partial \log f(y_n|x_n; \widehat{\theta})}{\partial \theta} - \frac{(1 - r_n)}{1 - \int_{\mathcal{Y}} \frac{h(y, x_{1n}; \widehat{\psi})}{Q(y, x_{1n}; \widehat{\theta})} f(y|x_n; \widehat{\theta}) dy} \int_{\mathcal{Y}} \frac{h(y, x_{1n}; \widehat{\psi})}{Q(y, x_{1n}; \widehat{\theta})} \frac{\partial f(y|x_n; \widehat{\theta})}{\partial \theta} dy \}, \tag{2.35}$$

and the corresponding moment indicator is

$$\theta : r \frac{\partial \log f(y|x; \theta)}{\partial \theta} - \frac{(1 - r)}{1 - \int_{\mathcal{Y}} \frac{h(y, x_1; \psi)}{Q(y, x_1; \theta)} f(y|x; \theta) dy} \int_{\mathcal{Y}} \frac{h(y, x_1; \psi)}{Q(y, x_1; \theta)} \frac{\partial f(y|x; \theta)}{\partial \theta} dy. \tag{2.36}$$

Like INRYX, the estimators of the nuisance functions in INRYX1 become more complex as the dimension of X increases. However, first-step estimation in INRYX1 is less complicated than that in INRYX since $\dim(x_1) < \dim(x)$ which justifies interest in the GMM estimator for INRYX1. Consider a situation where first-step estimation is nonparametric and X is of large dimension. INRYX1 simplifies the estimation procedures in INRYX by restricting the MDM to depend on the subvector of X_1 and using INRYX1-GMM instead of INRYX-GMM estimation.

2.5 Asymptotic Properties

Since the one-step GMM estimator for INRY in section 2.4.2 does not require an initial estimation, its asymptotic properties can therefore be proved using standard arguments. Thus, the proof is omitted.

For the other cases, we specify parametric models for the nuisance functions. Therefore, ML delivers a natural procedure for first-step estimation. Let $\gamma \in \Gamma$ denote the parameters involved in the initial stage. These parameters are $(\alpha', \psi')'$ for INRYX and ψ for INRY and INRYX1. Let W denote the triple (Y, X, R) and $\mathcal{W} = \mathcal{Y} \times \mathcal{X} \times \{0, 1\}$ the sample space. Also let $d(w, \gamma)$ denote the vector of related parametric density functions. For instance, $d(w, \gamma) = [f_X(x; \alpha), r \cdot h(y, x, r = 1; \psi)]'$ in INRYX. Thus, the ML estimator $\widehat{\gamma}$ satisfies the condition

$$N^{-1} \sum_{n=1}^{N} \frac{\partial \log d(w_n, \gamma)}{\partial \gamma} = 0. \tag{2.37}$$

Let $g(w; \theta, \gamma)$ denote the vector of moment indicators from either (2.29), (2.31) or (2.36). The GMM estimator, $\widehat{\theta}$, solves

$$N^{-1} \sum_{n=1}^{N} g(w_n; \theta, \widehat{\gamma}) = 0. \tag{2.38}$$

One can interpret this two-step GMM estimator as a joint GMM estimator by stacking (2.37) and (2.38) to form $\widetilde{g}(w, \theta, \gamma) = [\partial \log d(w, \gamma)/\partial \gamma', g(w, \theta, \gamma)']'$. Let $g(w) = g(w, \theta_0, \gamma_0)$, $G_\theta = E[\partial g(w)/\partial \theta]$, $G_\gamma = E[\partial g(w)/\partial \gamma]$, $D = E[\partial^2 \log d(w, \gamma_0)/\partial \gamma \partial \gamma']$ and $\xi(w) = -D^{-1}[\partial \log d(w, \gamma_0)/\partial \gamma]$. The following assumptions describe standard regularity conditions which are sufficient for the consistency and asymptotic normality of $(\theta', \gamma')'$. See Newey and McFadden (1994, Theorems 2.6, 3.4 and 6.1)

Assumption 2.5.1: *Suppose that* $\{y_n, x_n, r_n\}_{n=1}^N$ *are i.i.d. and (i)* $(\theta_0', \gamma_0')' \in int(\Theta \times \Gamma)$ *where* Θ and Γ *are compact; (ii)* $d(w, \gamma)$ *is twice continuously differentiable in* $\gamma \in \Gamma$;

(iii) $d(w;\gamma)$ *and* $\partial d(w;\gamma)/\partial\gamma$ *are continuous at each* $\gamma \in \Gamma$*; (iv)* $d(w,\gamma) > 0$ *for all* $w \in \mathcal{W}$ and γ *in an open neighbourhood of* γ_0.

Assumption 2.5.2: *(i)* $f(y|x;\theta)$ *is twice continuously differentiable in* $\theta \in \Theta$*; (ii)* $f(y|x;\theta)$ *and* $\partial f(y|x;\theta)/\partial\theta$ *are continuous at each* $\theta \in \Theta$*; (iii)* $f(y|x;\theta) > 0$ *for all* $y \in \mathcal{Y}$, $x \in \mathcal{X}$ *and* θ *in an open neighbourhood of* θ_0*; (iv)* $f_X(x) > 0$ *for all* $x \in \mathcal{X}$.

Assumption 2.5.3: *(i)* $(\theta_0', \gamma_0')'$ *is the unique solution to* $E[\widetilde{g}(w,\theta_0,\gamma_0)] = 0$ *and* $E[\partial\widetilde{g}(w,\theta_0,\gamma_0)/\partial(\theta',\gamma')]$ *is full column rank; (ii)* $E[\sup_{(\theta',\gamma')'} \|\widetilde{g}(w,\theta,\gamma)\|^2] < \infty$ *and* $E[\sup_{(\theta',\gamma')'\in\mathcal{N}} \|\partial\widetilde{g}(w,\theta,\gamma)/\partial(\theta',\gamma')\|] < \infty$ *where* N *is a neighbourhood of* $(\theta_0', \gamma_0')'$ *in* $\Theta \times \Gamma$*; (iii)* $\widetilde{G}'\widetilde{G}$ *is nonsingular for* $\widetilde{G} = E[\partial\widetilde{g}(w,\theta_0,\gamma_0)/\partial(\theta',\gamma')']$.

Assumption 2.5.1(iv) implies that the sampling process is valid because $h(y, r = 1)$, $h(y, x_1, r = 1)$ and $h(y, x, r = 1)$ are strictly positive. Assumptions 2.5.2(iii) and (iv) ensure that Q_y and Q_{yx_1} are also strictly positive. For INRYX, Assumption 2.5.2(iv) can be dropped since $f_X(x;\alpha) > 0$ by Assumption 2.5.1(iv). All other conditions are standard for a joint GMM estimator. These conditions lead to the following result.

Theorem 2.5: *(Consistency and Asymptotic Normality.) If Assumptions 2.5.1-2.5.3 are satisfied then* $\widehat{\theta} \xrightarrow{p} \theta_0$ *and* $\widehat{\gamma} \xrightarrow{p} \gamma_0$. *Also,* $\widehat{\theta}$ *and* $\widehat{\gamma}$ *are asymptotic normal and, in particular,*

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N(0, V),$$

where $V = G_\theta^{-1} E[\{g(w) + G_\gamma\xi(w)\}\{g(w) + G_\gamma\xi(w)\}']G_\theta^{-1\prime}$ *and where* $\xrightarrow{p}$ *and* $\xrightarrow{d}$ *denote convergence in probability and distribution respectively.*

The proof of this theorem follows exactly that for Newey and McFadden (1994, Theorem 6.1).

2.6 INRYX and the RS Approach

Let Y and X be continuously distributed with respective sample spaces $\mathcal{Y}$ and $\mathcal{X}$. Let $\mathcal{Y}_i$, $i \in \mathcal{I}, \mathcal{I} = \{1, ..., C\}$, and $\mathcal{X}_j$, $j \in \mathcal{J}, \mathcal{J} = \{1, ..., M\}$, be finite partitions of $\mathcal{Y}$ and $\mathcal{X}$.

RS applies the methodology of Imbens (1992) for choice-based sampling to deal with the missing data problem when Y is discrete and the MDM is (2.2). Moreover, RS suggests that this approach can be extended to cover the case when Y and X are continuous and the MDM is

$$\begin{aligned} P_{ij} &= \mathcal{P}\{R = 1 | Y = y, X = x\}. \\ &= \mathcal{P}\{R = 1 | Y \in \mathcal{Y}_i, X \in \mathcal{X}_j\}, i \in \mathcal{I}, j \in \mathcal{J}. \end{aligned} \tag{2.39}$$

In other words, although Y and X are continuous, P_{ij} is constant within a particular set $\mathcal{Y}_i \times \mathcal{X}_j$ of values of Y and X.

In this section, we derive an estimator for (2.39) using RS's approach that is then compared to the two-step GMM estimator for INRYX since the MDM in (2.39) is a special case of (2.1).

2.6.1 The RS GMM Estimator

Given (2.39), the observed data likelihood for a respondent unit is

$$\begin{aligned}\mathcal{P}\{Y=y, X=x, R=1\} &= P_{ij} f(y|x;\theta) f_X(x) \\ &= \frac{H_{ij}}{Q_{ij}} f(y|x;\theta) f_X(x), y \in \mathcal{Y}_i, x \in \mathcal{X}_j,\end{aligned}$$

where $H_{ij} = \mathcal{P}\{Y \in \mathcal{Y}_i, X \in \mathcal{X}_j, R=1\}$ and $Q_{ij} = \int_{\mathcal{Y}_i \times \mathcal{X}_j} f(y|x;\theta) f_X(x) dy dx$.

Define the indicator $I_i = 1$ if $Y \in \mathcal{Y}_i$ and 0 otherwise. Then the conditional probability of $I_i = 1$ given X is

$$R(i|x;\theta) = \int_{\mathcal{Y}_i} f(y|x;\theta) dy.$$

Hence $Q_{ij} = \int_{\mathcal{X}_j} R(i|x;\theta) f_X(x) dx$. Let $H_{ij}^{nr} = \mathcal{P}\{Y \in \mathcal{Y}_i, X \in \mathcal{X}_j, R=0\}$. Then, the observed data likelihood for a non-respondent unit is

$$\begin{aligned}\mathcal{P}\{X=x, R=0\} &= \sum_{i \in \mathcal{I}} \int_{\mathcal{Y}_i} \frac{H_{ij}^{nr}}{Q_{ij}} f(y|x;\theta) f_X(x) dy \\ &= (1 - \sum_{i \in \mathcal{I}} \frac{H_{ij}}{Q_{ij}} R(i|x;\theta)) f_X(x), \quad x \in \mathcal{X}_j.\end{aligned}$$

Accordingly, the joint observed data likelihood of a sample unit is

$$\left[\frac{H_{ij}}{Q_{ij}} f(y|x;\theta) f_X(x)\right]^r [(1 - \sum_{i \in \mathcal{I}} \frac{H_{ij}}{Q_{ij}} R(i|x;\theta)) f_X(x)]^{1-r}.$$

Following Cosslett (1981) and Imbens (1992) we regard X as discrete with support $\mathcal{X}$, each mass point associated with probability mass $\mathcal{P}\{X=x\} = \pi_x$, $x \in \mathcal{X}$. Thus, the objective function based on the above likelihood is

$$\begin{aligned}\mathcal{L} &= \sum_{n=1}^{N}\{r_n[\log H_{i_n j_n} + \log f(y_n|x_n;\theta) + \log \pi_{x_n} - \log Q_{i_n j_n}] \\ &\quad + (1-r_n)\left[\log(1 - \sum_{i\in\mathcal{I}} \frac{H_{ij_n}}{Q_{ij_n}} R(i|x_n,\theta)) + \log \pi_{x_n}\right]\} \\ &\quad -\mu(\sum_{x\in\mathcal{X}} \pi_x - 1),\end{aligned}$$

where $Q_{ij} = \sum_{x\in\mathcal{X}_j} R(i|x,\theta)\pi_x$.

The following moment conditions may be obtained from the above objective function, the detailed derivation of which is given in the appendix:

$$\theta : 0 = \sum_{n=1}^{N}\left(r_n \frac{\partial \log f(y_n|x_n;\widehat{\theta})}{\partial\theta} - \frac{(1-r_n)}{1 - \sum_{i\in\mathcal{I}} \frac{\widehat{H}_{ij_n}}{\widehat{Q}_{ij_n}} R(i|x_n,\widehat{\theta})} \sum_{i\in\mathcal{I}} \frac{\widehat{H}_{ij_n}}{\widehat{Q}_{ij_n}} \frac{\partial R(i|x_n,\widehat{\theta})}{\partial\theta}\right), \tag{2.40}$$

$$H_{ij} : 0 = \sum_{n=1}^{N}\left(r_n \frac{1[i_n = i]\cdot 1[j_n = j]}{H_{ij}} - (1-r_n)\frac{1[j_n = j]}{1 - \sum_{i\in\mathcal{I}} \frac{\widehat{H}_{ij}}{\widehat{Q}_{ij}} R(i|x_n,\widehat{\theta})} \frac{R(i|x_n,\widehat{\theta})}{\widehat{Q}_{ij}}\right), \tag{2.41}$$

$$Q_{ij} : 0 = N\widehat{Q}_{ij} - \sum_{n=1}^{N} 1\left[x_n \in \mathcal{X}_j\right] R(i|x_n,\widehat{\theta}). \tag{2.42}$$

2.6.2 Comparison

None of (2.40), (2.41), and (2.42) depend on the marginal distribution of X. Thus, although X is assumed to be discrete with finite support in the above derivation, these moment indicators remain applicable even when X is continuously distributed. We now compare

the RS GMM estimator (RSGMM) to the two-step GMM estimator for INRYX (INRYX-GMM) from section 2.4.1.

In term of the sample density function $h(y, x, r = 1)$, RSGMM estimates H_{ij} using (2.41) whereas, in INRYX-GMM, $h(y, x, r = 1)$ must be specified and estimated. This suggests that RSGMM is more robust than INRYX-GMM in this regard. Nonetheless, a nonparametric estimator for $h(y, x, r = 1)$ obtained directly from sample respondents may be used in the first-step. Thus, it is possible to use the moment indicator for θ from INRYX-GMM without specifying a parametric model for $h(y, x, r = 1)$. The prospect of nonparametric first-step estimation makes INRYX-GMM competitive with RSGMM in this respect.

The main advantage of RSGMM over INRYX-GMM is that it does not require the specification and estimation of $f_X(x)$. The moment indicators for RSGMM depend loosely on the marginal distribution of X through the indicator function $1[j_n = j]$. To use RSGMM, we only need to know from which set in the partition of $\mathcal{X}$ each observation was drawn, but not the density function, $f_X(x)$, itself, since the MDM is assumed constant within $\mathcal{Y}_i \times \mathcal{X}_j$, $i \in \mathcal{I}, j \in \mathcal{J}$.

On the other hand, (2.39) is more restrictive than (2.1) which underpins INRYX-GMM. Thus, by comparing these two estimators, we can see that the cost of relaxing (2.39) to (2.1) is the specification and estimation of $f_X(x)$. This cost may, however, be acceptable if one can use a nonparametric method to estimate $f_X(x)$ from the data which is possible because there are no missing values in X.

To sum up, RSGMM and INRYX-GMM are operable under different assumptions. The decision whether to use RSGMM or INRYX-GMM depends on one's preference concerning these assumptions. If one believes that (2.39) is unrealistic and nonparametric first-step estimation is adopted, then INRYX-GMM may be preferable to RSGMM.

2.7 INRY, RS and TLR

TLR propose pseudolikelihood (PL) estimators for dealing with missing data in a INRY setting. RS develop GMM estimation for the special case when Y is discrete *or*, if Y is continuously distributed, the MDM depends on a finite partition of $\mathcal{Y}$. In this section, we demonstrate a connection between PL, RS GMM and the two-step INRY-GMM estimators of Section 2.4.2. To simplify the notation, we again use RSGMM estimator to denote this particular RS estimator.

2.7.1 Pseudolikelihood Estimators

Under (2.2), respondent units are a random sample from the population distribution of X given Y. Thus, inference may be based on the following observed data likelihood

$$\prod_{n=1}^{N} \left[\frac{f(y_n|x_n;\theta) f_X(x_n)}{\int_{\mathcal{X}} f(y_n|x;\theta) f_X(x) dx} \right]^{r_n}.$$

TLR propose a two-step procedure for estimating θ: (i) estimate the marginal density of X, $f_X(x)$, based on the full sample, and (ii) replace $f_X(x)$ with the estimator $\widehat{f}_X(x)$ and maximise the resultant pseudolikelihood with respect to θ. Different methods of estimating $f_X(x)$ lead to a different PL estimator. Among these PL estimators, TLR show that the

most efficient one is the one where $f_X(x)$ is estimated by its empirical counterpart, i.e., the most efficient PL estimator maximises

$$\mathcal{L} = \sum_{n=1}^{N} r_n \log f(y_n|x_n;\theta) - \sum_{n=1}^{N} r_n \log[\int_{\mathcal{X}} f(y_n|x;\theta) d\widehat{F}_n(x)], \tag{2.43}$$

where $\widehat{F}_n(x) = N^{-1}\sum_{n=1}^{N} 1[x_n \leq x]$. Since it is more efficient than the others, we use the maximiser of (2.43) as the benchmark for any comparison between pseudolikelihood and other estimators.

An alternative procedure, which yields an equivalent expression to (2.43), to the two-step estimation procedure in TLR is as follows. Suppose X is discrete with support $\mathcal{X}$, with each mass point having associated probability mass $\mathcal{P}\{X = x\} = \pi_x$, $x \in \mathcal{X}$. Instead of using $\widehat{F}_n(x)$, one can directly estimate π_x using the empirical estimator $\widehat{\pi}_x = \sum_{n=1}^{N} 1[x_n = x]/N$. Accordingly, the estimator of $Q_y = \int_{\mathcal{X}} f(y|x;\theta) f_X(x) dx$ is

$$\begin{aligned} \widehat{Q}_y &= \sum_{x\in\mathcal{X}} \widehat{\pi}_x \cdot f(y|x;\theta) \\ &= \frac{1}{N}\sum_{n=1}^{N}\sum_{x\in\mathcal{X}} 1[x_n = x] f(y|x;\theta) = \frac{1}{N}\sum_{n=1}^{N} f(y|x_n;\theta), y \in \mathcal{Y}. \end{aligned}$$

Since the estimator $\widehat{Q}_y$ is valid even if X is continuous because it does not depend on the nuisance parameter π_x, an alternative expression to (2.43) is

$$\mathcal{L} = \sum_{n=1}^{N} r_n \log f(y_n|x_n;\theta) - \sum_{n=1}^{N} r_n \log[\frac{1}{N}\sum_{n'=1}^{N} f(y_n|x_{n'};\theta)]. \tag{2.44}$$

The maximisers of (2.43) and (2.44) are theoretically identical. An advantage of using (2.44) is that it does not require the first-step estimation of $f_X(x)$. The PL estimator in (2.44) resembles a Maximum Simulated Likelihood (MSL) estimator where the direct Monte Carlo integral estimator is used. The only difference is that the values of X are from

the actual sample, rather than arising from a Monte Carlo sample drawn from a known density of X.

From (2.44), the first-order condition with respect to θ is

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{n=1}^{N} r_n \left(\frac{\partial \log f(y_n|x_n;\widehat{\theta})}{\partial \theta} - \frac{1}{\sum_{n'=1}^{N} f(y_n|x_{n'};\widehat{\theta})} \sum_{n''=1}^{N} \frac{\partial f(y_n|x_{n''};\theta)}{\partial \theta} \right) = 0. \tag{2.45}$$

If pseudolikelihood estimation is applied to a case where Y and X are *discrete*, then (2.45) becomes

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{n=1}^{N} r_n \left(\frac{\partial \log \mathcal{P}\{y_n|x_n;\widehat{\theta}\}}{\partial \theta} - \frac{1}{\sum_{n'=1}^{N} \mathcal{P}\{y_n|x_{n'};\widehat{\theta}\}} \sum_{n''=1}^{N} \frac{\partial \mathcal{P}\{y_n|x_{n''};\widehat{\theta}\}}{\partial \theta} \right), \tag{2.46}$$

Thus, the PL estimator can be interpreted as a GMM estimator whose moment indicators are

$$\begin{aligned} \theta &: r \left(\frac{\partial \log \mathcal{P}\{y|x;\theta\}}{\partial \theta} - \frac{1}{Q_y} \sum_{n''=1}^{N} \frac{\partial \mathcal{P}\{y|x_{n''};\theta\}}{\partial \theta} \right); \\ Q_y &: Q_y - \mathcal{P}\{y|x;\theta\}, y \in \mathcal{Y}. \end{aligned} \tag{2.47}$$

Equations (2.45), (2.46) and (2.47) are useful to compare this PL estimator to the RS GMM estimator.

2.7.2 Comparison

The RSGMM estimator in this setting is in fact the GMM estimator based on moment indicators (2.16) in Section 2.3.2. This establishes a connection between the RSGMM and INRY-GMM estimators. Furthermore, notice that although the derivation in section 2.3.2 is carried out by assuming that X is discrete, (2.16) remains valid whether X is continuous or discrete since they are independent of the marginal distribution of X.

To compare the RSGMM estimator to the PL estimator, notice that, by using the ML estimators for π_x and Q_y in section 2.3.2, one can re-express the terms associated with the respondent indicator r_n in (2.13) from section 2.3.2 as

$$\sum_{n=1}^{N} r_n \left(\frac{\partial \log \mathcal{P}\{y_n|x_n;\widehat{\theta}\}}{\partial \theta} - \frac{1}{\sum_{n'=1}^{N} \mathcal{P}\{y_n|x_{n'};\widehat{\theta}\}} \sum_{n''=1}^{N} \frac{\partial \mathcal{P}\{y_n|x_{n''};\widehat{\theta}\}}{\partial \theta} \right). \tag{2.48}$$

A comparison of (2.48) and (2.46) confirms that they are identical.

Moreover, the expectation of (2.48) with respect to the sample distribution is zero. Thus, rather than using (2.16), H_y, θ and Q_y, $y \in \mathcal{Y}$, can be consistently estimated using the following GMM moment indicators.

$$\begin{aligned} H_t &: r1[y=t] - \frac{H_t}{Q_t} \frac{(1-r)\mathcal{P}\{t|x;\theta\}}{1-\sum_{y\in\mathcal{Y}} \frac{H_y}{\widehat{Q}_y} \mathcal{P}\{y|x;\theta\}}, t \in \mathcal{Y}; \\ \theta &: r \left(\frac{\partial \log \mathcal{P}\{y|x;\theta\}}{\partial \theta} - \frac{1}{Q_y} \sum_{m=1}^{N} \frac{\partial \mathcal{P}\{y|x_m;\theta\}}{\partial \theta} \right); \\ Q_y &: Q_y - \mathcal{P}\{y|x;\theta\}, y \in \mathcal{Y}. \end{aligned} \tag{2.49}$$

The moment indicators for θ and Q_y in (2.49) are identical to those in (2.47). The additional moment indicators in (2.49) are for the estimation of the extra parameters H_y, $y \in \mathcal{Y}$. Since the moment indicators for θ and Q_y are independent of H_y, one can ignore the first set of moment indicators in (2.49) if the focus is on θ. This means that, in terms of estimating θ, using the GMM estimator based on (2.49) is the same as using the PL estimator in (2.46).

The RSGMM estimator is based on moment indicators in (2.16), rather than those in (2.49), because they lead to more efficient GMM estimation. In deriving (2.49), we neglect the information on θ from non-respondent units. Thus, using (2.16) is more efficient than using (2.49) because we extract information from the full likelihood, i.e., the likelihood

functions of both respondent and nonrespondent units. Since the GMM estimator based on (2.49) is as efficient as the PL estimator in (2.46), the RSGMM estimator must be more efficient than the PL estimator whenever Y and X are discrete. Moreover, this efficiency argument comparing the approaches in RS and TLR must remain true if only X is allowed to be continuous. In such case, RS has shown that the RSGMM estimator attains the semiparametric efficiency bound.

However, if both Y and X are continuously distributed and the MDM is (2.2), the RSGMM estimator is no longer appropriate whereas the PL estimator is still applicable. In such a case, the PL estimator should be compared to the INRY-GMM estimator. By considering (2.44) or (2.45) against (2.31), one can see that both estimators estimate $f_X(x)$ and Q_y by $N^{-1}\sum_{n=1}^{N} 1[x_n = x]$ and $N^{-1}\sum_{n=1}^{N} f(y|x_n;\theta)$ respectively. Nonetheless, by additionally estimating $h(y, r = 1)$, the INRY-GMM estimator can incorporate the information on θ from non-respondent units into the estimation procedure. Thus, we suspect that the INRY-GMM estimator is semiparametrically more efficient than the PL estimator in (2.44). Even though we do not have a mathematical proof of this conjecture at present, the Monte Carlo evidence supporting it is presented in Chapter 3.

2.8 Identification

We discuss two kinds of identifying assumptions, nonparametric and parametric. We show that, although identification in RS and TLR is a combination of the two, they rely more on the latter identifying assumption. Moreover, we discuss the relationship between all esti-

mators discussed above and the score functions of various fully parametric ML estimators. Then, primitive conditions for identification and some related conditions are presented.

2.8.1 Nonparametric and Parametric Identifying Assumptions

The objective of any estimator which is based on a sample with missing data is to extract information on the population conditional distribution of Y given X from the joint distribution of Y, X and R. Manski (2003) notes the following total probability relationship

$$\mathcal{P}\{Y|X=x\} = \mathcal{P}\{Y|X=x,R=1\}\mathcal{P}\{R=1|X=x\} \tag{50}$$
$$+\mathcal{P}\{Y|X=x,R=0\}\mathcal{P}\{R=0|X=x\}.$$

Except for $\mathcal{P}\{Y|X=x,R=0\}$, all quantities on the right hand side of (2.50) are nonparametrically identified from an observed sample of Y, X and R. Hence, the fundamental problem of the inference with missing values (on the dependent variable) is that the sampling process cannot reveal any information on $\mathcal{P}\{Y|X=x,R=0\}$. To obtain point-identification, distributional assumptions must therefore be imposed.[4] In this subsection, we consider two ways of solving this identification problem. One solution leads to nonparametric identification whereas the other leads to parametric identification.

MAR is a distributional assumption which can yield nonparametric identification and it can be stated formally as follows.

---

[4] Manski (2003) studies partial identification of population parameters avoiding maintaining strong distributional assumptions for identification by either imposing no restrictions or adopting those which are weak or are refutable. Such approaches only place the parameter of interest within a set-valued identification region. Hence, identification in the usual sense is referred to as point-identification.

Assumption MAR: (*Missing At Random*)

$$\mathcal{P}\{R=1|Y=y,X=x\}=\mathcal{P}\{R=1|X=x\},x\in\mathcal{X}.$$

This assumption maintains a statistical independence between missingness R and values of Y. The MAR assumption implies

$$\mathcal{P}\{Y|X=x\}=\mathcal{P}\{Y|X=x,R=1\}=\mathcal{P}\{Y|X=x,R=0\}. \tag{2.51}$$

In fact, (2.51) $\Leftrightarrow$ MAR. The relationship in (2.51) clearly solves the identification problem in (2.50). One can, as a consequence of MAR, make inferences regarding $\mathcal{P}\{Y|X=x\}$ directly from $\mathcal{P}\{Y|X=x,R=1\}$. See Wooldridge (2002a, section 5) for a discussion of this point in the context of M-estimation.

Manski (2003) refers to (2.51) as Outcomes Missing at Random and refers to MAR as a type of Statistical Independence (SI) assumption. We choose not to distinguish between MAR and (2.51) because they are equivalent. Assumption MAR is presented here to show that an exclusion restriction or SI imposed on the MDM can yield nonparametric point-identification.

While maintaining MAR, one may also choose to specify a parametric family for $\mathcal{P}\{Y|X=x\}$ and estimate the associated parameters from respondent units, i.e., from $\mathcal{P}\{Y|X=x,R=1\}$. However, the identification problem is not solved by specifying a parametric family, but rather by maintaining MAR.

The second approach in (2.50) is to parameterise fully the observed data likelihood. Suppose the joint distribution of $\mathcal{P}\{Y,X,R\}$ is factorised into three parts, $\mathcal{P}\{R|Y,X\}$,

$\mathcal{P}\{Y|X\}$ and $\mathcal{P}\{X\}$. The joint observed data likelihood of a sample unit is thus given by

$$[f(r=1|y,x;\psi)f(y|x;\theta)f_X(x;\alpha)]^r\left[\left(1-\int_{\mathcal{Y}} f(r=1|y,x;\psi)f(y|x;\theta)dy\right)f_X(x;\alpha)\right]^{1-r},$$

where $f(r=1|\cdot,\cdot;\psi)$, $f(\cdot|\cdot;\theta)$ and $f_X(\cdot;\alpha)$ denote known parametric density functions with associated parameters ψ, θ and α respectively. The first-order conditions with respect to ψ, θ and α for the random sample $\{y_n,x_n,r_n\}_{n=1}^N$ are

$$\begin{aligned}\frac{\partial\mathcal{L}}{\partial\psi} &= \sum_{n=1}^{N}\left\{r_n\frac{\partial\log f(r=1|y_n,x_n;\widehat{\psi})}{\partial\psi}-\right.\\ &\left.\frac{1-r_n}{1-\int_{\mathcal{Y}}f(r=1|y,x_n;\widehat{\psi})f(y|x_n;\widehat{\theta})dy}\int_{\mathcal{Y}}f(y|x_n;\widehat{\theta})\frac{\partial f(r=1|y,x_n;\widehat{\psi})}{\partial\psi}dy\right\}=0;\\ \frac{\partial\mathcal{L}}{\partial\theta} &= \sum_{n=1}^{N}\left\{r_n\frac{\partial\log f(y_n|x_n;\widehat{\theta})}{\partial\theta}-\right.\\ &\left.\frac{1-r_n}{1-\int_{\mathcal{Y}}f(r=1|y,x_n;\widehat{\psi})f(y|x_n;\widehat{\theta})dy}\int_{\mathcal{Y}}f(r=1|y,x_n;\widehat{\psi})\frac{\partial f(y|x_n;\widehat{\theta})}{\partial\theta}dy\right\}=0;\end{aligned} \tag{2.52}$$

$$\frac{\partial\mathcal{L}}{\partial\alpha}=\sum_{n=1}^{N}\frac{\partial\log f_X(x;\widehat{\alpha})}{\partial\alpha}=0.$$

The derivation of the above first-order conditions with respect to ψ and θ assumes that all related functions are smooth enough to allow the interchangeability between integration and differentiation. This is not an issue if Y and X are discrete.

In the full-parametric approach, one calculates $\widehat{\psi}$, $\widehat{\theta}$ and $\widehat{\alpha}$ simultaneously using the first-order conditions in (2.52). Notice that α can be estimated independently because X is fully observed. However, ψ and θ can never be separately estimated using the observed sample because of the missing values in Y. In contrast to the result based on Assumption MAR in (2.51), since the MDM is now NMAR, $\mathcal{P}\{Y|X=x,R=1\}\neq\mathcal{P}\{Y|X=x,R=0\}$ here; the sample distributions of Y given X among respondent and non-respondent units

are

$$g(y|x, r = 1) = \frac{f(r = 1|y, x; \psi) f(y|x; \theta)}{\int_y f(r = 1|y, x; \psi) f(y|x; \theta) dy},$$

and

$$g(y|x, r = 0) = \frac{(1 - f(r = 1|y, x; \psi)) f(y|x; \theta)}{1 - \int_y f(r = 1|y, x; \psi) f(y|x; \theta) dy},$$

where $g(\cdot)$ denotes a generic sample density function.

2.8.2 Identification with Missing Data

The identification problem in RS and TLR is solved by a combination of nonparametric and parametric identifying assumptions. The power of the nonparametric identifying assumption comes from asserting (2.2), which is inverts Assumption MAR. Under (2.2), one can show that

$$\mathcal{P}\{X|Y = y\} = \mathcal{P}\{X|y = y, R = 1\} = \mathcal{P}\{X|Y = y, R = 0\}. \tag{2.53}$$

This relationship implies that $\mathcal{P}\{X|Y = y\}$ is nonparametrically identified from the observed sample. To recover information on $\mathcal{P}\{Y|X = x\}$ from $\mathcal{P}\{X|Y = y\}$ is, by itself, another form of identification problem. To show this, let the sample space $\mathcal{Y} \times \mathcal{X}$ be finite. By Bayes Theorem and the Law of Total Probability, we can write

$$\mathcal{P}\{Y = y|X = x\} = \frac{\mathcal{P}\{X = x|Y = y\}\mathcal{P}\{Y = y\}}{\sum_{y' \in \mathcal{Y}} \mathcal{P}\{X = x|Y = y'\}\mathcal{P}\{Y = y'\}}. \tag{2.54}$$

The sampling process under (2.2) is informative about $\mathcal{P}\{X = x|Y = y\}$ but does not reveal $\mathcal{P}\{Y = y\}$ due to the missing values in Y. This identification problem is the subject of the choice-based or response-based sampling literature. Thus, we can conclude that

although (2.2) cannot solve the identification problem in (2.50) in itself, it reduces the identification missing data problem to that arising in the choice-based sampling literature.

The full parametric approach provides a simple solution to the identification problem highlighted in (2.54). Let $f(\cdot|\cdot;\theta)$ and $f_X(\cdot;\alpha)$ denote parametric density functions known up to the finite-dimensional parameters θ and α. Under (2.2), the observed data likelihood of a respondent unit is given by

$$\mathcal{P}\{X|y=y,R=1\}=\frac{f(y|x;\theta_0)f_X(x;\alpha_0)}{\int_{\mathcal{X}}f(y|x;\theta_0)f_X(x;\alpha_0)dx}.$$

ML can then be applied to estimate θ_0 and α_0 simultaneously. This approach combines nonparametric and parametric identifying assumptions because it maintains (2.2) and specifies parametric models for $\mathcal{P}\{Y|X\}$ and $\mathcal{P}\{X\}$. It offers two advantages: (i) one can avoid the specification of the MDM and (ii) one can apply the same method to unit nonresponse since the information used is from respondent units only.

For INRY, TLR relaxes the above approach by estimating $f_X(x)$ in the first stage and maximising the resultant pseudolikelihood with respect to θ in the second stage. This is possible because X is observed in all sampling units.

One can, however, show that the above two approaches, in fact, do not completely ignore the MDM. From (2.2), the MDM is

$$\mathcal{P}\{R=1|Y=y\}=\frac{\mathcal{P}\{R=1,Y=y\}}{\int_{\mathcal{X}}f(y|x;\theta_0)f_X(x;\alpha_0)dx}.$$

The joint likelihood $\mathcal{P}\{Y=y,X=x,R=1\}$ can then be written as

$$\mathcal{P}\{Y=y,X=x,R=1\}=\mathcal{P}\{R=1,Y=y\}\frac{f(y|x;\theta_0)f_X(x;\alpha_0)}{\int_{\mathcal{X}}f(y|x;\theta_0)f_X(x;\alpha_0)dx}.$$

However, even if $\mathcal{P}\{R = 1, Y = y\}$ can be consistently estimated from the observed sample, $\mathcal{P}\{R = 1, Y = y\}$ is independent of θ. The above two approaches choose to ignore $\mathcal{P}\{R = 1, Y = y\}$ and specify only a part of the MDM which is dependent on θ, that is, $\int_{\mathcal{X}} f(y|x;\theta_0) f_X(x;\alpha_0)dx$. By doing so, it prevents them from using information on θ from nonrespondent units.

In contradistinction to the previous two approaches, RS estimates nonparametrically both $\mathcal{P}\{Y = y, R = 1\} = H_y$ and $f_X(x)$ and uses the full likelihood as the basis for inference on θ. In the case where Y and X are discrete, it is clear that, while imposing the same set of assumptions, the approach of RS must be semiparametrically more efficient than the other two approaches.

There is a clear distinction between parametric and nonparametric identifying assumptions in the first approach described. This distinction becomes less visible in RS and TLR since some related functions can be estimated from the observed sample without parametric specification. Despite these attempts to weaken the parametric assumption, the identification of both RS and TLR nevertheless rely heavily on the parametric identifying identification, i.e., the specification of $\mathcal{P}\{Y|X\}$.

An alternative way to gain an insight into the estimation procedures in RS and TLR is to consider the fully parametric score function based only on *respondent units*, which, given (2.2), is

$$0 = \sum_{n=1}^{N} r_n \left(\frac{\partial \log f(y_n|x_n;\widehat{\theta})}{\partial\theta} - \frac{1}{\int_{\mathcal{X}} f(y_n|x;\widehat{\theta}) f_X(x;\widehat{\alpha})dx} \int_{\mathcal{X}} \frac{\partial f(y_n|x;\widehat{\theta})}{\partial\theta} f_X(x;\widehat{\alpha})dx \right).$$

The form of this score function holds if the interchangeability between integration and differentiation is allowed. The score function of TLR in (2.45) is almost identical to this. The

only difference is that the integral $\int_{\mathcal{X}} f(y_n|x;\theta)f_X(x;\alpha)dx$ and its derivative are approximated by summations of the relevant quantities with respect to the empirical distribution of X.

Similarly, by (2.2), the fully parametric score function with respect to θ based on *the full sample* in (2.52) can be re-written as

$$0 = \sum_{n=1}^{N}\left\{ r_n \frac{\partial \log f(y_n|x_n;\widehat{\theta})}{\partial\theta} - \frac{1-r_n}{1-\int_{\mathcal{Y}} f(r=1|y;\widehat{\psi})f(y|x_n;\widehat{\theta})dy}\int_{\mathcal{Y}} f(r=1|y;\widehat{\psi})\frac{\partial f(y|x_n;\widehat{\theta})}{\partial\theta}dy\right\}.$$

We compare this expression to (2.15) which is the basis of the RS moment indicators for θ in an INRY setting. It is clear that (2.15) is the discrete version of the above score function where $f(r=1|y;\psi)$ is estimated by H_y/Q_y and where integration with respect to $\mathcal{Y}$ is approximated by summation across $y \in \mathcal{Y}$. In fact, one can estimate θ consistently using only the moment indicators based on (2.15) by replacing $\widehat{H}_y$ and $\widehat{Q}_y$ by $N^{-1}\sum_{n=1}^{N} r_n 1[y_n = y]$ and $N^{-1}\sum_{n=1}^{N}\mathcal{P}\{y|x_n;\theta\}$.

By considering RS and TLR from this perspective, it is clear that both depend markedly on the parametric identifying assumption. Effectively, RS and TLR modify the fully parametric score function using information from the observed sample. Nonparametric identification from (2.2) only affects the form of the estimable MDM. The modification of the fully parametric score function is possible because of parametric specification of the conditional distribution of Y given X, $f(y|x)$. Unlike $f(x)$, $f(y|x)$ cannot be estimated from the observed data due to the missing values in Y.

In accordance with this view, it is straightforward to see how our approach extends those of RS and TLR. That is, the INRYX-GMM estimator modifies the fully parametric score function with respect to θ in (2.52) using the relationship in (2.27). This results in the moment indicators for θ given in (2.29).

A similar idea has been used by Chatterjee, Chen and Breslow (2003) which proposes a pseudoscore (PS) estimator for two-phase sampling. The setting of this PS estimator is different to that presented here. The two main differences are (i) the MDM there depends on observed variables and (ii) a subset of covariates is missing. Nevertheless, the intuition underpinning the PS estimator and that of our GMM estimators is the same. Chatterjee, Chen and Breslow (2003) derive firstly a score function from the observed data likelihood. Then, they show that any unknown and unspecified part of the score function can be estimated from the observed sample. The resultant score function is referred to as pseudoscore because it has neither an associated log-likelihood nor log-pseudolikelihood.

2.8.3 Conditions for Identification in Practice and Related Issues

Section 2.8.2 shows that identification is possible because we parametrically specify the conditional density function for Y given X. In this section, we examine primitive conditions for identification in a specific setting. Then we show that a certain specification of this density function can lead to an identification problem if the available data is unit nonresponse (UNR). This problem does not concern the application of our GMM estimators because, as explained below, it does not occur if data is INR.

The identification conditions appropriate for the implementation of pseudolikelihood estimation are also investigated. We then elaborate conditions under which (2.2) can non-parametrically identify $\mathcal{P}\{Y = y|X = x\}$.

An Identified Example

In the GMM framework adopted here, Assumption 2.5.3(i) is the identification assumption. It is comprised of two conditions: uniqueness of $(\theta_0', \gamma_0')'$ and full column rank of $E[\partial\widetilde{g}(w, \theta_0, \gamma_0)/\partial(\theta', \gamma')]$. The uniqueness condition is needed to ensure that there is no local optimum satisfying $E[\widetilde{g}(w, \theta, \gamma)] = 0$. The full rank condition implies that the moment indicators in $\widetilde{g}(w, \theta, \gamma)$ are nonredundant and, as a consequence, the number of moment indicators is equal to the number of parameters. To identify θ, both conditions must hold.

Since $\widetilde{g}(w, \theta, \gamma)$ is nonlinear in parameters, it is difficult to state primitive conditions for a general case; see Newey and McFadden (1994, p.2127). Nevertheless it is possible to show identification in special cases. Below, identification occurs for the use of INRY-GMM estimation in a specific setting and we discuss primitive conditions which can guarantee identification.

Suppose a random sample $\{y_n, x_{1n}, x_{2n}, r_n\}_{n=1}^N$ is drawn from the population of interest. Consider first the moment indicators for first-step estimation. For INRY, $\mathcal{P}\{Y = y, R = 1\}$ must be estimated in the first step. Now $\mathcal{P}\{Y = y, R = 1\} = \mathcal{P}\{Y = y|R = 1\}\mathcal{P}\{R = 1\}$ and $\mathcal{P}\{R = 1\}$ can be consistently estimated by N_r/N, where N_r is the number of respondent units. Suppose Y given R is normally distributed and let $\phi(\cdot)$ de-

note the density function of the standard normal distribution. Thus, in this specific setting, $d(w,\gamma) = \frac{1}{\sigma_y}\phi(\frac{y-\mu_y}{\sigma_y}), \gamma = (\mu_y, \sigma_y)'$ and the first-step estimator is ML based on respondent units.

The ML objective function is $N^{-1}\sum_{n=1}^{N}\left[r_n \log \frac{1}{\sigma_y}\phi(\frac{y-\mu_y}{\sigma_y})\right]$ and the moment indicators based on the first-order conditions are

$$\begin{aligned} \mu_y &: r\frac{(y-\mu_y)}{\sigma_y^2}; \\ \sigma_y &: \frac{r}{\sigma_y}\left[\frac{(y-\mu_y)^2}{\sigma_y^2} - 1\right]. \end{aligned} \tag{2.55}$$

To show first-step identification, the expectation of the ML objective function can be written as

$$-\frac{N_r}{2}\left[\log(2\pi\sigma_y^2) + \frac{\sigma_{y0}^2 + (\mu_y - \mu_{y0})^2}{\sigma_y^2}\right]. \tag{2.56}$$

Thus, (μ_{y0}, σ_{y0}) uniquely maximises (2.56) and this implies that (μ_{y0}, σ_{y0}) is also the unique solution to equating the expectation of the moment indicators in (2.55) to zero.

Consider the moment indicators for second-step estimation. Let $f(y|x_1, x_2; \theta) = \frac{1}{\sigma}\phi(\frac{y-x'\beta}{\sigma})$ where $x'\beta = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Let $h(y)$ denote $\frac{1}{\sigma_y}\phi(\frac{y-\mu_y}{\sigma_y})\,(N_r/N)$, the estimator of the joint density of Y and $R = 1$ from the first step. Let $Q(y)$ denote $N^{-1}\sum_{n=1}^{N}\frac{1}{\sigma}\phi(\frac{y-x_n'\beta}{\sigma})$ and let φ denote $(\beta_0, \beta_1, \beta_2, \sigma, \mu_y, \sigma_y)'$.

To write the moment indicators more compactly, define

$$k_1(y,x,r) = r\left(\frac{y-x'\beta}{\sigma^2}\right) - \frac{(1-r)}{1-\int_{\mathcal{Y}}\frac{h(y)}{Q(y)}\left(\frac{1}{\sigma}\phi(\frac{y-x'\beta}{\sigma})\right)dy}\left[\int_{\mathcal{Y}}\frac{h(y)}{Q(y)}\left(\frac{y-x'\beta}{\sigma^3}\right)\phi(\frac{y-x'\beta}{\sigma})dy\right],$$

and

$$k_2(y,x,r) = r\frac{1}{\sigma}\left(\left(\frac{y-x'\beta}{\sigma}\right)^2-1\right) - \frac{(1-r)}{1-\int_{\mathcal{Y}}\frac{h(y)}{Q(y)}\left(\frac{1}{\sigma}\phi(\frac{y-x'\beta}{\sigma})\right)dy}\left[\int_{\mathcal{Y}}\frac{h(y)}{Q(y)}\left(\left(\frac{y-x'\beta}{\sigma}\right)^2-1\right)\frac{1}{\sigma^2}\phi(\frac{y-x'\beta}{\sigma})dy\right].$$

Thus, the moment indicators for $\beta_0, \beta_1, \beta_2$ and σ are

$$\beta_0 : k_1(y,x,r), \tag{2.57}$$

$$\beta_1 : k_1(y,x,r)\cdot x_1, \tag{2.58}$$

$$\beta_2 : k_1(y,x,r)\cdot x_2, \tag{2.59}$$

$$\sigma : k_2(y,x,r). \tag{2.60}$$

The form of these moment indicators indicates that if $E[k_1(y,x,r)|x] = 0$ and $E[k_2(y,x,r)|x] = 0$ then the unconditional expectation of these second-step moment indicators with respect to the sample distribution equals zero.

Now, $E[k_1(y,x,r)|x]$ and $E[k_2(y,x,r)|x]$ can be written as

$$E[k_1(y,x,r)|x] = \int_{\mathcal{Y}}\frac{h_0(y)}{Q_0(y)}\left(\frac{1}{\sigma_0}\phi(\frac{y-x'\beta_0}{\sigma_0})\right)\left(\frac{y-x'\beta}{\sigma^2}\right)dy - \int_{\mathcal{Y}}\frac{h(y)}{Q(y)}\left(\frac{1}{\sigma}\phi(\frac{y-x'\beta}{\sigma})\right)\left(\frac{y-x'\beta}{\sigma^2}\right)dy,$$

$$E[k_2(y,x,r)|x] = \int_{y}\frac{h_0(y)}{Q_0(y)}\left(\frac{1}{\sigma_0}\phi(\frac{y-x'\beta_0}{\sigma_0})\right)\left[\frac{1}{\sigma}\left(\left(\frac{y-x'\beta}{\sigma}\right)^2-1\right)\right]dy - \left[\int_{\mathcal{Y}}\frac{h(y)}{Q(y)}\left(\frac{1}{\sigma}\phi(\frac{y-x'\beta}{\sigma})\right)\left[\frac{1}{\sigma}\left(\left(\frac{y-x'\beta}{\sigma}\right)^2-1\right)\right]dy\right].$$

It is clear that $E[k_1(y,x,r)|x]$ and $E[k_2(y,x,r)|x]$ are zero if $\varphi = \varphi_0$. For φ_0 to be the unique solution, we need the following condition

$$\varphi \neq \varphi_0 \rightarrow \frac{h(y)}{Q(y)}\frac{1}{\sigma}\phi(\frac{y-x'\beta}{\sigma})dy \neq \frac{h_0(y)}{Q_0(y)}\frac{1}{\sigma_0}\phi(\frac{y-x'\beta_0}{\sigma_0}). \tag{2.61}$$

Notice that $\frac{1}{\sigma}\phi(\frac{y-x'\beta}{\sigma})$ is the conditional density of Y given X, $f(y|x;\theta)$, and $h(y)/Q(y)$ is the MDM or P_y. In a standard application of conditional ML, it must be assumed $\theta \neq \theta_0 \rightarrow f(y|x;\theta) \neq f(y|x;\theta_0)$ to ensure the identification of θ_0. Condition (2.61) is then seen as the counterpart of such assumption for the missing data problem. We can interpret $h(y)/Q(y)$ as a kind of distortion due to missing data. Thus, the requirement (2.61) means that, after adjusting for this distortion, there is no two values of (β, σ) which yield the same probability.

Alternatively, since we include the constant term β_0 in this model, we can demean all covariates such that they have zero mean. Then, unconditional expectations of the moment indicators in (2.58) and (2.59) are, in fact, $Cov[E[k_1(y,x,r)|x], x_1]$ and $Cov[E[k_1(y,x,r)|x], x_2]$, respectively. If we can prove that $E[k_1(y,x,r)|x]$, which is a function of (x_1, x_2), is monotonic in (x_1, x_2) then these covariances are nonzero for $\varphi \neq \varphi_0$.[5]

Under either condition (2.61) or monotonicity in (x_1, x_2) of $E[k_1(y,x,r)|x]$, φ_0 is the unique solution to equating the expectation of the second step moment indicators to zero. By combining this result with that shown in (2.56), the uniqueness condition in Assumption 2.5.3(i) is satisfied. Moreover, none of moment indicators in (2.55), (2.57), (2.58), (2.59)

---

[5] Newey and McFadden (1994, p.2128) note, in a footnote, that $Cov[x, f(x)]$ is nonzero for any monotonic and nonconstant function $f(x)$ of a random variable x.

and (2.60) can be expressed as a linear combination of others. Thus, the full rank condition is satisfied and, as a consequence, Assumption 2.5.3(i) holds in this particular example.

An Unidentified Example

The objective of this example is to show that there exists a family of parametric functions for the conditional distribution of Y given X that leads to a lack of identification in a specific circumstance. Then, we explain why the GMM estimators proposed above should avoid this type of identification problem.

There are three main factors contributing to a lack of identification: data is unit non-response (UNR), the MDM is a particular case of (2.2) where Y is discrete and, lastly, the parametric specification for the conditional distribution of Y given X belongs to the family of *multiplicative intercept* models (MIM); see Hsieh, Manski and McFadden (1985) or RS. The MIM family is defined formally as

$$\mathcal{P}\{y|x, v_y, \theta_y^1\} = \frac{\nu_y V_y(\theta_y^1)}{\sum_{y\in\mathcal{Y}} \nu_y V_y(\theta_y^1)}, \tag{2.62}$$

where $\nu_y = \nu_y(\theta_y^0)$, $\nu_0(\theta_0^0) = 1$, $V_0(\theta_0^1) = 1$, $V_y(\theta_y^1) > 0$, $\partial\nu_y(\theta_y^0)/\partial\theta_y^0 = \nu_y(\theta_y^0)$ and $\partial V_y(\theta_y^1)/\partial\theta_y^1 = x_y V_y(\theta_y^1)$ for all y. The multinomial logit model is a member of this family with $\nu_y(\theta_y^0) = \exp(\theta_y^0)$, $V_y(\theta_y^1) = \exp(x'\theta_y^1)$, $\theta_0^0 = 0$, and $\theta_0^1 = 0$.

RS have addressed MIM when data is UNR with unknown total sample size N. The RS-GMM estimator is the same as that proposed in Imbens(1992) for choice-based samples indicating that the identification problem is also well understood in the literature of choice-based sampling.

For simplicity, set $\mathcal{Y} \in \{0,1\}$; see Imbens (1992). Let $\mathcal{P}\{Y = 1|X = x\} = \mathcal{P}\{x'\beta\} = \mathcal{P}\{\beta_0 + \beta_1 x_1\}$ and $\mathcal{P}_\beta\{x'\beta\}$ denote the first order derivative. The moment indicators in this case are

$$\begin{aligned}
H_1 &: 1[y=1] - H_1; \\
Q_1 &: Q_1 - \mathcal{P}\{x'\beta\}\left[\frac{H_1}{Q_1}\mathcal{P}\{x'\beta\} + \frac{1-H_1}{1-Q_1}(1-\mathcal{P}\{x'\beta\})\right]^{-1}; \\
\beta &: \left[\frac{\mathcal{P}_\beta\{x'\beta\}}{\mathcal{P}\{x'\beta\}}\right]\cdot 1[y=1] - \left[\frac{\mathcal{P}_\beta\{x'\beta\}}{1-\mathcal{P}\{x'\beta\}}\right]\cdot 1[y=0] \\
&\quad -\mathcal{P}_\beta\{x'\beta\}\left[\frac{H_1}{Q_1} - \frac{1-H_1}{1-Q_1}\right]\left[\frac{H_1}{Q_1}\mathcal{P}\{x'\beta\} + \frac{1-H_1}{1-Q_1}(1-\mathcal{P}\{x'\beta\})\right]^{-1},
\end{aligned}$$

where now $H_y = \mathcal{P}\{Y = y|R = 1\}$ and, hence, $1 - H_1 = H_0$. The logit model, in which $\mathcal{P}\{x'\beta\} = \exp(x'\beta)(1+\exp(x'\beta))^{-1}$, is a special case of the MIM family with $\nu_0(\theta_0^0) = 1$, $V_0(\theta_0^1) = 1, \nu_1(\theta_0^1) = \exp(\beta_0)$ and $V_1(\theta_1^1) = \exp(\beta_1 x_1)$. Thus, the moment indicators for H_1, Q_1, β_0 and β_1 under the logit specification are

$$H_1 : 1[y=1] - H_1; \tag{2.63}$$

$$Q_1 : Q_1 - \exp(x'\beta)\left[\frac{H_1}{Q_1}\exp(x'\beta) + \frac{1-H_1}{1-Q_1}\right]^{-1}; \tag{2.64}$$

$$\beta_0 : 1[y=1] - \frac{H_1}{Q_1}\exp(x'\beta)\left[\frac{H_1}{Q_1}\exp(x'\beta) + \frac{1-H_1}{1-Q_1}\right]^{-1}; \tag{2.65}$$

$$\beta_1 : x_1 \cdot \{1[y=1] - \frac{H_1}{Q_1}\exp(x'\beta)\left[\frac{H_1}{Q_1}\exp(x'\beta) + \frac{1-H_1}{1-Q_1}\right]^{-1}\}; \tag{2.66}$$

The moment indicator for β_0 is a linear combination of (2.63) and (2.64), i.e.,

$$(2.65) = (2.63) + \frac{H_1}{Q_1}(2.64).$$

However, (2.66), which is equivalent to x_1 times (2.65), is linear independent of (2.63) and (2.64) due to the presence of x_1. Thus, β is unidentified here since the full rank condition

does not hold, i.e., the number of independent moment indicators is less than the number of parameters to be estimated.

Although we cannot identify all parameters of interest, other parameters apart from β_0 are identified due to the properties of the MIM family. Let C_0 denote $\exp(\beta_0)$. Suppose that we normalise the intercept term of choice $Y = 0$ to be C_0, instead of a unit. Then the above moment indicators ignoring (2.65) become

$$
\begin{aligned}
H_1 &: 1[y=1] - H_1; \\
Q_1 &: Q_1 - C_0 \exp(\beta_1 x_1) \left[\frac{H_1}{Q_1} C_0 \exp(\beta_1 x_1) + \frac{1-H_1}{1-Q_1} C_0 \right]^{-1}; \\
\beta_1 &: \left(1[y=1] - \frac{H_1}{Q_1} C_0 \exp(\beta_1 x_1) \left[\frac{H_1}{Q_1} C_0 \exp(\beta_1 x_1) + \frac{1-H_1}{1-Q_1} C_0 \right]^{-1} \right) x_1.
\end{aligned}
$$

It is clear that C_0 may be cancelled from all moment indicators. Thus, H_1,Q_1 and β_1 can be consistently estimated even though β_0 is unidentified. In the general case when number of choices exceeds one, this result implies that only the intercept term for each choice is unidentified.

It has been noted that identification in RS, TLR and our approach requires the specification of a parametric function for $\mathcal{P}\{Y = y|X = x\}$. Thus, the above example demonstrates a potential weakness in these approaches, suggesting there may be some parametric models for $\mathcal{P}\{Y = y|X = x\}$ which can lead to a lack of identification.

In the choice-based sampling literature, Cosslett (1981) suggests two set of conditions to deal with this problem. One restricts the sample design whereas the other restricts the class of parametric models for $\mathcal{P}\{Y = y|X = x\}$. Nonetheless, in the context of miss-

ing data, we cannot change the MDM and have to take the sample design as given. Thus, restricting the parametric families for $\mathcal{P}\{Y = y|X = x\}$ is the only available option for the missing data problem. E.g., one can adopt an assumption which prohibits the conditional parametric model for $\mathcal{P}\{Y = y|X = x\}$ from taking the MIM form (2.62).

Nevertheless, this example assumes that the available data is UNR with unknown total sample size N. If the data is INR, then the moment indicators derived in RS are no longer of the form in (2.63), (2.64), (2.65) and (2.66). Furthermore, the RS GMM estimator for INR data no longer coincides with that for choice-based sampling. Thus, if data is INR, it can be shown that all parameters are identified even if the conditional parametric model is of the form in (2.62); see RS, p. 23. This stresses the importance of being able to collect data on X from nonrespondent units. Since our approach is also based on INR data, our GMM estimators should not suffer the identification problem exemplified by this example.

Identification in TLR

It is noted above that the identification problem may be avoided by excluding the MIM parametric family as the specification for $\mathcal{P}\{Y = y|X = x\}$. TLR, however, adopts the opposite approach by restricting attention to a particular parametric family in which identification is guaranteed. Below, we present results which form the basis for the identification of θ_0 in TLR; *viz.*

Lemma 2.8.3 *(Identifiability): Suppose that the joint distribution of* (Y, X) *admits the density* $f(y|x;\theta_0)f_X(x;\alpha_0)$. *For any given* θ *and an arbitrary function* $c(y)$, *let*

$$D_\theta = \{y : f(y|x;\theta) = c(y)f(y|x;\theta_0) \textit{ for any } x\}.$$

If $\mathcal{P}\{R=1\} > 0$ *and* $\mathcal{P}\{D_\theta\} < 1$ *for any* $\theta \neq \theta_0$*, then*

$$E[r \cdot \log f(x|y;\theta,\alpha_0)] < E[r \cdot \log f(x|y;\theta_0,\alpha_0)], \theta \neq \theta_0,$$

where R *is the binary indicator function for complete cases and* $\mathbb{E}[\cdot]$ *denotes expectation taken with respect to the sample distribution.*

Proposition 2.8.3. *Suppose* α_0 *is known and a random sample is available from* $\mathcal{P}\{X|Y\}$*. Also, suppose* θ *can be reparameterised as* $\widetilde{\theta} = (\widetilde{\theta}_1, \widetilde{\theta}_2)$ *and the conditional density function may be written as*

$$f(y|x,\theta) = \exp\{f_1(y,x;\widetilde{\theta}_1) + f_2(y;\widetilde{\theta})\},$$

such that, for any $b \neq \widetilde{\theta}_{10}$ *in the domain of* $\widetilde{\theta}_1$*, the function* $f_1(y,x;b) - f_1(y,x;\widetilde{\theta}_{01})$ *is not a function of* y *alone. Then* θ_1 *is identifiable whereas* θ_2 *is not.*

Lemma 2.8.3 and Proposition 2.8.3 appear somewhat involved but may be motivated by the following simple example. Suppose that the joint distribution of Y and X is bivariate normal. Further suppose that the mean functions of the conditional distributions of Y given X and X given Y are linear. By properties of the bivariate normal distribution, both conditional distributions are also normal and the following conditions hold

$$\mu_x = \beta_{0,x|y} + \beta_{1,x|y}\mu_y, \ \sigma^2_{yx} = \beta_{1,x|y}\sigma^2_y, \ \sigma^2_x = \sigma^2_{x|y} + \beta^2_{1,x|y}\sigma^2_y,$$

$$\mu_y = \beta_{0,y|x} + \beta_{1,y|x}\mu_x, \ \sigma^2_{yx} = \beta_{1,y|x}\sigma^2_x, \ \sigma^2_y = \sigma^2_{y|x} + \beta^2_{1,y|x}\sigma^2_x,$$

where β_0 and β_1 denote conditional mean function parameters and μ and σ^2 mean and variance of the associated marginal distributions. If data is INR and the MDM is (2.2), $\beta_{0,x|y}$, $\beta_{1,x|y}$, $\sigma^2_{x|y}$ are identified from respondent units and μ_x, σ^2_x are identified from all sampling

units. Thus, using the above conditions to write the unknown parameters as functions of the identified parameters yields

$$\beta_{0,y|x} = \frac{\mu_x(\sigma^2_{x|y}/\sigma^2_x) - \beta_{0,x|y}}{\beta_{1,x|y}}, \beta_{1,y|x} = \frac{1 - \sigma^2_{x|y}/\sigma^2_x}{\beta_{1,x|y}}, \sigma^2_{y|x} = \frac{(\sigma^2_{x|y}/\sigma^2_x)(\sigma^2_x - \sigma^2_{x|y})}{\beta^2_{1,x|y}}. \tag{2.67}$$

Therefore, all parameters defining the conditional distribution of Y given X are identified.

Essentially, Lemma 2.8.3 and Proposition 2.8.3 in TLR generalise the above example. It is immediate that the example satisfies their conditions, i.e., since Y and X are dependent, $\beta_{1,y|x} \neq 0$ implying that $\mathcal{P}\{D_\theta\} < 1$ in Lemma 2.8.3. Moreover, $f(y|x;\theta)$ where $\theta = (\beta_{0,y|x}, \beta_{1,y|x}, \sigma^2_{y|x})$ is a normal density from the properties of bivariate normal distribution and

$$f(y|x;\theta) \propto \exp\left\{-\frac{(y - \beta_{0,y|x} - \beta_{1,y|x}x)^2}{2\sigma^2_{y|x}}\right\} = \exp\{-\frac{\beta_{0,y|x}\beta_{1,y|x}}{\sigma^2_{y|x}}x + \frac{\beta_{1,y|x}}{\sigma^2_{y|x}}yx - \frac{\beta^2_{1,y|x}}{2\sigma^2_{y|x}}x^2 + f_2(y;\theta)\}.$$

Thus, $\widetilde{\theta}_1$ of Proposition 2.8.3 is $(\beta_{0,y|x}\beta_{1,y|x}/\sigma^2_{y|x}, \beta_{1,y|x}/\sigma^2_{y|x}, \beta^2_{1,y|x}/2\sigma^2_{y|x})$. Hence, all parameters of interest are identified.

TLR claim that the normality of X in this example can be relaxed. All parameters can be identified as long as X is continuous or discrete with point mass at more than three points. As a consequence, the estimators in TLR require this further condition on X to be satisfied in additional to those in Lemma 2.8.3 and Proposition 2.8.3 even though it is not assumed by TLR in their proofs.

If X is Bernoulli with probability π, the parameters θ are not identified as noted in TLR. To explain this, consider another example. Suppose $f(y|x;\theta)$ is normal with the same

set of parameters as in the first example. Since X is Bernoulli, TLR asserts that $\mathcal{P}\{X|Y\}$ is now a logit model with a mean function $\gamma_0 + \gamma_1 Y$. Accordingly, the joint distribution $\mathcal{P}\{Y|X\}\mathcal{P}\{X\}$ is parameterised by $\beta_{0,y|x}, \beta_{1,y|x}, \sigma^2_{y|x}$ and π but only three parameters are identifiable from the distributions $\mathcal{P}\{X|Y\}$ and $\mathcal{P}\{X\}$, i.e., γ_0, γ_1 and π. Thus, θ is not identified.

In our opinion, there are, at least, two points which remain unclear in the second example. First, the parametric model for $\mathcal{P}\{X|Y\}$ is normal in the first example by the properties of the bivariate normal distribution. However, it is somewhat *ad hoc* to maintain that the model for $\mathcal{P}\{X|Y\}$ in the second example is logit since other binary choice models such as probit may be used. Secondly, TLR do not show explicit relationships between the two sets of parameters, such as those in (2.67), for the second example. It is unclear whether or not such relationships even exist. Furthermore, if one estimates $\mathcal{P}\{Y, R = 1\}$ from the observed sample as in RS and our approaches, it is not transparent in this framework what effect this information has on identification.

Conditions for Nonparametric Point-identification when MDM is (2.2)

Section 2.8.2 shows the power of non-parametric identification from the MDM in (2.2) is relatively limited in RS and TLR. However, by combining it with additional assumptions, Manski (1994) shows that the conditional distribution $\mathcal{P}\{Y|X\}$ can be non-parametrically identified. Although this result does not explain identification in TLR and RS, we include it here for completeness.

From (2.2) and (2.53), $\mathcal{P}\{X|Y=y\} = \mathcal{P}\{X|Y=y, R=1\}$ which implies that

$$\frac{\mathcal{P}\{Y=y|X=x\}}{\mathcal{P}\{Y=y\}} = \frac{\mathcal{P}\{Y=y, R=1|X=x\}}{\mathcal{P}\{Y=y, R=1\}}. \tag{2.68}$$

If data is INR, then $\mathcal{P}\{Y=y, R=1|X=x\}$ is identified because X is fully observed. Thus, the right hand side of (2.68) is non-parametrically identified. This also means that, under (2.2), the ratio $\mathcal{P}\{Y=y|X=x\}/\mathcal{P}\{Y=y\}$ is also non-parametrically point-identified. We have noted in (2.54) that assuming (2.2) cannot point-identify $\mathcal{P}\{Y=y|X=x\}$ because the sampling process is not informative about $\mathcal{P}\{Y=y\}$. We show below that, under some additional assumptions, one can non-parametrically identify $\mathcal{P}\{Y=y\}$ using (2.68).

Corollary 2.8.3: *Suppose the MDM is (2.2). Let Y be multinomial with sample space $\mathcal{Y} = \{y^1, ..., y^I\}$. Let the sample space of X contain $J+1$ points of support, i.e., $\mathcal{X} = \{x^0, ..., x^J\}$. Let $\mathcal{P}\{Y=y, R=1\} > 0$ for all $y \in \mathcal{Y}$. Let A be the $(J+1) \times I$ matrix whose ij*th *element is*

$$a_{ij} = \frac{\mathcal{P}\{Y=y_j, R=1|X=x_i\}}{\mathcal{P}\{Y=y_j, R=1\}}.$$

Then $\mathcal{P}\{Y=y\}$ is point-identified if A has rank I.

Corollary 2.8.3 implies that $\mathcal{P}\{Y=y\}$, $y \in \mathcal{Y}$, is point-identified if (i) the MDM is (2.2), (ii) Y is discrete and (iii) X has at least as many support points as does Y. As a result, $\mathcal{P}\{Y=y|X=x\}$ is also point-identified from either (2.54) or (2.68).

Even though RS is interested mainly on cases where Y is discrete, the result from Corollary 2.8.3 cannot explain the identification in RS. This is because the GMM estimator in RS is applicable even when data is UNR but Corollary 2.8.3 does not hold in such case.

Corollary 2.8.3 is invalid for UNR data since $\mathcal{P}\{Y = y, R = 1|X = x\}$ in (2.68) is not non-parametric identified whenever X is jointly missing with Y.

Nevertheless, Corollary 2.8.3 indicates some interesting conclusions. It suggests that (2.2) may be powerful in identifying $\mathcal{P}\{Y = y|X = x\}$. This point is not obvious in the RS and TLR approaches. It also emphasises that $\mathcal{P}\{Y = y, R = 1\}$, $y \in \mathcal{Y}$, must be positive, an important point also stressed in section 2.5 by Assumption 2.5.1(iv), since it differentiates the sampling process underpinning RS and our approach from that of the Tobit model. Moreover, Corollary 2.8.3 shows formally that the support points of X may by important for the identification of $\mathcal{P}\{Y = y|X = x\}$. A similar point is claimed in TLR, but with no formal proof provided there.

2.9 Summary

In this chapter, the RS estimation procedure is extended such that the MDM can now be an arbitrary function of either a continuous response variable or both the continuous response variable and covariates. The cost of the relaxation is that some nuisance functions must be estimated prior to the GMM estimation. Thus, the resultant estimation procedures can be referred to as two-step GMM estimations. Three types of two-step GMM estimators are proposed, namely, INRYX-GMM, INRY-GMM and INRYX1-GMM estimators. These GMM estimators are associated with the MDMs in (2.1), (2.2) and (2.3) respectively.

The parametric estimation is used in the first step of these two-step GMM estimators to simplify the discussion of their asymptotic properties. This implies, for example, that the INRYX-GMM estimator requires the correct parametric specifications of

$f(y|x), h(y, x, r = 1)$ and $f_X(x)$. It is therefore restrictive and may not be more attractive than the fully parametric approach. However, it is theoretically possible to estimate both $h(y, x, r = 1)$ and $f_X(x)$ by nonparametric methods based only on sample information. Provided these first-step nonparametric estimations, the INRYX-GMM estimator requires only the correct specification of the population conditional density function $f(y|x; \theta)$ of Y given X for consistent estimation of θ_0. This possibility represents a considerable advantage of it over the fully parametric approach. A shortcoming of the INRYX-GMM estimator, however, is that estimation of $h(y, x, r = 1)$ and $f_X(x)$ becomes more complex as the dimension of X increases.

If one is willing to assume that the MDM is (2.2), the above shortcoming can be overcome by applying the INRY-GMM estimator instead of the INRYX-GMM estimator. An increase in the dimension of X does not complicate its first-step estimation because the MDM in this case does not depend on X. Under (2.2), it is also possible to apply the one-step version of the INRY-GMM estimator which allows $h(y, r = 1)$ to be unspecified. Moreover, the INRYX1-GMM estimators can be considered as a compromise between the INRYX-GMM estimator and the INRY-GMM estimator. It is more general than the INRY-GMM estimator since it permits the MDM to depend on the continuous response variable and a subset of covariates X_1. Its first-step estimation is also less complicated than that of the INRYX-GMM estimator since $\dim(x_1) < \dim(x)$.

The RS approach can cover the case where MDM depends on finite partitions of the continuous response variable and covariates, that is, the MDM is (2.39). The INRYX-GMM estimator relaxes this estimator by allowing the MDM to be (2.1) rather than (2.39).

The cost of such relaxation is the specification and estimation of $f_X(x)$. This cost may, however, be acceptable if one can use a nonparametric method to estimate $f_X(x)$ from the observed sample.

It is also shown that if the MDM is (2.2), the RS GMM estimator should be more efficient than the PL estimator whenever Y and X are discrete because the former extracts more information from the likelihood of the nonrespondent units. This efficiency argument should remain true if only X is allowed to be continuous. However, the RS GMM estimator is no longer applicable whenever both Y and X are continuously distributed. In such a case, the INRY-GMM estimator should be compared to the PL estimator and we suspect that the former is semiparametrically more efficient than the latter due to the same reasoning. Even though there is no mathematical proof of this conjecture at present, the Monte Carlo evidence supporting it is presented in Chapter 3.

Nonparametric and parametric identifying assumptions are then discussed. Although identification in RS, TLR and our approach is a combination of the two, they rely more on the parametric identifying assumption. In fact, all three approaches can be considered as modifications of the associated fully parametric score functions using information from the observed sample. Furthermore, identification in these approaches is possible because the conditional density function for Y given X is parametrically specified. Unlike $f(x)$, $f(y|x)$ cannot be estimated from the observed data due to the missing values in Y. In the context of RS GMM estimation, certain specification of $f(y|x)$ can lead to an identification problem if the available data is unit nonresponse (UNR). However, this problem does not concern the application of our GMM estimators because it does not occur if data is INR.

2.A Appendix A: Derivation of Moment Indicators in Section 2.6.1

The objective function is

$$\begin{aligned}\mathcal{L} &= \sum_{n=1}^{N}\{r_n[\log H_{i_n j_n} + \log f(y_n|x_n;\theta) + \log \pi_{x_n} - \log Q_{i_n j_n}] \\ &\quad + (1-r_n)\left[\log(1 - \sum_{i\in\mathcal{I}} \frac{H_{ij_n}}{Q_{ij_n}} R(i|x_n,\theta)) + \log \pi_{x_n}\right]\} \\ &\quad -\mu(\sum_{x\in\mathcal{X}} \pi_x - 1),\end{aligned}$$

where $Q_{ij} = \sum_{x\in\mathcal{X}_j} R(i|x,\theta)\pi_x$. The first-order conditions with respect to θ, π_x and H_{ij} are

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \theta} &= \sum_{n=1}^{N}\left\{ r_n \left(\frac{\partial \log f(y_n|x_n;\theta)}{\partial \theta} - \frac{1}{Q_{i_n j_n}} \sum_{x\in\mathcal{X}_{j_n}} \frac{\partial R(i_n|x,\theta)}{\partial \theta}\pi_x \right) \right. \\ &\quad \left. - \frac{(1-r_n)}{1 - \sum_{i\in\mathcal{I}} \frac{H_{ij_n}}{Q_{ij_n}} R(i|x_n,\theta)} \sum_{i\in\mathcal{I}} \frac{H_{ij_n}}{Q_{ij_n}} \left(\frac{\partial R(i|x_n,\theta)}{\partial \theta} - \frac{1}{Q_{ij_n}} R(i|x_n,\theta) \sum_{x\in\mathcal{X}_{j_n}} \frac{\partial R(i|x,\theta)}{\partial \theta}\pi_x \right) \right\},\end{aligned} \tag{2.69}$$

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \pi_x} &= \sum_{n=1}^{N}\left\{ r_n \left(\frac{1[x_n = x]}{\pi_x} - 1[j_n = j] \frac{R(i_n|x,\theta)}{Q_{i_n j}} \right) \right. \\ &\quad \left. + (1-r_n) \left(\frac{1[j_n = j] \sum_{i\in\mathcal{I}} \frac{H_{ij}}{Q_{ij}^2} R(i|x_n,\theta) R(i,x,\theta)}{1 - \sum_{i\in\mathcal{I}} \frac{H_{ij}}{Q_{ij}} R(i|x_n,\theta)} + \frac{1[x_n = x]}{\pi_x} \right) \right\} - \mu, x \in \mathcal{X}_j,\end{aligned} \tag{2.70}$$

$$\frac{\partial \mathcal{L}}{\partial H_{ij}} = \sum_{n=1}^{N}\left(r_n \frac{1[i_n = i]\cdot 1[j_n = j]}{H_{ij}} - (1-r_n) \frac{1[j_n = j]}{1 - \sum_{i\in\mathcal{I}} \frac{H_{ij}}{Q_{ij}} R(i|x_n,\theta)} \frac{R(i|x_n,\theta)}{Q_{ij}} \right). \tag{2.71}$$

From (2.71),

$$\widehat{H}_{ij} = \widehat{Q}_{ij} \sum_{n=1}^{N} r_n 1[i_n = i] \cdot 1[j_n = j] / \sum_{n=1}^{N} \frac{(1 - r_n)\, 1[j_n = j] R(i|x_n, \widehat{\theta})}{1 - \sum_{i \in \mathcal{I}} \frac{\widehat{H}_{ij}}{\widehat{Q}_{ij}} R(i|x_n, \widehat{\theta})}. \tag{2.72}$$

Equating (2.71) to zero implies that the second and third terms in (2.70) cancel. Thus, (2.70) becomes

$$\begin{aligned} 0 &= \sum_{n=1}^{N} 1[x_n = x] \left(\frac{r_n}{\widehat{\pi}_x} + \frac{1 - r_n}{\widehat{\pi}_x} \right) - \widehat{\mu} \\ &= \sum_{n=1}^{N} 1[x_n = x] \frac{1}{\widehat{\pi}_x} - \widehat{\mu}. \end{aligned} \tag{2.73}$$

Multiplying through by $\widehat{\pi}_x$ and summing over $x \in \mathcal{X}$ gives

$$\widehat{\mu} = N.$$

Substituting $\widehat{\mu} = N$ into (2.73), we obtain the ML estimator for $\widehat{\pi}_x$

$$\widehat{\pi}_x = \sum_{n=1}^{N} 1[x_n = x]/N. \tag{2.74}$$

As a result, the population stratum occupancy probability estimator becomes

$$\begin{aligned} \widehat{Q}_{ij} &= \sum_{x \in \mathcal{X}_j} R(i|x, \widehat{\theta}) \widehat{\pi}_x \\ &= \sum_{x \in \mathcal{X}_j} R(i|x, \widehat{\theta}) \sum_{n=1}^{N} 1[x_n = x]/N \\ &= \sum_{n=1}^{N} 1\left[x_n \in \mathcal{X}_j\right] R(i|x_n, \widehat{\theta})/N. \end{aligned}$$

Likewise, (2.71) implies that the second and fourth terms of (2.69) cancel. Thus, first order conditions for $\widehat{\theta}$ become

$$\frac{\partial \mathcal{L}}{\partial \theta} = \sum_{n=1}^{N} \left(r_n \frac{\partial \log f(y_n|x_n; \widehat{\theta})}{\partial \theta} - \frac{(1 - r_n)}{1 - \sum_{i \in \mathcal{I}} \frac{\widehat{H}_{ijn}}{\widehat{Q}_{ijn}} R(i|x_n, \widehat{\theta})} \sum_{i \in \mathcal{I}} \frac{\widehat{H}_{ijn}}{\widehat{Q}_{ijn}} \frac{\partial R(i|x_n, \widehat{\theta})}{\partial \theta} \right) = 0.$$

The moment conditions for GMM estimation are then given by

$$0 = \sum_{n=1}^{N}\left(r_n\frac{\partial \log f(y_n|x_n;\widehat{\theta})}{\partial\theta} - \frac{(1-r_n)}{1-\sum_{i\in\mathcal{I}}\frac{\widehat{H}_{ijn}}{\widehat{Q}_{ijn}}R(i|x_n,\widehat{\theta})}\sum_{i\in\mathcal{I}}\frac{\widehat{H}_{ijn}}{\widehat{Q}_{ijn}}\frac{\partial R(i|x_n,\widehat{\theta})}{\partial\theta}\right),$$

$$0 = \sum_{n=1}^{N}\left(r_n\frac{1[i_n=i]\cdot 1[j_n=j]}{H_{ij}} - (1-r_n)\frac{1[j_n=j]}{1-\sum_{i\in\mathcal{I}}\frac{\widehat{H}_{ij}}{\widehat{Q}_{ij}}R(i|x_n,\widehat{\theta})}\frac{R(i|x_n,\widehat{\theta})}{\widehat{Q}_{ij}}\right),$$

$$0 = N\widehat{Q}_{ij} - \sum_{n=1}^{N} 1\left[x_n\in\mathcal{X}_j\right]R(i|x_n,\widehat{\theta}).$$

2.B Appendix B: Proofs

Proof of Lemma 2.8.3 TLR shows that

$$\begin{aligned} E[r\cdot\log f(x|y,\theta,\alpha_0)] &= E_{y,x}[E[r\cdot\log f(x|y,\theta,\alpha_0)|y,x]] \\ &= E_{y,x}[E[f(r=1|y;\psi_0)\log f(x|y,\theta,\alpha_0)|y,x]] \\ &= -E_y[f(r=1|y;\psi_0)E[-\log f(x|y,\theta,\alpha_0)|y]], \end{aligned}$$

where $f(r=1|y;\psi_0)$ is the true MDM under (2.2). For any fixed y, Jensen's inequality implies

$$E[\log\frac{f(x|y,\theta,\alpha_0)}{f(x|y,\theta_0,\alpha_0)}|y] \leq \log E[\frac{p(x|y,\theta,\alpha_0)}{p(x|y,\theta_0,\alpha_0)}|y].$$

The RHS is equal to zero because $\log(1)=0$. Thus,

$$E[\log f(x|y,\theta_0,\alpha_0)|y] \geq E[\log f(x|y,\theta,\alpha_0)|y].$$

Equality holds if and only if $f(x|y,\theta_0,\alpha_0) = f(x|y,\theta,\alpha_0)$ or

$$\frac{f(y|x;\theta_0)f_X(x;\alpha_0)}{\int_{\mathcal{X}} f(y|x;\theta_0)f_X(x;\alpha_0)dx} = \frac{f(y|x;\theta)f_X(x;\alpha_0)}{\int_{\mathcal{X}} f(y|x;\theta)f_X(x;\alpha_0)dx}.$$

Re-arranging

$$\frac{f(y|x;\theta_0)}{f(y|x;\theta)} = \frac{\int_{\mathcal{X}} f(y|x;\theta_0) f_X(x;\alpha_0)dx}{\int_{\mathcal{X}} f(y|x;\theta) f_X(x;\alpha_0)dx} = c(y;\theta,\theta_0).$$

By assumption $\mathcal{P}\{D_\theta\} < 1$ for any $\theta \neq \theta_0$. Since $f(r=1|y;\psi_0)$ lies in $(0,1)$ almost surely,

$$E_y[f(r=1|y;\psi_0)E[\log f(x|y,\theta_0,\alpha_0)|y]] > E_y[f(r=1|y;\psi_0)E[\log f(x|y,\theta,\alpha_0)|y]],$$

for any $\theta \neq \theta_0$. □

Proof of Proposition 2.8.3 If $f(y|x,\theta)$ takes this form, then

$$\begin{aligned} f(x|y,\theta,\alpha_0) &= \frac{\exp\{f_1(y,x;\widetilde{\theta}_1)\}}{\int_{\mathcal{X}} \exp\{f_1(y,x;\widetilde{\theta}_1)\} f_X(x;\alpha_0)dx} \\ &= \exp\{f_1(y,x;\widetilde{\theta}_1) + \overline{f}_1(y)\}, \end{aligned}$$

where $\overline{f}_1(y) = -\log \int_{\mathcal{X}} \exp\{f_1(y,x;\widetilde{\theta}_1)\} f_X(x;\alpha_0)dx$. Hence,

$$\frac{f(x|y,\theta,\alpha_0)}{f(x|y,\theta_0,\alpha_0)} = \exp\{f_1(y,x;\widetilde{\theta}_1) - f_1(y,x;\widetilde{\theta}_{10}) + \overline{f}_1(y) - \overline{f}_{10}(y)\}.$$

Since $f_1(y,x;\widetilde{\theta}_1) - f_1(y,x;\widetilde{\theta}_{10})$ cannot be written as a function of y alone, the argument of the exponential function on the right hand side can never be zero. Thus $f(x|y,\theta,\alpha_0)/f(x|y,\theta_0,\alpha_0) \neq 1$. □

Proof of Corollary 2.8.3 Equation (2.68) can be re-arranged as

$$\mathcal{P}\{Y = y_j|X = x_i\} = a_{ij}\mathcal{P}\{Y = y_j\}.$$

Now

$$\sum_{y_j \in \mathcal{Y}} \mathcal{P}\{Y = y_j|X = x_i\} = 1, x \in \mathcal{X}.$$

Let p denote the $I \times 1$ vector with elements for $\mathcal{P}\{Y = y_j\}$, $y_j \in \mathcal{Y}$. Let ι denote the $(J+1)$-vector of units. Hence, the system of linear equations from the above two equations can be

written as $Ap = \iota$. Therefore, there exists a unique p solving this system if and only if A has rank I. □

Chapter 3
A Monte Carlo Comparison of Alternative Estimators

3.1 Introduction

This chapter presents Monte Carlo evidence on the finite sample performance of a subset of the proposed GMM estimators in comparison to other estimators for missing data. A unified model specification framework for all considered estimators is provided in Section 3.2. Inverse probability weighted estimators, unweighted estimators, sample selection model estimators, PL estimators and GMM estimators are then discussed in the following four sections. In each case the assumptions required for consistent estimation are highlighted. Section 3.7 examines the results from a set of Monte Carlo experiments. The experiments are designed to demonstrate the contrasting performance of the estimators in a variety of circumstances. No estimator dominates in all circumstances. In line with the earlier sections, the differing assumptions required of the underlying model for the estimators to display satisfactory finite sample performance are stressed. The summaries are then given in Section 3.8.

3.2 Model Specification for the Missing Data Problem

Let (Y, X) be a $(1 + p_1) \times 1$ random vector and let $\beta_0 \in \mathcal{B} \subset R^{p_1}$ denote the vector of parameters of interest. Consider a linear regression model for the conditional mean of y,

$$y = x'\beta_0 + \varepsilon, \tag{3.75}$$

where ε is an unobserved disturbance. The objective is consistent estimation of, and inference on, β_0.

To provide a unified treatment of the different estimation procedures, we consider models specified by a finite number of moment restrictions. Let $g(\cdot, \cdot, \cdot)$ be a p-vector of known functions. Let $\theta_0 \in \Theta \subset R^p$ denote the true parameter where $p \geq p_1$ and $\mathcal{B}$ is a proper subset of Θ. The dimension of Θ is larger than that of $\mathcal{B}$ because estimation of nuisance parameters is required for some of the procedures analysed. The dimensions of $g(\cdot, \cdot, \cdot)$ and θ are taken to be the same since the focus here is on just-identified cases.

Many conventional estimation procedures for random sampling are based on the fact that θ_0 satisfies uniquely the moment condition

$$E[g(y, x, \theta)] = 0. \tag{3.76}$$

Let $\{(y_n, x_n) : n = 1, ..., N\}$ denote a set of independent, identically distributed $(1 + p_1) \times 1$ random vectors. The sample analog of (3.76) is then $N^{-1} \sum_{n=1}^{N} g(y_n, x_n, \theta) = 0$. Suppose there exists $\widehat{\theta}$ satisfying the sample analog. Then, under regularity conditions which guarantee the uniform convergence in probability of the sample analog to (3.76), $\widehat{\theta}$ is a consistent estimator for θ_0.

The focus here is on cases where data on Y, the endogenous variable, is missing, but the X variables are completely recorded. As in Chapter 2, let R be the response indicator variable whose value is one if Y is recorded and is zero otherwise. The sample analog based on the complete data is then $N^{-1}\sum_{n=1}^{N} r_n \cdot g(y_n, x_n, \theta) = 0$. Under the usual regularity conditions, this sample analog converges uniformly in probability to

$$E[r \cdot g(y, x, \theta)] = 0. \tag{3.77}$$

Without additional assumptions, θ_0, solving (3.76), does not satisfy the moment condition in (3.77). As a consequence, the solution to the sample analog of (3.77) is not a consistent estimator for θ_0 without these additional assumptions.

This chapter compares a number of estimation procedures for correcting the missing-data bias. In essence, each of the estimators suggests an alternative p-vector of known functions, $g_{mis}(\cdot, \cdot, \cdot, \cdot)$, such that θ_0 satisfies uniquely the moment condition

$$E[g_{mis}(y, x, r, \theta)] = 0. \tag{3.78}$$

The form of $g_{mis}(\cdot, \cdot, \cdot, \cdot)$ is usually a modification of $g(\cdot, \cdot, \cdot)$ and is related to a set of extra assumptions imposed to solve the missing data problem. Thus, $g_{mis}(\cdot, \cdot, \cdot, \cdot)$ varies across estimators as they employ different sets of assumptions. The estimators investigated in this chapter are presented in this framework in the next four sections. For each estimator, the assumptions required for consistency are highlighted. However, for simplicity, the assumptions required for identification are not explicitly stated and it will be assumed that θ_0 uniquely solves (3.78) for all estimators.

3.3 Least Squares Estimators

3.3.1 Inverse Probability Weighted Estimators

For the intuition behind this class of estimators, note that the left hand side of (3.77) can be alternatively written as

$$E[\mathcal{P}\{R = 1|Y = y, X = x\} \cdot g(y, x, \theta)]$$

where $\mathcal{P}\{R = 1|Y = y, X = x\}$ is the true MDM as in (2.1) and the unconditional expectation is taken with respect to the joint distribution of (Y, X). This suggests correcting missing-data bias by weighting $g(y, x, \theta)$ with the inverse of the true MDM.

In practice, $\mathcal{P}\{R = 1|Y = y, X = x\}$ is usually unknown and has to be estimated. Since Y is missing for the portion of the sample with $R = 0$, it cannot be included in the estimation of the MDM. Thus, a use of an Inverse Probability Weighted (IPW) estimator implicitly assumes that the MDM is independent of Y. Let Z denote a vector of additional fully observed variables which are related to R. Then, suppose that the model for the response indicator variable R is of the following form

$$r = 1[w'\gamma_0 + \upsilon > 0], \tag{3.79}$$

where υ is an unobserved disturbance and let $W = (X', Z')'$. The presence of Z in (3.79) adds a degree of additional flexibility to the approach. It allows the use of information on other fully observed variables, which are not included in X, to estimate the true MDM. Its presence is not required for identification and its absence from X should not be viewed as a set of identification exclusion restrictions.

Let $p(\cdot,\cdot)$ be a known parametric function for the conditional distribution of υ given W. Let $\gamma_0 \in \Gamma \subset R^{p2}$ denote the true parameter such that $p(w'\gamma_0) = \mathcal{P}\{R = 1|W = w\}$. Let $\widehat{\gamma}$ denote a consistent estimator of γ_0. Then, the IPW estimator, $\widehat{\theta}$, satisfies the sample condition

$$N^{-1}\sum_{n=1}^{N}[r_n/p(w_n'\widehat{\gamma})]\cdot g(y_n,x_n,\theta) = 0. \tag{3.80}$$

Under regularity conditions which ensure the uniform convergence of (3.80) to (3.76), the IPW estimator is consistent for θ_0. Wooldridge (2002a) presents such conditions, shows the asymptotic normality of this class of IPW estimators and gives a consistent estimator for the asymptotic variance. A review of the literature on IPW estimators is given in Wooldridge (2002a, 2003).

IPW estimation can be applied to a general class of nonlinear models. It is also interesting to note that if $\widehat{\gamma}$ is a ML estimator, Wooldridge (2002a) shows that using $\widehat{\gamma}$ leads to a more efficient IPW estimator than using γ_0, the true value.

The required conditions for this estimator can be collected in the following assumption:

Assumption 3.3.1: *(i)* $\mathcal{P}\{R = 1|Y = y, W = w\} = \mathcal{P}\{R = 1|W = w\}$, *i.e., the MDM is MAR; (ii)* $p(\cdot,\cdot)$ *is correctly specified; (iii)* $g(\cdot,\cdot,\cdot)$ *is correctly specified; (iv)* $E[\varepsilon|x] = 0$; *(v)* υ is independent of ε and W.

Assumption 3.3.1(i) is a crucial assumption for IPW estimation. It holds if R is a deterministic function of W, i.e., υ is a constant. Assumption 3.3.1(v) is needed because υ is usually stochastic. If Assumption 3.3.1(v) does not hold, MAR is violated because Y is related to R through the relationship between two unobserved disturbances, ε and

υ. Intuitively, Assumption 3.3.1(i) does not allow R to depend on Y because we do not observe Y for nonrespondent units and, as a result, we cannot use it to estimate the MDM.

Given Assumptions 3.3.1(i) and 3.3.1(v), the MDM can depend on any fully observed variable. Theoretically, such a variable can even be correlated with ε. For example, suppose there exists a completely recorded variable Z^* such that $corr(z^*, \varepsilon) \neq 0$, that is, it is endogenous for the structural equation. Because υ is independent of ε, Z^* may be included in W if it is also independent of υ. In this situation, we can see that the model allows Y to relate to R through the observed variable Z^*, but not through the unobserved disturbance υ.

Variables such as Z^* are often encountered in practice. Skinner et al. (2002), for example, studied the use of the Labour Force Survey (LFS) wage rate variable. Evidence suggests that this new "direct" measure of hourly pay, introduced in March 1999, is more accurate than the more standardly used measure which is "derived" from questions on earnings and working hours. The difficulty in using the direct measure is that it is observed only on a subset of individuals in the LFS. The old derived measure is however treated as completely recorded. This problem can be interpreted as a missing data problem where the direct measure is the dependent variable in our setting. Thus, the derived measure is a good candidate for Z^* because it is fully observed and is correlated with Y (and ε). This example is explored further in Chapter 4.

Assumptions 3.3.1(ii) and 3.3.1(iii) can be relaxed in some settings. Scharfstein, Rotnitzky and Robins (1999) refer to such a property of IPW estimators as double robustness. That is, for certain choices of the objective function and of the response probabil-

ity function, the IPW estimator is still consistent even when one of them is misspecified. Wooldridge (2003) discusses an example in the econometric literature where the property is applicable.

In our Monte Carlo investigation, we focus on a special case of IPW estimators where $\theta = \beta$ and $g(y, x, \theta) = \frac{\partial}{\partial \beta}(y - x'\beta)^2$. In other words, $\widehat{\theta}$ is an IPW Least Squares (IPWLS) estimator. In this setting, the form of $g_{mis}(y, w, r, \theta)$ is thus $[r/p(w, \widehat{\gamma})] \cdot \frac{\partial}{\partial \beta}(y - x'\beta)^2$.

An advantage of this choice of $g(y, x, \theta)$ is that, except for Assumptions 3.3.1(iv) and 3.3.1(v), no distributional assumption is imposed on ε. Moreover, a constant conditional variance assumption for ε given X is not required for the consistency and asymptotic normality of the IPWLS estimator and is therefore not included in Assumption 3.3.1. If heteroskedasticity is suspected, heteroskedasticity-robust standard errors can be used.

3.3.2 Unweighted Estimators

The solution of (3.77) can be referred to as an unweighted estimator. Provided some additional assumptions are satisfied, it can be shown to be consistent for θ_0. These extra assumptions are grouped below:

Assumption 3.3.2: *(i) for each* $x \in \mathcal{X}$, θ_0 *satisfies the moment condition* $E[g(y, x, \theta)|x] = 0$; *(ii)* θ_0 uniquely satisfies (3.77); *(iii)* $\mathcal{P}\{R = 1|Y = y, X = x\} = \mathcal{P}\{R = 1|X = x\}$; *(iv)* $E[\varepsilon|x] = 0$.

Assumption 3.3.2(i) is stronger than (3.76) because it implies (3.76) but not vice versa. In practice, it holds if the econometric model of interest is correctly specified. It is therefore equivalent to Assumption 3.3.1(iii) but it cannot be relaxed as in the case of IPW

estimation. Assumption 3.3.2(ii) is needed to ensure that the information from respondent units is rich enough to identify θ_0. Assumption 3.3.2(iii) is a version of MAR. It is more restrictive than Assumption 3.3.1(i) since the MDM can depend only on the conditioning variables. Under some regularity conditions and Assumption 3.3.2, Wooldridge (2002a) proves the consistency of the unweighted estimators.

The main virtue of unweighted estimators is that it does not require specification and estimation of the MDM. The standard estimation and inference procedures based on the censored sample are valid under these assumptions. On the other hand, the main weakness of unweighted estimators is the strength of the assumptions required. Wooldridge (2002a, 2003) gives an extensive and detailed comparison, in terms of consistency and efficiency, between IPW and unweighted estimators. In the Monte Carlo investigation below, we consider the LS version of the unweighted estimators and the form of $g_{mis}(y, x, r, \theta)$, which is the same as $g(y, x, \theta)$, is $\frac{\partial}{\partial \beta}(y - x'\beta)^2$.

3.4 Sample Selection Model Estimators

This section examines the Sample Selection (SS) model estimators made popular by Heckman (1976). The method also maintains the structure in (3.75) and (3.79) as same as IPW estimation. However, IPW estimation assumes that ε is independent of υ whereas the SS model estimators allow ε to be correlated with υ. Thus, MDM for the SS model is NMAR. To put it differently, Y is related to R through correlation between unobservables, ε and υ. There are two alternative sets of assumptions leading to two different methods of estimating the SS model.

3.4.1 Heckman's Two-Step Estimators

Given (3.75) and (3.79), Wooldridge (2002b) specifies the following set of assumptions:

Assumption 3.4.1: *(i)* (ε, υ) *is independent of* W *with zero mean; (ii)* υ *is normally distributed with zero mean and unit variance; (iii)* $E[\varepsilon|\upsilon] = c_1\upsilon$ *where* c_1 *is a constant parameter.*

Assumption 3.4.1(i) guarantees that W is exogenous in both equations. Assumptions 3.4.1(ii) and 3.4.1(iii) are required for the derivation of the conditional expectation of Y given W and $R = 1$. Assumption 3.4.1(ii) demands both correct specification and normality, which makes it stronger than Assumption 3.3.1(ii). Assumption 3.4.1(iii) is implied if (ε, υ) is bivariate normal. However Assumption 3.4.1(iii) also holds under weaker assumption than bivariate normality.

Under Assumption 3.4.1, the conditional expectation of Y given W and $R = 1$ is

$$E[y|w, r = 1] = x'\beta_0 + c_1\frac{\phi(w'\gamma_0)}{\Phi(w'\gamma_0)}, \tag{3.81}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote respectively the probability density and the cumulative distribution functions of the standard normal distribution. The term $\phi(\cdot)/\Phi(\cdot)$ is called *the inverse Mills ratio*. From (3.81), Heckman proposes a two-step estimator: (i) estimate γ_0 using the probit model based on all observations; (ii) obtain $(\widehat{\beta}, \widehat{c}_1)$ by regressing y on x and $\phi(w'\widehat{\gamma})/\Phi(w'\widehat{\gamma})$ using the completely recorded observations only, where $\widehat{\gamma}$ is the probit estimate from the first step.

The model allows other fully recorded variables, Z, to enter (3.79) via W as long as Assumption 3.4.1(i) holds. These additional variables assist the identification of the SS

model and they are therefore *exclusion restrictions*. In this regard, the SS model is more restrictive than IPW estimation since a variable such as Z^*, from the previous section, is now endogenous for (3.79). IPW estimation gains this flexibility by imposing that ε and υ are independent, which is contrary to the underlying assumptions of the SS model.

The exclusion restrictions are not theoretically required in the estimation of the SS model. In their absence, as can be seen from (3.81), the identification of $\widehat{\beta}$ depends on the nonlinearity of the inverse Mills ratio. If the variation of $x'\widehat{\gamma}$ is relatively low in the sample, $\phi(x'\widehat{\gamma})/\Phi(x'\widehat{\gamma})$ can be highly correlated with X, leading to a collinearity problem. Exclusion restrictions can solve this problem and, in practice, at least one exclusion restriction is desired in the application of the SS model.

The form of $g(y, x, \theta)$ in this setting is $\frac{\partial}{\partial\beta}(y - x'\beta)^2$ and estimation based on it is inconsistent because the term $\phi(w'\gamma)/\Phi(w'\gamma)$ is omitted. The appropriate objective function $g_{mis}(y, w, r, \theta)$ is given by $\frac{\partial}{\partial\theta}[r \cdot (y - x'\beta - c_1[\phi(w'\widehat{\gamma})/\Phi(w'\widehat{\gamma})])^2]$, where $\theta = (\beta', c_1)'$. Because of this, Heckman (1979) indicates that the missing data problem can be seen as an omitted regressor problem.

3.4.2 The Partial Maximum Likelihood Estimator

Following Wooldridge (2002b), given (3.75) and (3.79), partial Maximum Likelihood (ML) estimation can be used if the following stronger assumption holds:

Assumption 3.4.2: *(i) Assumption 3.4.1(i) hold; (ii)* (ε, υ) *is bivariate normal with mean zero,* $Var(\varepsilon) = \sigma_\varepsilon^2, Cov(\varepsilon, \upsilon) = \sigma_{\varepsilon,\upsilon}$ *and* $Var(\upsilon) = 1$.

Assumption 3.4.2(ii) is stronger than Assumptions 3.4.1(ii) and 3.4.1(iii). If it holds then the partial ML estimator will be more efficient than the two-step estimator. To use the ML estimation, the joint likelihood of observed data, $[f(y, w, r = 1)]^r [f(w, r = 0)]^{1-r}$, is factorised as

$$[f(y|r = 1, w)f(r = 1|w)f(w)]^r [f(r = 0|w)f(w)]^{1-r}, \tag{3.82}$$

where $f(w)$ can be dropped out because it is ancillary for the parameters of interest. This method of estimation is called partial ML because we can only use the density $f(y|r, w)$ when $R = 1$. This is because $\int_y f(y|r = 0, w)dy = 1$ and it therefore does not make any contribution for the observed likelihood of nonrespondent units.

Moreover, $f(y|r = 1, w)$ can be written as $[f(r = 1|y, w)f(y|w)]/f(r = 1|w)$. Notice that $f(r = 1|y, w)$ is the MDM and all of these densities are known under Assumption 3.4.2. Firstly, $f(y|w)$ is $\frac{1}{\sigma_\varepsilon}\phi\left(\frac{y-x'\beta}{\sigma_\varepsilon}\right)$ because ε is normally distributed and Z is ignorable in the conditional mean equation. Secondly, $f(r = 1|w)$ is $\Phi(w'\gamma)$ due to the structure of (3.79). Lastly, the bivariate normality of (ε, υ) implies that the specification of the MDM is

$$f(r = 1|y, w) = \Phi\left(\frac{1}{\sqrt{1 - (\sigma^2_{\varepsilon,\upsilon}/\sigma^2_\varepsilon)}}\left(w'\gamma + \frac{\sigma_{\varepsilon,\upsilon}}{\sigma_\varepsilon}\left(\frac{y - x'\beta}{\sigma_\varepsilon}\right)\right)\right). \tag{3.83}$$

Thus, under Assumption 3.4.2, the individual log-likelihood function can be written as

$$\begin{aligned}\log L_n^{SS}(\theta) &= r_n \log\left[\frac{1}{\sigma_\varepsilon}\phi\left(\frac{y_n - x_n'\beta}{\sigma_\varepsilon}\right)\right] + (1-r)\log\left[1-\Phi(w_n'\gamma)\right] + \\ & r_n \log\Phi\left(\frac{1}{\sqrt{1-(\sigma_{\varepsilon,v}^2/\sigma_\varepsilon^2)}}\left(w_n'\gamma + \frac{\sigma_{\varepsilon,v}}{\sigma_\varepsilon}\left(\frac{y_n - x_n'\beta}{\sigma_\varepsilon}\right)\right)\right). \quad (3.84)\end{aligned}$$

The form for $g_{mis}(y, w, r, \theta)$ in this setting is thus $\frac{\partial}{\partial\theta}[\log L^{SS}(\theta)]$, where θ denotes now $(\beta, \gamma, \sigma_\varepsilon, \sigma_{\varepsilon,v})$. If the missing-data bias is ignored, the objective function $g(y, x, \theta)$ will be $\frac{\partial}{\partial\theta}[\log(\sigma_\varepsilon^{-1}\phi(\sigma_\varepsilon^{-1}[y - x'\beta]))]$ since the conditional distribution of Y given X is normal. The form of $g(y, x, \theta)$ is exactly the same as the first term on right hand side of (3.84). We can therefore view the other terms in (3.84) as bias-correcting terms for the SS model estimator relative to the normal linear-regression model estimator.

3.5 The Pseudolikelihood Estimators

Tang, Little and Raghunathan (2003) or TLR proposes PL estimators, which are described in Chapter 2, for dealing with missing data on the dependent variable. In contrast to IPW estimation, the PL estimators maintains that $\mathcal{P}\{R = 1|Y = y, X = x\} = \mathcal{P}\{R = 1|Y = y\}$, that is, the MDM, conditional on Y, is independent of X.

It is shown in Section 2.7.1 that the most efficient PL estimator maximises two alternative pseudolikelihood functions, namely, (2.43) and (2.44). (2.44) is adopted for the Monte Carlo experiments since it does not require first-step estimation of $f_X(x)$. For convenience, it is reproduced here:

$$\log L_n^{PL}(\theta) = r \log f(y_n|x_n;\beta,\sigma_\varepsilon) - r \log[N^{-1}\sum_{j=1}^{N} f(y_n|x_j;\beta,\sigma_\varepsilon)]. \tag{3.85}$$

Moreover, the required conditions for the PL estimator are presented in the following assumption:

Assumption 3.5: *(i)* $\mathcal{P}\{R = 1|Y = y, X = x\} = \mathcal{P}\{R = 1|Y = y\}$*; (ii)* $f(\cdot|\cdot;\beta,\sigma_\varepsilon)$ *is correctly specified; (iii)* $E[\varepsilon|x] = 0$.

The estimator does not require the MAR assumption, but does assume missingness to be independent of X in Assumption 3.5(i). This assumption can be regarded as a strong version of the NMAR Assumption. Assumption 3.5(ii) is quite strong and is equivalent to maintaining that the distribution of $\varepsilon|x$ in (3.75) is known.

From (3.85), $\frac{\partial}{\partial\theta}[\log L^{PL2}(\theta)]$ is then given by

$$\theta : r\left[\frac{\partial \log f(y|x;\beta,\sigma_\varepsilon)}{\partial\theta} - \frac{1}{\sum_{n=1}^{N} f(y|x_n;\beta,\sigma_\varepsilon)}\sum_{j=1}^{N}\frac{\partial f(y|x_j;\beta,\sigma_\varepsilon)}{\partial\theta}\right]. \tag{3.86}$$

The form of $g_{mis}(y,x,r,\theta)$ in this setting is the moment indicator in (3.86), giving a GMM interpretation of the PL estimator. If the nonrespondent units are discarded and conventional ML estimation used, the objective function is $\frac{\partial}{\partial\theta}[\log f(y|x;\beta,\sigma_\varepsilon)]$, which is $g(y,x,\theta)$ in this case. By comparing this to (3.86), one can see that the second term in (3.86) acts as the bias-correcting term for the PL estimator in comparison to the standard conditional ML estimator. In simulation studies, we specify $f(y|x;\beta,\sigma_\varepsilon)$ as $\frac{1}{\sigma_\varepsilon}\phi\left(\frac{y-x'\beta}{\sigma_\varepsilon}\right)$ which means that $\varepsilon|x$ is assumed to be normally distributed.

3.6 GMM Estimators

3.6.1 Discrete Y

For the setting where Y is discrete, Ramalho and Smith (2003) or RS also exploits Assumption 3.5(i) and propose GMM estimators for various types of missing data. In this chapter we focus on the case where only values of Y are missing. The RS GMM estimator for this case is in fact the GMM estimator which is based on moment indicators (2.16) in Section 2.3.2 and has already been examined in Section 2.7.2. For comparison purposes, some results from Chapter 2 regarding this estimator are reiterated below.

Suppose that Y and X are discrete with respective sample spaces $\mathcal{Y}$ and $\mathcal{X}$. The joint observed data likelihood of a sampling unit is then given by

$$\left[\frac{H_y}{Q_y}\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}\mathcal{P}_X\{x\}\right]^r\left[(1-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\})\mathcal{P}_X\{x\}\right]^{1-r}. \tag{3.87}$$

Notice the differences between (3.87) and the observed data likelihood under the SS model in (3.82). The forms of both observed data likelihoods are dissimilar which is expected because different assumptions are imposed. One of the crucial dissimilarities is that, whereas (3.82) is a partial likelihood, (3.87) is a full likelihood since the density $\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}$ is used both when $R=1$ and $R=0$. The linear structure such as (3.79) is also not imposed on any part of (3.87). Unlike the SS model, the density of covariates, $\mathcal{P}_X\{x\}$, cannot be dropped out of (3.87) since Q_y is modeled as $\sum_{x\in\mathcal{X}}\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}\mathcal{P}_X\{x\}$. Furthermore, the MDM of the SS model is shown to be (3.83) due to the assumption of bivariate normality. Here, the MDM is H_y/Q_y and only a part of Q_y, the conditional density of Y given X, is parametrically specified.

From (3.87), since X is discrete, we can replace the unknown $\mathcal{P}_X\{\cdot\}$ by the (unknown) probability π_x associated with each mass point x, $x \in \mathcal{X}$. Thus, the individual log-likelihood contribution can be written as

$$\begin{aligned}\log L_n^{RS}(\theta) &= r[\log H_{y_n} + \log \mathcal{P}\{y_n|x_n;\beta,\sigma_\varepsilon\} + \log \pi_{x_n} - \log Q_{y_n}] \\ &\quad + (1-r)\,[\log(1 - \sum_{y\in\mathcal{Y}} \frac{H_y}{Q_y}\mathcal{P}\{y|x_n;\beta,\sigma_\varepsilon\}) + \log \pi_{x_n}], \qquad (3.88)\end{aligned}$$

where $Q_y = \sum_{x\in\mathcal{X}} \mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}\pi_x$.

To obtain the first-order conditions, we maximises (3.88) with respect to $(H_y, \pi_x, \beta, \sigma_\varepsilon)$ and subject to the constraint that $\sum_{x\in\mathcal{X}} \pi_x = 1$. Clearly, these first-order conditions are dependent on the nuisance parameter π_x. However, RS show that it is possible to rewrite these conditions such that they are no longer dependent on π_x. The resultant conditions can then be used as moment conditions in the GMM estimation. These GMM moment indicators are given by

$$\begin{aligned}H_t &: r1[y=t] - \frac{H_t}{Q_t}\frac{(1-r)\mathcal{P}\{t|x;\beta,\sigma_\varepsilon\}}{1-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}}, t\in\mathcal{Y}; \qquad (3.89)\\ \beta &: r\frac{\partial \log \mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}}{\partial\beta} - \frac{(1-r)}{1-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}}\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\frac{\partial\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}}{\partial\beta};\\ \sigma_\varepsilon &: r\frac{\partial \log \mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}}{\partial\sigma_\varepsilon} - \frac{(1-r)}{1-\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}}\sum_{y\in\mathcal{Y}}\frac{H_y}{Q_y}\frac{\partial\mathcal{P}\{y|x;\beta,\sigma_\varepsilon\}}{\partial\sigma_\varepsilon};\\ Q_y &: Q_y - \mathcal{P}\{y|x;\beta\}, y\in\mathcal{Y}.\end{aligned}$$

Notice that the parameters of the optimisation problem are changed from $(H_y, \pi_x, \beta, \sigma_\varepsilon)$ to $(H_y, \beta, \sigma_\varepsilon, Q_y)$. Thus, the moment indicators in (3.89) are free from the dependence on $\mathcal{P}_X\{\cdot\}$ and, therefore, remain valid even if X is allowed to be continuous. In such a case,

RS also prove that the GMM estimator attains the semiparametric efficiency bound. To implement it, the following assumption is required:

Assumption 3.6.1: *(i) Assumption 3.5 holds*; *(ii)* Y *is discrete*.

The GMM estimator requires Assumption 3.5, the same as that required for the PL estimator. An important difference between them is that the RS GMM estimator is inapplicable when Y is continuous. Furthermore, a considerable advantage of the GMM and PL estimators is that they do not require the specification of a MDM such as that in (3.79). However, some may argue that it is strong to assume that X has no effect on R conditional on Y as in Assumption 3.5(i). As show in Section 2.6.1, this assumption can be weakened within RS's framework to allow the MDM to depend on discrete Y and on a finite partition of the continuous covariates.

In this setting, $g_{mis}(y, x, r, \theta)$ is obtained by stacking the moment indicators in (3.89) and $\theta = (H_y, \beta, \sigma_\varepsilon, Q_y)$. Observe the moment indicator for β and σ_ε in (3.89). The first terms are exactly the score functions for $(\beta, \sigma_\varepsilon)$ of the standard conditional ML estimator applied to the completely recorded observations. Thus, the second terms are the bias correcting terms and the form of $g(y, x, \theta)$ is $\frac{\partial}{\partial(\beta', \sigma_\varepsilon)'}[\log \mathcal{P}\{y|x; \beta, \sigma_\varepsilon\}]$.

In the simulation studies, Y is generated as a continuous variable. To utilise this GMM estimator, we discretise values of Y from respondent units into ten groups; a method which entails certainly a loss in information. Our process of discretisation relies on the deciles of Y, conditional on $R = 1$, to ensure that each group has similar number of observations. Let Y^d denote the resultant discrete dependent variable. This variable is

related to Y by

$$y^d = \begin{cases} 1 \text{ if } y < Dc_1 \\ 2 \text{ if } Dc_1 \leq y < Dc_2 \\ \vdots \\ 10 \text{ if } Dc_9 < y, \end{cases}$$

where Dc_i is the ith decile, $i = 1, ..., 9$. By defining $Dc_0 = -\infty$ and $Dc_{10} = \infty$, we can write this relationship more compactly as

$$y^d = \sum_{j=1}^{10} j \cdot 1[Dc_{j-1} \leq y < Dc_j].$$

In the Monte Carlo investigation, we assume tha $\varepsilon|x$ is normally distributed. Thus, the probabilities of the discrete outcomes in (3.89) are given by

$$\mathcal{P}\{y^d = j|x; \beta, \sigma_\varepsilon\} = \Phi(\frac{Dc_j - x'\beta}{\sigma_\varepsilon}) - \Phi(\frac{Dc_{j-1} - x'\beta}{\sigma_\varepsilon}). \tag{3.90}$$

This specification will be used to construct the moment indicators in (3.89). Unlike the ordered probit model, σ_ε is identified in this setting because the cutoff points, Dc_i $(i = 1, ..., 9)$ are known (Stewart, 1983).

3.6.2 Continuous Y

A limitation of the RS GMM estimator is that Y cannot be continuous. A solution of this problem has been proposed in Chapter 2. As before, some results are restated here for comparison purposes. Let Y and X be continuous variables and let $h(y, r = 1)$ denote the joint population probability of Y and $R = 1$. This probability can be estimated using the completely recorded observations. Let $h(y; \psi_0)$ denote the parametric estimate for $h(y, r = 1)$ based on the respondent units. Let θ denote now $(\beta', \sigma_\varepsilon)'$. Thus, the moment

indicator for β and σ_ε in (3.89) can be rewritten as

$$\theta : r\frac{\partial \log f(y|x;\beta,\sigma_\varepsilon)}{\partial\theta} - \frac{(1-r)}{1-\int_{\mathcal{Y}}\frac{h(y;\psi)}{Q(y;\beta,\sigma_\varepsilon)}f(y|x;\beta,\sigma_\varepsilon)dy}\left[\int_{\mathcal{Y}}\frac{h(y;\psi)}{Q(y;\beta,\sigma_\varepsilon)}\frac{\partial f(y|x;\beta,\sigma_\varepsilon)}{\partial\theta}dy\right], \tag{3.91}$$

where $Q(y;\beta,\sigma_\varepsilon) = N^{-1}\sum_{n=1}^{N} f(y|x_n;\beta,\sigma_\varepsilon)$. This GMM estimator is referred to as the two-step INRY-GMM estimator in Section 2.4.2. The conditions in (3.91) are the continuous version of the moment indicators for β and σ_ε in (3.89). By comparing (3.91) to (3.86), one can see that the two-step INRY-GMM estimator extracts more information from nonrespondent units than the PL estimator at the cost of estimating $h(y, r = 1)$.

The GMM estimator based on (3.91) is operational under the following assumption:

Assumption 3.6.2: *(i) Assumption 3.5 holds*; *(ii)* $h(\cdot;\psi)$ *is correctly specified.*

Assumption 3.6.2(ii) can be relaxed and this GMM estimator can be extended to be a two-step semiparametric estimator if $h(y, r = 1)$ is estimated by a nonparametric method such as kernel or series estimators. In fact, it is shown that if $h(y, r = 1)$ is replaced by its empirical estimator, $N^{-1}\sum_{n=1}^{N} r \cdot 1[y_n = y]$, then (3.91) can be re-expressed as

$$\theta : r\frac{\partial \log f(y|x;\beta,\sigma_\varepsilon)}{\partial\theta} - \frac{(1-r)}{1-\frac{1}{N}\sum_{j=1}^{N_r}\frac{f(y_j|x;\beta,\sigma_\varepsilon)}{Q(y_j;\beta,\sigma_\varepsilon)}dy}\left[\frac{1}{N}\sum_{k=1}^{N_r}\frac{1}{Q(y_k;\beta,\sigma_\varepsilon)}\frac{\partial f(y_k|x;\beta,\sigma_\varepsilon)}{\partial\theta}dy\right], \tag{3.92}$$

where N_r is the number of nonrespondent units. Intuitively, in (3.92), the integration with respect to Y and $h(y, r = 1)$ is approximated by the summation across values of Y from respondent units only. The GMM estimator based on (3.92) is the one-step INRY-GMM estimator as illustrated in Section 2.4.2. It does not require the specification of $h(y, r = 1)$ and, therefore is valid under Assumption 3.5, which is weaker than Assumption 3.6.2.

Within this framework, it is even possible to relax Assumption 3.5(i) which is imposed on the MDM. Notice that both the marginal distribution of X, $f_X(x)$, and $h(y, x, r = 1)$ can be estimated from the observed sample since only values of Y are missing among nonrespondent units. For simplicity, suppose there exists two parametric models $f_X(x; \widehat{\alpha})$ and $h(y, x; \widehat{\psi})$ such that $f_X(x; \alpha_0) = f_X(x)$ and $h(y, x; \psi_0) = h(y, x, r = 1)$. Thus, the MDM or $\mathcal{P}\{R = 1|Y = y, X = x\}$ can be modelled as $h(y, x; \psi_0)/[f(y|x; \theta_0) \cdot f_X(x; \alpha_0)]$. This knowledge of the MDM enables use to write the moment indicators in this case as

$$\theta : r\frac{\partial \log f(y|x; \beta, \sigma_\varepsilon)}{\partial \theta} - (1 - r)\frac{1}{1 - \int_{\mathcal{Y}} \frac{h(y,x;\psi)}{f_X(x;\alpha)} dy} \left[\int_{\mathcal{Y}} \frac{h(y, x; \psi)}{f_X(x; \alpha)} \frac{\partial \log f(y|x; \beta, \sigma_\varepsilon)}{\partial \theta} dy\right]. \tag{3.93}$$

This GMM estimator is the INRYX-GMM estimator presented in Section 2.4.1. It suggests that one can relax Assumption 3.5(i) at the cost of assuming correct specification of both $f_X(x; \widehat{\alpha})$ and $h(y, x; \widehat{\psi})$. Nevertheless, it is possible to estimate $h(y, x, r = 1)$ and $f_X(x)$ by a nonparametric method. In such a case, consistent estimation of θ_0 only requires correct specification of $f(y|x; \beta, \sigma_\varepsilon)$, which is weaker than Assumption 3.5.

As in the previous subsection, $g_{mis}(y, x, r, \theta)$ is obtained by stacking the moment indicators in either (3.91), (3.92) or (3.93). Again the form of $g(y, x, \theta)$ is $\frac{\partial}{\partial \theta}[\log f(y|x; \beta, \sigma_\varepsilon)]$ and the second terms in both (3.91), (3.92) and (3.93) are the bias correcting terms relative to the standard conditional ML estimator. In addition, $f(y|x; \beta, \sigma_\varepsilon)$ is specified as $\frac{1}{\sigma_\epsilon}\phi\left(\frac{y - x'\beta}{\sigma_\epsilon}\right)$ in our Monte Carlo experiments. In these experiments, the INRYX-GMM estimator is excluded and the one-step INRY-GMM estimator is implemented in the baseline experiment only. The reason, which is further explained in Section 3.7.1, is that the computing time required for these estimators is considerable.

3.7 A Monte Carlo Investigation

This section presents Monte Carlo evidence on the finite sample performance of the estimators discussed above. The Monte Carlo experiments presented are designed to illustrate the performance of the estimators in a variety of circumstances. For IPW estimation, we will analyse the IPWLS estimator specified in Section 3.3.1. Similarly, we will examine the special cases of PL and GMM estimators which are described in Sections 3.5 and 3.6 respectively. For simplicity, the RS GMM estimator in (3.89) will be referred to as DGMM estimator to reflect its *discrete* nature. Accordingly, the two-step and one-step INRY-GMM estimators in (3.91) and (3.92) are termed as CGMM and CGMM1 estimators since they allow the response variable to be *continuous*. $h(y|r = 1)$ in (3.91) is assumed to be normal distributed for the first-step estimation. The integration in the bias correcting term of (3.91) is evaluated using Gauss-Hermite Quadrature with 32 points. The SS model will be estimated using both two-step estimation (SSTS) and partial ML estimation (SSML).

Each of the aforementioned estimators makes use of information available on non-respondent units in some way. Additional estimators to be analysed, which use only information on the complete observations, are unweighted Least Squares (ULS) and Interval Regression (INTREG) estimators. The ULS estimator is discussed in Section 3.3.2. We would like to compare its finite sample performance to that of the IPWLS estimator since, under Assumption 3.3.1, the IPWLS estimator can correct the bias of the ULS estimator which is caused by missing data.

The INTREG estimator is a conditional ML estimator maximising

$$N^{-1}\sum_{n=1}^{N} r_n \log \mathcal{P}\{y_n^d|x_n;\beta,\sigma_\varepsilon\},$$

where $\mathcal{P}\{y^d|x;\beta,\sigma_\varepsilon\}$ is as defined in (3.90). This estimator can be considered as an unweighted estimator where the distribution of $\varepsilon|x$ is specified. It is included for comparison since the DGMM estimator can be regarded as correcting the missing-data bias in the INTREG estimator. In other words, the form of $N^{-1}\sum_{n=1}^{N} r_n \cdot g(y_n, x_n, \theta)$, associated with the DGMM estimator, is the first-order conditions of the objective function of the INTREG estimator with respect to β and σ_ε. Note that the similar relationship exists between the ULS and CGMM estimators and it is another justification for the inclusion of the ULS estimator[6].

3.7.1 Structure of Monte Carlo Experiments and the Baseline Experiment

In each experiment, the structural model is generated as

$$y = 1 + \beta_1 x_1 + \beta_2 x_2 + \sigma_\varepsilon \cdot \varepsilon, \tag{3.94}$$

where ε has zero mean and unit variance. The Data Generating Process (DGP) uses $\beta_1 = -1$ and $\beta_2 = 1$. The focus of the investigation is on the estimates for β_1 and β_2. The number of replications is set to 500. Moreover, tables of Monte Carlo results are shown in the appendix 3.A.

---

[6] Strictly speaking, the CGMM estimator should remove the missing-data bias in the normal linear regression. However, although the ULS estimator does not require normality, it gives the same results as those of the normal linear regression. These two estimators are therefore equivalent for comparison purposes.

For the baseline experiment, ε is *i.i.d.* normally distributed with $\sigma_\varepsilon = 1$ and the sample size is 3000. Both covariates are generated to be non-normally distributed to prevent special results which might occur if they were normally distributed. They are also generated such that the explanatory power of the covariates in the model, measured by R^2, is roughly 20% to reflect the nature of micro-level data found in practice. These covariates are generated as follows. The variable X_1 has mean 0.7411 and variance 0.2272. It is generated as a mixture of two normals. The first has mean 0.502 and variance 0.0814 with a probability of 0.7. The second has mean 1.3386 and variance 0.0814 with a probability of 0.3. The scalar covariate X_2 is uniformly distributed over $(0, 0.53)$. The linear correlation between X_1 and X_2 is approximately zero. Both covariates are fixed across replications and experiments.

In the baseline experiment, the MDM is

$$r^* = 0.5 - 0.9y + \upsilon, \tag{3.95}$$

where $r = 1[r^* > 0]$ and υ is normally distributed with zero mean and unit variance. The specification of the MDM in (3.95) is chosen because the primary interest is in non-ignorable missing data. The intercept and the coefficient on Y in (3.95) are selected such that the proportion of missing observation is 0.50. This means that, on average, the number of respondent units is 1500.

There is no correlation between ε and υ in the baseline experiment (because they are independently generated). Nonetheless, we can use (3.94) to rewrite (3.95) as

$$r^* = -0.4 + 0.9x_1 - 0.9x_2 + \upsilon^*, \tag{3.96}$$

where $\upsilon^* = \upsilon - 0.9\varepsilon$. This implies that υ^* is correlated with ε and is normally distributed. The relationship between (3.95) and (3.96) is useful in analysing the performance of SS model.

Table 3.1 shows the Monte Carlo results for the baseline experiment. The table reports the mean bias and root mean square error (RMSE) for each estimator. Although the CGMM1 estimator is also implemented, its results are not reported in the table. The mean bias and RMSE for β_1 from the CGMM1 estimator are 0.0081 and 0.0763 whereas these statistics for β_2 are 0.0029 and 0.1701. We decide to present these results separately and to exclude the CGMM1 estimator from other experiments since it is time-consuming and its behaviour with regard to the factors of interest should be similar to that of the CGMM estimator. We use Stata/Intercool 9.0, which is available on a server, to run all experiments on a 2.8 GHz Pentium 4 Windows XP machine. The time spent in running the experiment with PL estimator is 16.5929 minutes for a replication. This means that an experiment with 500 replications takes almost 6 days. The time spent on the experiment of the CGMM1 estimator is even longer, which is why it is excluded from the subsequent experiments. This point is also of considerable concern if a researcher wishes to use these estimators on a large data set.

Accordingly, the DGMM, CGMM, CGMM1 and PL estimators show small mean bias and RMSE for both β_1 and β_2 as expected, since all assumptions underlying these estimators are satisfied in the baseline experiment; especially the MDM in (3.95). The sizes of RMSE for the GMM estimators imply that the variance of the DGMM estimator

is bigger than those of the CGMM and CGMM1 estimators. This indicates the anticipated loss of information from discretising Y.

The CGMM1 estimator outperforms the PL estimator in terms of both mean bias and RMSE. Interestingly, although the PL estimator produces considerably smaller bias than the CGMM estimator, its variance is bigger than that of the CGMM estimator. This supports our theoretical speculation that the framework of RS and HS should yield estimators which are more efficient than the PL estimation because it uses more information from the nonrespondent units. Notice that the DGMM estimator is designed for data with discrete Y and should not be compared directly to the PL estimator. In the table, the performance of DGMM estimator is inferior to the PL estimator and this may due partly to the discretisation.

It is clear from the table that the DGMM and CGMM estimators correct the missing data bias of the INTREG and ULS estimators. The results from the INTREG and ULS estimators are also very alike. Since the INTREG estimator can be considered as a discretised version of the ULS estimator, the results imply that the loss of information from discretisation in this case is small. The results suggest that discretisation affects the DGMM estimator more than INTREG. Note that, the IPWLS estimator does not remove the bias in the ULS estimator because Assumptions 3.3.1(i) and 3.3.1(ii) do not hold in this experiment.

Since (3.95) can be written as (3.96), the underpinning assumptions of the SS model are satisfied. One should therefore anticipate the SSML and SSTS estimators to perform well in this experiment. However, this is not found to be the case in Table 3.1. For the

SSTS estimator, although the size of its mean bias is as small as that of the CGMM estimator, it has the biggest RMSE in the table. Similarly, the RMSE of the SSML estimator is as large as those of INTREG, ULS and IPWLS estimators which are supposed to perform poorly.

There are two possible explanations for the poor performance of the SSML and SSTS estimators. Firstly, the explanatory power of X_1 and X_2 in the MDM is very low. The pseudo R-squared of the probit model of R^* on X_1 and X_2 is 0.0495. This means that the variation of R^* is explained mostly by the variation in v^*. Secondly, there is no exclusion restriction in (3.96). These points are, of course, not mutually exclusive. They are explored further in Monte Carlo experiments below.

3.7.2 Deviations from the Baseline Experiment

The finite sample performance of the estimators for missing data considered here is likely to be influenced by:

- the proportion of missing observations;
- the correlation between R^* and Y;
- the explanatory power of covariates in the structural model;
- the specification of the MDM;
- the distribution of ε;
- the distribution of v;

- the correlation between R^* and covariates;
- the differences between the covariates in the MDM and structural model;
- the number of observations.

This list is used to structure the Monte Carlo investigation. Deviations from the baseline experiment in this list are considered, one at a time, to see the change in the finite sample performance of the estimators. The characteristics of following Monte Carlo experiments are compared and presented in Table 3.2.

The Proportion of Missing Observations

In the baseline experiment, the proportion of missing observations is 0.5. In this section, we show results from two other experiments where the proportions of missing observations are set to 0.25 and 0.75 while other aspects, in the above list, are similar to the baseline experiment. The MDMs of these experiments are

$$\begin{aligned} \text{Ex 1} &: r^* = 1.45 - 0.9y + v, \\ \text{Ex 2} &: r^* = -0.48 - 0.9y + v. \end{aligned} \tag{3.97}$$

The proportions of missing data are changed while maintaining the same level of $Corr(r^*, y)$ by manipulating only the intercepts of the MDMs.

Tables 3.3 and 3.4 report the mean bias and RMSE of the estimates for β_1 and β_2, respectively. The relationship between the estimators in Experiments 1 and 2 remains the same as that in the baseline experiment. The DGMM, CGMM and PL estimators still per-

form better than other estimators in both experiments. These findings do not appear to be sensitive to the degree of missingness. The most apparent effect of changing the proportion of missing observations is on the sizes of mean bias and of RMSE for all estimators. That is, a lower (higher) proportion of missing data corresponds to a smaller (bigger) bias and RMSE. This result is as expected since lesser information should lead to greater imprecision in estimation.

The Correlation between R^* and Y

Experiments 3 and 4 depart from the baseline experiment in terms of correlation between R^* and Y. We increase the correlation measured by $Corr(r^*, y)$ relative to the baseline experiment in Experiment 4 and decrease it in Experiment 3. The MDMs are specified as follows

$$\text{Ex 3} \quad : \quad r^* = 0.25 - 0.5y + v,$$

$$\text{Ex 4} \quad : \quad r^* = 0.88 - 1.6y + v.$$

Table 3.5 and 3.6 show the Monte Carlo results from the deviations. The GMM and PL estimators still dominate other estimators; especially in terms of RMSE. The effect of these deviations on the INTREG, ULS and IPWLS is considerable. Their performance improves remarkably as the correlation decreases. In Experiment 3, their RMSE is even smaller than that of the DGMM estimator. This outcome is attributable to the fact that, as

the correlation decreases, the MDM is influenced more by υ and the intercept term. Thus, it is closer to being missing completely at random or MCAR.

The RMSE of the SSTS estimator is quite sensitive to the changes in the correlation between R^* and Y. It becomes the biggest in each table in Experiment 3 where pR^2_1, the explanatory power of the covariates in the MDM, is at its lowest due to the decline in the correlation between R^* and Y.

The Explanatory Power of Covariates in the Structural Model

The standard deviation of ε, σ_ε, is manipulated to change the explanatory power of X_1 and X_2 in the structural model. In the baseline experiment, $\sigma_\varepsilon = 1$ and the explanatory power, which is measured by R^2, is 0.2040. To see the effects of positive and negative deviations, σ_ε is set to be 1.45 and 0.25 respectively in Experiments 5 and 6. Accordingly, R^2 in these experiments becomes 0.1504 and 0.5053. To fix other factors, the MDMs of Experiments 5 and 6 are

$$\text{Ex 5} \quad : \quad \sigma_\varepsilon = 1.45,\ r^* = 0.45 - 0.76y + \upsilon,$$

$$\text{Ex 6} \quad : \quad \sigma_\varepsilon = 0.25,\ r^* = 0.75 - 1.4y + \upsilon.$$

As can be seen in Tables 3.7 and 3.8, changes in the explanatory power do not affect the dominance of the GMM and PL estimators. An increase in R^2 obviously boosts the finite sample performance of all estimators, especially in terms of RMSE. Nevertheless, the effect is more evident among the GMM, PL, SSML and SSTS estimators than others. The RMSE of the SSTS estimator in Experiment 6 becomes smaller than that of INTREG,

ULS and IPWLS estimators. This is because an increase in R^2 leads to a rise in pR^2_1 through the correlation between R^* and Y. It also confirms the conjecture in Section 3.7.1 that low pR^2_1 is a factor causing the poor performance of the SSML and SSTS estimators in the base line experiment.

A Nonlinear MDM

The MDM of Experiment 7 is set to be nonlinear in Y, as opposed to the linearity of the MDM in the baseline experiment. The specification of this MDM is

$$\text{Ex } 7 : r^* = 0.45 - 0.76y - 0.05y^3 + v. \tag{3.98}$$

Because of the additional variable Y^3, this MDM can no longer be rewritten in a linear form as in (3.96). The SSML and SSTS estimators assume that the selection equation is linear in X_1 and X_2. Thus, (3.98) implies that the selection equations of the SSML and SSTS estimators are misspecified in Experiment 7.

Tables 3.9 and 3.10 illustrate mean bias and RMSE of all estimators in both Monte Carlo experiments. The GMM and PL estimators exhibit only a small mean bias and dominate other estimators in this experiment. These estimators do not require specification of the MDM and, consequently, nonlinearity of (3.98) does not affect their finite sample behaviour.

There is no dramatic change in the results of the INTREG, ULS and IPWLS estimators. They still show comparatively large bias and RMSE in the experiment. The performance of the SSTS estimator in Experiment 7 is worse than that in the baseline ex-

periment as expected. However, surprisingly the performance of the SSML estimator has improved despite the misspecification of selection equation.

The Distribution of ε

In Experiments 8 and 9, we depart from the baseline experiment by changing the distribution of ε. This is a type of misspecification as it is equivalent to asserting that the conditional density function of Y given X is misspecified. To hold other factors relatively constant, the MDMs for these experiments are formulated as

$$\text{Ex 8} \quad : \quad \varepsilon \sim \text{ Gamma}, \qquad r^* = 0.45 - 0.95y + \upsilon,$$

$$\text{Ex 9} \quad : \quad \varepsilon \sim \text{ Normal Mixture}, \; r^* = 0.4 - y + \upsilon.$$

In Experiment 8, the gamma distribution is used to generate ε, which is standardised to zero mean and unit variance. This results in ε being positively skewed. In Experiment 9, ε is generated as a mixture of two normals. The first has $\mu = -0.5, \sigma = 0.25$ with a probability of 0.75. The second has $\mu = 1.5, \sigma = 0.9$ with a probability of 0.25. The combination is chosen to produce ε with zero mean and unit variance.

Note that, in addition to the missing data problem, there is also a misspecification problem. Tables 3.11, 3.12 and 3.13 show the Monte Carlo results for β_1, β_2 and β_0. In all three tables, the SSML and SSTS estimators produce a sizeable mean bias because neither ε nor υ^* is normal.

The results for β_0 are included in this section because the GMM and PL estimators demonstrate an interesting property. That is, although there is a clear bias in the estimates

of β_0, these estimators have only a small bias in β_1 and β_2[7]. This suggests that they may be able to give a consistent estimate for all slope coefficients even when the conditional model is misspecified. This result is unexpected and should motivate a further investigation into the properties of these estimators.

Observe also that, in Experiments 8 and 9, the CGMM estimator is outperformed by both the DGMM and PL estimators. In applying the CGMM estimator, $h(y|r = 1)$ is assumed to be normal in the first-step estimation. Since the conditional model of Y given X is no longer normal in both experiments, it is unlikely that $h(y|r = 1)$ will be well approximated by the normal distribution. This may explain the poor performance of the CGMM estimator in these settings.

Another unanticipated result is that the INTREG, ULS and IPWLS estimators also exhibit the same property as the GMM and PL estimators when the distribution of ε is normal mixture. This means that they remarkably overcome both missing data and misspecification problems. For β_1 and β_2, their RMSE are even smaller than that of the CGMM estimator. However, unlike the GMM and PL estimators, these estimators do not exhibit this property in Experiment 8 where ε is gamma distributed.

---

[7] As well as β_0, β_1 and β_2, the GMM and PL estimators also produce an estimate of σ_ε for every experiment but we do not show such results here. In Experiments 8 and 9, there is a clear bias in the estimates of σ_ε. Thus, the GMM and PL estimators cannot estimate consistently both β_0 and σ_ε when the conditional distribution of y given x is misspecified.

The Distribution of v

This section investigates the effect of misspecification in the MDM by changing the distribution of v. The same specifications of gamma and normal mixture distributions as in the previous section are used. The MDMs for Experiments 10 and 11 are given by

$$\text{Ex 10} \quad : \quad v \sim \text{Gamma}, \qquad r^* = 0.6 - 0.9y + v,$$

$$\text{Ex 11} \quad : \quad v \sim \text{Normal Mixture}, r^* = 0.63 - 0.9y + v.$$

Since the GMM and PL estimators do not require correct specification for the MDM, they are robust against these changes as shown in Tables 3.14 and 3.15. In contrast, the SSML and SSTS estimators give inconsistent estimates because v and v^* are no longer normally distributed. The INTREG, ULS and IPWLS estimators still display a considerable bias as in the baseline experiment because the MDMs are not MAR.

The Correlation between R^* and X_1

In this section, we will allow the MDM to depend on a covariate, namely, X_1 and then vary the correlation between R^* and X_1 across experiments. As in the baseline experiment, ε and v are independent in the experiments analysed here. In Experiment 12, the MDM depends on both Y and X_1 but the correlation between R^* and Y is higher, in absolute term, than the correlation between R^* and X_1. In Experiment 13, the correlation between R^* and X_1 is increased such that it becomes higher than $corr(r^*, y)$. In Experiment 14,

the MDM depends only on X_1 and the correlation between R^* and X_1 is -0.6901 which is close to $corr(r^*, y)$ in the baseline experiment. The MDMs are specified as

$$\text{Ex 12} \quad : \quad r^* = 0.08 - 0.8y + 0.5x_1 + v,$$

$$\text{Ex 13} \quad : \quad r^* = -1.5 - 0.8y + 2.8x_1 + v,$$

$$\text{Ex 14} \quad : \quad r^* = 1.43 - 2x_1 + v.$$

The results of these experiments are shown in Tables 3.16 and 3.17. Notice that, for all estimators, the mean bias and RMSE of β_2 are generally smaller than those of β_1.

The estimates from the GMM and PL estimators show clear bias and sizable RMSE in all three experiments since all MDMs are dependent on X_1. An exception is CGMM estimates for β_2 in Experiments 12 and 13 although the corresponding RMSE is considerable.

The INTREG, ULS and IPWLS estimators perform poorly in Experiments 12 and 13 because the MDMs are NMAR. Their performances improve markedly in Experiment 14. For the INTREG and ULS estimators, this is expected because (i) the conditional mean model is correctly specified and (ii) the MDM depends on a conditioning variable. The IPWLS estimator also performs well in Experiment 14 since the MDM is MAR and is correctly specified. In other words, both Assumptions 3.3.1 and 3.3.2 are satisfied. In this case, Wooldridge (2002a, Theorem 5.3) proves that an unweighted estimator is more efficient than a weighted estimator. This point is confirmed in Experiment 14 since the RMSE of the IPWLS estimator is bigger than those of the unweighted estimators.

The SSML and SSTS estimators demonstrate relatively small bias in all experiments. For Experiments 12 and 13, this is because their MDMs can be shown to satisfy the assumptions of SS model, i.e., one can rewrite the MDMs as

$$\begin{aligned} r^* &= -0.72 + 1.3x_1 - 0.8x_2 + \upsilon^*, \\ r^* &= -2.3 + 3.6x_1 - 0.8x_2 + \upsilon^*, \end{aligned} \tag{3.99}$$

where $\upsilon^* = \upsilon - 0.8\varepsilon$ and is therefore correlated with ε. For Experiment 14, even though ε and υ are independent, only X_1 is included in the selection equation and, thus, the MDM is correctly specified. Moreover, the correction for missing-data bias is not needed in this experiment as indicated by the results of the ULS estimator. The RMSE of both SSML and SSTS estimators in Experiments 13 is smallest because the explanatory power of covariates in the selection equation increases to 40.27% as a result of a rise in $corr(r^*, x_1)$.

The Differences between Sets of Covariates

There are a number of interesting issues in this deviation. There are thus four experiments conducted to explore these issues. Consider first Experiments 15 and 16 which are constructed to see the effect of having an additional variable in the MDM or selection equation. It is also shown here that Assumption MAR in IPW estimation is not as restrictive as it may seem. The MDM of Experiment 15 depends on all covariates and an extra variable Z_1, which is normally distributed with zero mean and unit variance. In Experiment 16, Z_1 is replaced by $Z_2 = Y + \kappa$, where κ is also normally distributed with zero mean and unit variance. The disturbances, ε and υ, are independent in both experiments. Accordingly, the MDMs are given by

$$\text{Ex 15} \quad : \quad r^* = 1.1 - 0.9x_1 - 1.5x_2 - z_1 + \upsilon,$$

$$\text{Ex 16} \quad : \quad r^* = 1.85 - 1.5x_1 - 0.9x_2 - 0.9z_2 + \upsilon.$$

The Monte Carlo results are reported in Tables 3.18 and 3.19. There is a clear bias in the GMM and PL estimators. The only exception is the bias of the DGMM estimates for β_1, which is relatively small. However, the RMSE of all estimates from these three estimators is considerable. The unsatisfactory performance of the estimators is due to the fact that $\mathcal{P}\{R = 1|Y = y, X = x\} \neq \mathcal{P}\{R = 1|Y = y\}$ in these experiments.

The IPWLS estimator should outperform the INTREG and ULS estimators in Experiments 15 since the MDM also depends on Z_1, which is not a conditioning variable in the conditional mean model. Nevertheless, the tables show that the finite sample performance of the INTREG and ULS estimators is better than that of the IPWLS estimator, especially in terms of RMSE. This may be because the influence of X_1 and X_2 on the MDM is stronger than that of Z_1.

In Experiment 16, Z_2 can be considered as Y with a measurement error. Here, $\mathcal{P}\{R = 1|Y = y, X = x, Z_2 = z_2\} = \mathcal{P}\{R = 1|X = x, Z_2 = z_2\}$, i.e., Assumption MAR is satisfied because ε and υ are independent. As a result, the IPWLS estimator has only a small bias in Experiment 16. Notice that we can rewrite the MDM of Experiment 16 as

$$\begin{aligned} r^* &= 1.85 - 1.5x_1 - 0.9x_2 - 0.9(y + \kappa) + \upsilon, \\ &= 0.95 - 0.6x_1 - 1.8x_2 + \upsilon^*, \end{aligned} \tag{3.100}$$

where $\upsilon^* = \upsilon - 0.9\varepsilon - 0.9\kappa$. In the rewritten MDM, υ^* is correlated with ε, the disturbance of the structural model. Thus, if variables such as Z_2 can be observed, Assumption MAR in IPW estimation does allow the correlation between υ^* and ε or, alternatively, between R^* and Y. This is one of the most attractive properties of the IPW estimation. Moreover, the fact that $R^*$is correlated with Y may partly cause the INTREG and ULS estimators to have a sizable bias in Experiment 16.

If Z_2 is observed in the baseline experiment and is used in the first-step MDM estimation of the IPWLS estimator, the performance of the estimator will be improved even though the MDM in (3.95) is a function of Y only. The mean bias for β_1 and β_2 reduces to 0.1721 and -0.1755 and their RMSE becomes 0.1821 and -0.1755. Note that these mean bias and RMSE are smaller than those of the INTREG and ULS estimators in the baseline experiment. This implies that, with the availability of variables such as Z_2, the IPWLS estimator can even cope with the endogenous selection[8].

The SSML and SSTS estimators perform well in Experiment 15 even though ε and υ are independent. This should be a result of having Z_1 as an exclusion restriction and specifying correctly the MDM. On the other hand, both estimators have large bias in Experiment 16. This is because Z_2 is an endogenous variable and Assumption 3.4.1(i) is consequently violated when it is included in the MDM. To use the SS model, variables such as Z_2 must be dropped from the selection equation.

---

[8] The correlation between Z_2 and Y from the baseline experiment are 0.7460. If variables with similar or higher degree of correlation are used in the MDM estimation, the IPWLS estimator can produce even smaller mean bias.

Moreover, we have previously suggested two reasons why, given the equivalence between (3.95) and (3.96), the SSML and SSTS estimators do not perform well in the baseline experiment. That is, the explanatory power of X_1 and X_2 in the selection equation is low and there is no exclusion restriction. It is then shown in Experiments 5 and 6 that low explanatory power of the covariates is indeed a reason of the poor performance of the SS model. Accordingly, the following two experiments are designed to investigate these two reasons further. Note that it is difficult to untangle the effects of these two reasons because adding an exclusion restriction also leads to an increase in the explanatory power of the covariates in the MDM. Nevertheless, exclusion restriction should improve both performance and identification of the SS model. In contrast, an increase in the explanatory power cannot help identifying the SS model.

In Experiments 17 and 18, the intercept and the coefficients for X_1 and X_2 are the same as those in (3.96). However, υ and ε are generated as bivariate normal to explicitly justify the use of the SS model. Both disturbances have zero mean and unit variance and the correlation between them is 0.9. Thus, the explanatory power of X_1 and X_2 in Experiment 17 is higher than that in the baseline experiment because υ has only a unit variance whereas the variance of υ^* in (3.96) is greater than one . In Experiment 18, we also introduce an additional variable Z, which is normally distributed with zero mean and unit variance, as an exclusion restriction. This results in even higher pseudo R-squared. To sum up, the MDMs for these experiments are given by

$$\text{Ex 17} \quad : \quad r^* = -0.4 + 0.9x_1 - 0.9x_2 + \upsilon,$$

$$\text{Ex 18} \quad : \quad r^* = -0.4 + 0.9x_1 - 0.9x_2 + 0.5z + \upsilon.$$

Tables 3.20 and 3.21 report the Monte Carlo results from the experiments. The INTREG, ULS and IPWLS estimators show large mean bias in both experiments. This is because Assumption MAR does not hold in these experiments as a consequence of nonzero correlation between υ and ε.

As observed in Experiment 6, the mean bias of the SSML and SSTS estimators reduces significantly in Experiment 17 due to an increase in the explanatory power of the covariates in the selection equation. However, the RMSE of both estimators is still relatively high considering that the SS model is the correct specification. By adding an exclusion restriction in Experiment 18, the explanatory power increases from 0.0848 to 0.1612, which is close to that of Experiment 6, and the resultant RMSE of the SSML and SSTS estimators become smaller than those of their counterparts in Experiments 6 and 17. These two estimators also outperform other estimators in Experiment 18. Although this result is not clear-cut, it seems to suggest that having exclusion restrictions has a bigger effect on the finite sample performance of the SS model than an increase in the explanatory power of the covariates in the MDM.

The GMM and PL estimators perform reasonably well in both experiments even if $\mathcal{P}\{R = 1|Y = y, X = x\} \neq \mathcal{P}\{R = 1|Y = y\}$ in Experiment 18 due to the presence of Z in the MDM. This is because these estimators do not require the correct specification

of the MDM. In addition, the pseudo R-squared of the MDM is relatively low in both experiments whereas the correlation between υ and ε is very high. This implies that the variation in R^* is well explained by the variation in Y as shown in the tables by the high correlation between R^* and Y. As a result, the assumption, imposed on the MDM, of the GMM and PL estimators is indirectly satisfied in Experiments 17 and 18.

The Number of Observations

In Experiment 19, we decrease the sample size from 3000 to 1500 observations while using the same specification of the MDM as in the base line experiment. Thus, the summary statistics of Experiment 19, reported in Table 3.22, are almost the same as those of the base line experiment except the number of the observations. Even though an increase of the sample size is also interesting, it is omitted here because it would be time-consuming to apply some estimators in such a case. Also, some anticipated asymptotic properties can already be observed even at the sample size of 3000 observations. Thus, it is more worthwhile to consider the effect of a smaller sample size.

Table 3.22 and 3.23 present the Monte Carlo results of Experiment 19. As can be seen, the relationships between all estimators are unchanged in comparison to the baseline experiment. The GMM and PL estimators still dominate other estimators in the table. The most apparent effect of reducing the sample size is larger RMSE in all estimators. However, this effect is smaller in the INTREG, ULS and IPWLS estimators.

3.8 Summary

This chapter presents Monte Carlo evidence on the finite sample performance of various estimators for missing data. As expected, no estimator dominates in all experiments. For experiments where the underlying assumptions of the GMM and PL estimators hold, the CGMM estimator dominates the DGMM and PL estimators in terms of RMSE. Although this is a finite sample property, it supports the theoretical conjecture in Chapter 2 that the framework proposed should produce more efficient estimator than the PL estimator since it extracts more information from the nonrespondent units. It also does not lose any information to the discretisation of values of Y as in the case of RS GMM estimators, which explains the aforementioned result. Moreover, if only the conditional density function of Y given X is misspecified, it seems that the estimates for the slope coefficients from the CGMM, DGMM and PL estimators remain consistent. In such a case, the performance of the CGMM estimator is however inferior to that of the DGMM and PL estimators.

There are several experiments for which all of the GMM, PL and SS model estimators are applicable as indicated by, for example, the equivalence between (3.95) and (3.96). In such experiments, if there is no exclusion restriction and the explanatory power of the covariates in the MDM is low, the SS model estimators are dominated and perform poorly especially in terms of RMSE. However, the Monte Carlo investigation also suggests that if there are exclusion restrictions and the MDM is explained well by the variation of the covariates, the SS model estimators will outperform the CGMM, DGMM and PL estimators. This is due partly to the fact that these GMM and PL estimators do not allow the MDM to depend on the covariates. Thus, it will be interesting to see in a future study a compari-

son between the SS model estimators and the INRYX-GMM estimator which permits the MDM to be a function of both response variable and covariates.

In addition, the Monte Carlo study shows that the INTREG and ULS estimators are more efficient than the IPWLS estimator if all of their associated assumptions are satisfied. This finding confirms the mathematical proof of the same result in Wooldridge (2002a). In such a circumstance, the MAR assumption only allows the MDM to depend on the covariates in the conditional mean model, which is restrictive. If the MDM is correlated with Y and a variable such as Z_2 in Experiment 16 is available, the IPWLS estimator will perform better than the INTREG and ULS estimators. This is because, in IPW estimation, the MDM is estimated and Z_2 can be used in the estimation to take account of the correlation between R^* and Y.

Note that, with the availability of variables such as Z_2, the IPWLS estimator is able to cope with the endogenous selection. An advantage of the IPWLS estimator over the SS model estimators is that it is easier to find a fully observed variable which is correlated with Y than finding an exclusion restriction. In contrast, the GMM and PL estimators require neither exclusion restriction nor Z_2 in dealing with the endogenous selection.

A natural extension of this study is to replace some of the estimators under consideration with their semiparametric counterparts. Various semiparametric SS model estimators are already available in the literature. For IPWLS estimation, one can apply a number of semiparametric binary response models as the first-step estimation. As above, it will also be interesting to compare these semiparametric estimators to the INRYX-GMM estimator in (3.93) which allows the MDM to depend on both Y and X.

3.A Appendix A: Tables of Results from the Monte Carlo Experiments

Table 3.1
Monte Carlo results for the baseline experiment

Estimator	β_1		β_2	
	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	0.0472	0.2221
CGMM	0.0293	0.0903	−0.0194	0.1671
PL	−0.0094	0.1256	0.0208	0.1985
INTREG	0.2711	0.2754	−0.2607	0.3023
ULS	0.2673	0.2712	−0.2580	0.2984
IPWLS	0.2738	0.2778	−0.2665	0.3077
SSML	0.1166	0.2177	−0.1046	0.2576
SSTS	−0.0344	0.5832	0.0414	0.6117
$corr(r^*, y)$	−0.7097			
$corr(\varepsilon, \upsilon^*)$	−0.6685			
$P\{r = 0\}$	0.4934			
R^2	0.2040			
pR^2	0.2822			
pR^2_1	0.0495			

Notes:

1. Estimators: DGMM = RS's GMM estimator, CGMM = two-step INRY-GMM estimator, PL = Pseudolikelihood estimator, INTREG = Interval regression, ULS = Unweighted least squares regression, IPWLS = Inverse probability weighted least squares estimator, SSML = Maximum likelihood sample selection model estimator, SSTS = Two-step sample selection model estimator
2. pR^2 is a short term for a pseudo R-squared
3. pR^2 is obtained from a probit model of R^* on Y
4. pR^2_1 is obtained from a probit model of R^* on X_1 and X_2
5. All descriptive statistics are the averages from 500 replications

Table 3.2
The Characteristics of Monte Carlo Experiments

No.	Missing Prop.	R^2	pR^2	*Corr* (r^*, y^*)	γ_0	γ_1	γ_2	γ_3	Other variations/Notes
Baseline	0.50	0.20	0.28	−0.7097	0.50	−0.90	0	0	
1	0.25	0.20	0.28	−0.7097	1.45	−0.90	0	0	
2	0.75	0.20	0.29	−0.7097	−0.48	−0.90	0	0	
3	0.50	0.20	0.12	−0.4883	0.25	−0.50	0	0	
4	0.50	0.20	0.50	−0.8731	0.88	−1.60	0	0	
5	0.50	0.15	0.28	−0.7040	0.40	−0.76	0	0	$\sigma = 1.45$
6	0.50	0.51	0.28	−0.7050	0.75	−1.40	0	0	$\sigma = 0.25$
7	0.50	0.20	0.28	−0.7332	0.45	−0.76	−0.05	0	$r^* = f(y^*, (y^*)^3)$
8	0.50	0.20	0.27	−0.7285	0.45	−0.95	0	0	$\varepsilon \sim$ Gamma
9	0.50	0.20	0.27	−0.7452	0.40	−1.00	0	0	$\varepsilon \sim$ Normal Mixture
10	0.50	0.20	N/A	−0.7088	0.60	−0.90	0	0	$\upsilon \sim$ Gamma
11	0.50	0.20	N/A	−0.7094	0.63	−0.90	0	0	$\upsilon \sim$ Normal Mixture
12	0.50	0.20	0.29	−0.6983	0.08	−0.80	0.50	0	$r^* = f(y^*, x_1), Corr(r^*, x) = 0.4363$
13	0.50	0.20	0.53	−0.6843	−1.50	−0.80	2.80	0	$r^* = f(y^*, x_1), Corr(r^*, x) = 0.8009$
14	0.50	0.20	0.27	0.2970	1.43	0	−2.00	0	$r^* = f(x_1), Corr(r^*, x) = -0.6901$
15	0.51	0.20	0.32	0.1012	1.10	−0.90	−1.50	−1.00	$r^* = f(x, z_1), Corr(r^*, z_1) = -0.6990$
16	0.50	0.20	0.38	−0.4330	1.85	−1.50	−0.90	−0.90	$r^* = f(x, z_1), Corr(r^*, z_1) = -0.6826, z_1 = f(y^*)$
17	0.51	0.20	0.08	−0.9178	−0.40	0.90	−0.90	0	$\varepsilon, \upsilon \sim Bi, rho = -0.9, r^* = f(x_1, x_2)$
18	0.51	0.20	0.16	−0.8353	−0.40	0.90	−0.90	0.50	$\varepsilon, \upsilon \sim Bi, rho = -0.9, r^* = f(x_1, x_2, z_1)$
19	0.51	0.21	0.29	−0.7119	0.50	−0.90	0	0	1,500 observations

Table 3.3
The effect of changes in the proportion of missing observations

Estimator	β_1 Baseline		Ex 1		Ex 2	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.0172	0.0954	0.0096	0.2801
CGMM	0.0293	0.0903	0.0243	0.0651	0.0544	0.1429
PL	−0.0094	0.1256	−0.0028	0.0786	−0.0193	0.2018
INTREG	0.2711	0.2754	0.1906	0.1951	0.3275	0.3334
ULS	0.2673	0.2712	0.1879	0.1920	0.3252	0.3308
IPWLS	0.2738	0.2778	0.1917	0.1958	0.3332	0.3397
SSML	0.1166	0.2177	0.0630	0.1368	0.1998	0.3256
SSTS	−0.0344	0.5832	0.0055	0.2741	−0.1076	1.3759
$corr(r^*, y)$	−0.7097		−0.7097		−0.7097	
$corr(\varepsilon, \upsilon^*)$	−0.6685		−0.6685		−0.6685	
$P\{r = 0\}$	0.4934		0.2464		0.7493	
R^2	0.2040		0.2040		0.2040	
pR^2	0.2822		0.2875		0.2906	
pR^2_1	0.0495		0.0494		0.0540	

Notes:
1. pR^2 is obtained from a probit model of R^* on Y
2. pR^2_1 is obtained from a probit model of R^* on X_1 and X_2

Table 3.4
The effect of changes in the proportion of missing observations

Estimator	β_2 Baseline		Ex 1		Ex 2	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	0.0195	0.1664	−0.0013	0.3392
CGMM	−0.0194	0.1671	−0.0230	0.1440	−0.0418	0.2305
PL	0.0208	0.1985	0.0073	0.1612	0.0287	0.2732
INTREG	−0.2607	0.3023	−0.1897	0.2320	−0.3213	0.3814
ULS	−0.2580	0.2984	−0.1891	0.2294	−0.3179	0.3764
IPWLS	−0.2665	0.3077	−0.1937	0.2337	−0.3289	0.3942
SSML	−0.1046	0.2576	−0.0613	0.1927	−0.1973	0.3738
SSTS	0.0414	0.6117	−0.0080	0.3105	0.0925	1.4250

Table 3.5
The effect of changes in the correlation between R^* and Y

Estimator	β_1 Baseline		Ex 3		Ex 4	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.0241	0.1582	−0.0517	0.1715
CGMM	0.0293	0.0903	0.0369	0.1057	0.0345	0.0949
PL	−0.0094	0.1256	−0.0080	0.1297	−0.0050	0.1229
INTREG	0.2711	0.2754	0.1233	0.1341	0.4268	0.4288
ULS	0.2673	0.2712	0.1221	0.1317	0.4192	0.4212
IPWLS	0.2738	0.2778	0.1237	0.1335	0.4336	0.4356
SSML	0.1166	0.2177	0.0915	0.1605	0.0813	0.2180
SSTS	−0.0344	0.5832	−0.0331	0.9931	−0.0062	0.4156
$corr(r^*, y)$	−0.7097		−0.4883		−0.8731	
$corr(\varepsilon, v^*)$	−0.6685		−0.4467		−0.8477	
$P\{r = 0\}$	0.4934		0.5049		0.4942	
R^2	0.2040		0.2040		0.2040	
pR^2	0.2822		0.1194		0.4991	
pR^2_1	0.0495		0.0233		0.0764	

Notes:
1. pR^2 is obtained from a probit model of R^* on Y
2. pR^2_1 is obtained from a probit model of R^* on X_1 and X_2

Table 3.6
The effect of changes in the correlation between R^* and Y

Estimator	β_2 Baseline		Ex 3		Ex 4	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	0.0313	0.2234	0.0554	0.2313
CGMM	−0.0194	0.1671	−0.0258	0.1897	−0.0275	0.1617
PL	0.0208	0.1985	0.0193	0.2100	0.0124	0.1938
INTREG	−0.2607	0.3023	−0.1091	0.2039	−0.4242	0.4450
ULS	−0.2580	0.2984	−0.1108	0.2018	−0.4170	0.4374
IPWLS	−0.2665	0.3077	−0.1121	0.2039	−0.4380	0.4577
SSML	−0.1046	0.2576	−0.0771	0.2268	−0.0758	0.2650
SSTS	0.0414	0.6117	0.0398	1.0291	0.0072	0.4384

Table 3.7
The effect of changes in the explanatory power of covariates

Estimator	β_1 Baseline		Ex 5		Ex 6	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.0119	0.2567	−0.0129	0.0484
CGMM	0.0293	0.0903	0.0357	0.1144	0.0179	0.0362
PL	−0.0094	0.1256	−0.0119	0.1784	−0.0012	0.0370
INTREG	0.2711	0.2754	0.2822	0.2883	0.1861	0.1879
ULS	0.2673	0.2712	0.2788	0.2844	0.1835	0.1851
IPWLS	0.2738	0.2778	0.2840	0.2897	0.1951	0.1969
SSML	0.1166	0.2177	0.1343	0.2341	0.0483	0.1290
SSTS	−0.0344	0.5832	−0.0522	0.8528	0.0029	0.1712
$corr(r^*, y)$	−0.7097		−0.7040		−0.7050	
$corr(\varepsilon, \upsilon^*)$	−0.6685		−0.6746		−0.5729	
$P\{r = 0\}$	0.4934		0.5004		0.5000	
R^2	0.2040		0.1504		0.5053	
pR^2	0.2822		0.2765		0.2798	
pR^2_1	0.0495		0.0357		0.1295	

Notes:
1. pR^2 is obtained from a probit model of R^* on Y
2. pR^2_1 is obtained from a probit model of R^* on X_1 and X_2

Table 3.8
The effect of changes in the explanatory power of covariates

Estimator	β_2 Baseline		Ex 5		Ex 6	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	0.0178	0.3111	0.0133	0.0928
CGMM	−0.0194	0.1671	−0.0233	0.2045	−0.0132	0.0831
PL	0.0208	0.1985	0.0260	0.2597	0.0064	0.0878
INTREG	−0.2607	0.3023	−0.2704	0.3275	−0.1838	0.2013
ULS	−0.2580	0.2984	−0.2660	0.3219	−0.1799	0.1958
IPWLS	−0.2665	0.3077	−0.2721	0.3284	−0.1975	0.2147
SSML	−0.1046	0.2576	−0.1182	0.2925	−0.0442	0.1504
SSTS	0.0414	0.6117	0.0623	0.9182	0.0013	0.1889

Table 3.9
The effect of a nonlinear missing data mechanism

Estimator	β_1 Baseline		Ex 7	
	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.0439	0.1663
CGMM	0.0293	0.0903	0.0347	0.0991
PL	−0.0094	0.1256	−0.0081	0.1279
INTREG	0.2711	0.2754	0.2752	0.2796
ULS	0.2673	0.2712	0.2791	0.2830
IPWLS	0.2738	0.2778	0.2893	0.2932
SSML	0.1166	0.2177	0.0424	0.1902
SSTS	−0.0344	0.5832	−0.1946	0.6310
$corr(r^*, y)$	−0.7097		−0.7332	
$P\{r = 0\}$	0.4934		0.5019	
R^2	0.2040		0.2040	
pR^2	0.2822		0.2780	

Notes:
1. pR^2 for Experiment 7 is obtained from a probit model of R^* on Y and Y^3

Table 3.10
The effect of a nonlinear missing data mechanism

Estimator	β_2 Baseline		Ex 7	
	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	0.0494	0.2234
CGMM	−0.0194	0.1671	−0.0250	0.1734
PL	0.0208	0.1985	0.0205	0.2009
INTREG	−0.2607	0.3023	−0.2675	0.3083
ULS	−0.2580	0.2984	−0.2738	0.3112
IPWLS	−0.2665	0.3077	−0.2883	0.3250
SSML	−0.1046	0.2576	−0.0303	0.2525
SSTS	0.0414	0.6117	0.2037	0.6684

Table 3.11
The effect of changes in the distribution of ϵ

Estimator	β_1 Baseline		Ex 8		Ex 9	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	0.0303	0.0815	0.0263	0.0439
CGMM	0.0293	0.0903	0.0888	0.1196	0.0762	0.1660
PL	−0.0094	0.1256	0.0252	0.0637	0.0225	0.0363
INTREG	0.2711	0.2754	0.1622	0.1661	0.0819	0.0845
ULS	0.2673	0.2712	0.1702	0.1737	0.0982	0.1007
IPWLS	0.2738	0.2778	0.1704	0.1737	0.0950	0.0974
SSML	0.1166	0.2177	0.3493	0.4204	0.2576	0.2882
SSTS	−0.0344	0.5832	0.1573	0.4157	0.2230	0.3549
Distribution	Normal		Gamma		Mix	
$corr(r^*, y)$	−0.7097		−0.7285		−0.7452	
$corr(\varepsilon, v^*)$	−0.6685		−0.6884		−0.7063	
$P\{r = 0\}$	0.4934		0.4911		0.4990	
R^2	0.2040		0.2040		0.2040	
pR^2	0.2822		0.2748		0.2743	
pR^2_1	0.0495		0.0557		0.0612	

Notes:
1. pR^2 is obtained from a probit model of R^* on Y
2. pR^2_1 is obtained from a probit model of R^* on X_1 and X_2

Table 3.12
The effect of changes in the distribution of ϵ

Estimator	β_2 Baseline		Ex 8		Ex 9	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	−0.0388	0.1253	−0.0188	0.0705
CGMM	−0.0194	0.1671	−0.0823	0.1459	−0.0887	0.2076
PL	0.0208	0.1985	−0.0318	0.1188	−0.0383	0.0819
INTREG	−0.2607	0.3023	−0.1641	0.1938	−0.0757	0.0992
ULS	−0.2580	0.2984	−0.1643	0.1935	−0.0947	0.1181
IPWLS	−0.2665	0.3077	−0.1632	0.1918	−0.0891	0.1119
SSML	−0.1046	0.2576	−0.3486	0.4386	−0.2625	0.3073
SSTS	0.0414	0.6117	−0.1541	0.4429	−0.2257	0.3789

Table 3.13
The effect of changes in the distribution of ϵ

Estimator	β_0					
	Baseline		Ex 8		Ex 9	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0770	0.1683	−0.2903	0.2981	−0.4179	0.4191
CGMM	−0.0224	0.1052	−0.2997	0.3072	−0.4181	0.4871
PL	0.0036	0.1168	−0.2707	0.2770	−0.3732	0.3748
INTREG	−0.6566	0.6593	−0.5863	0.5878	−0.5310	0.5315
ULS	−0.6612	0.6638	−0.5745	0.5760	−0.5224	0.5231
IPWLS	−0.6647	0.6672	−0.5750	0.5764	−0.5209	0.5216
SSML	−0.2906	0.5286	−0.9781	1.1068	−0.8777	0.9201
SSTS	0.0862	1.4492	−0.5454	1.0461	−0.8078	1.0194

Table 3.14
The effect of changes in the distribution of v

	β_1					
Estimator	Baseline		Ex 10		Ex 11	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.0350	0.1602	−0.0304	0.1592
CGMM	0.0293	0.0903	0.0072	0.0651	−0.0174	0.0710
PL	−0.0094	0.1256	−0.0074	0.1238	−0.0089	0.1244
INTREG	0.2711	0.2754	0.2669	0.2711	0.2667	0.2711
ULS	0.2673	0.2712	0.2537	0.2578	0.2417	0.2462
IPWLS	0.2738	0.2778	0.2521	0.2567	0.2266	0.2323
SSML	0.1166	0.2177	0.2768	0.3154	0.3915	0.4289
SSTS	−0.0344	0.5832	0.2851	0.6123	0.7059	0.8735
Distribution	Normal		Gamma		Mix	
$corr(r^*, y)$	−0.7097		−0.7088		−0.7094	
$corr(\varepsilon, v^*)$	−0.6685		−0.6688		−0.6686	
$P\{r = 0\}$	0.4934		0.4944		0.5012	
R^2	0.2040		0.2040		0.2029	

Notes:
1. pR^2 is not calculated because v is not normally distributed and, as a result, the probit model cannot be used

Table 3.15
The effect of changes in the distribution of v

	β_2					
Estimator	Baseline		Ex 10		Ex 11	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	0.0359	0.2172	0.0299	0.2167
CGMM	−0.0194	0.1671	0.0008	0.1593	0.0264	0.1613
PL	0.0208	0.1985	0.0142	0.1959	0.0076	0.1969
INTREG	−0.2607	0.3023	−0.2547	0.2964	−0.2585	0.3009
ULS	−0.2580	0.2984	−0.2406	0.2847	−0.2306	0.2784
IPWLS	−0.2665	0.3077	−0.2374	0.2859	−0.2063	0.2675
SSML	−0.1046	0.2576	−0.2627	0.3342	−0.3771	0.4380
SSTS	0.0414	0.6117	−0.2740	0.6186	−0.6840	0.8690

Table 3.16
The effect of changes in the correlation between R^* and X

Estimator	β_1							
	Baseline		Ex 12		Ex 13		Ex 14	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.4346	0.4849	0.9889	0.9896	0.7126	0.7140
CGMM	0.0293	0.0903	−0.3085	0.3453	−1.4182	3.8838	0.9967	0.9992
PL	−0.0094	0.1256	−0.2968	0.3345	−6.4636	6.5902	−2.1578	2.1688
INTREG	0.2711	0.2754	0.3663	0.3696	0.6624	0.6651	−0.0015	0.0787
ULS	0.2673	0.2712	0.3626	0.3656	0.6574	0.6599	−0.0003	0.0748
IPWLS	0.2738	0.2778	0.3784	0.3817	0.8705	0.8779	−0.0028	0.1493
SSML	0.1166	0.2177	0.1090	0.2595	0.0027	0.0972	0.0119	0.3149
SSTS	−0.0344	0.5832	−0.0066	0.4068	0.0046	0.1635	0.0235	0.3403
$corr(r^*, y)$	−0.7097		−0.6983		−0.6843		0.2970	
$corr(\varepsilon, v^*)$	−0.6685		−0.6242		−0.6242			
$corr(r^*, x_1)$			0.4363		0.8009		−0.6901	
$P\{r = 0\}$	0.4934		0.4944		0.4981		0.5007	
R^2	0.2040		0.2040		0.2040		0.2040	
pR^2	0.2822		0.2884		0.5314		0.2731	
pR^2_1			0.0994		0.4027			

Notes:

1. The selection equations in Experiments 12 and 13 include X_1 and X_2, not Y and X_1, because Y is not observed
2. pR^2 for Experiments 12 and 13 is obtained from a probit model of R^* on Y and X_1
3. pR^2 for Experiment 14 is obtained from a probit model of R^* on X_1
4. pR^2_1 is obtained from a probit model of R^* on X_1 and X_2

Table 3.17
The effect of changes in the correlation between R^* and X

Estimator	β_2							
	Baseline		Ex 12		Ex 13		Ex 14	
	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	0.1727	0.3010	−0.6249	0.6406	−0.6673	0.6693
CGMM	−0.0194	0.1671	0.0675	0.2052	−0.0031	1.6514	−0.4647	0.5048
PL	0.0208	0.1985	0.1926	0.2884	1.4734	1.5897	0.1251	0.3112
INTREG	−0.2607	0.3023	−0.2151	0.2671	−0.1208	0.2013	0.0108	0.1853
ULS	−0.2580	0.2984	−0.2126	0.2624	−0.1173	0.1978	0.0104	0.1794
IPWLS	−0.2665	0.3077	−0.2308	0.2821	−0.1922	0.3227	0.0203	0.2874
SSML	−0.1046	0.2576	−0.0588	0.2270	0.0041	0.1669	0.0105	0.1796
SSTS	0.0414	0.6117	0.0111	0.2996	0.0033	0.1684	0.0104	0.1794

Table 3.18
The effect of the differences between sets of covariates

Estimator	β_1 Baseline		Ex 15		Ex 16	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	0.0351	0.4638	0.6680	0.6693
CGMM	0.0293	0.0903	0.5297	0.5947	0.4120	0.4173
PL	−0.0094	0.1256	−0.2758	0.3120	−0.2138	0.2394
INTREG	0.2711	0.2754	−0.0006	0.0620	−0.1328	0.1427
ULS	0.2673	0.2712	0.0007	0.0584	−0.1309	0.1399
IPWLS	0.2738	0.2778	−0.0045	0.1171	−0.0440	0.1805
SSML	0.1166	0.2177	0.0000	0.0616	−0.2576	0.2623
SSTS	−0.0344	0.5832	0.0000	0.0616	−0.3077	0.3122
$corr(r^*, z_1)$			−0.6690			
$corr(r^*, z_2)$					−0.6826	
$corr(r^*, y)$	−0.7097		0.1012		−0.4330	
$P\{r = 0\}$	0.4934		0.4904		0.4939	
R^2	0.2040		0.2032		0.2032	
pR^2	0.2822		0.3162		0.3847	

Notes:
1. pR^2 for Experiment 15 is obtained from a probit model of R^* on Y, X_1 and Z_1
2. pR^2 for Experiment 16 is obtained from a probit model of R^* on Y, X_1 and Z_2

Table 3.19
The effect of the differences between sets of covariates

Estimator	β_2 Baseline		Ex 15		Ex 16	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	−0.8056	0.8389	−0.9196	0.9212
CGMM	−0.0194	0.1671	−0.5541	0.6985	−0.8409	0.8520
PL	0.0208	0.1985	−0.2873	0.3864	−0.5283	0.5614
INTREG	−0.2607	0.3023	−0.0009	0.1851	−0.3956	0.4280
ULS	−0.2580	0.2984	−0.0022	0.1815	−0.3964	0.4270
IPWLS	−0.2665	0.3077	−0.0145	0.3210	−0.1114	0.6084
SSML	−0.1046	0.2576	−0.0036	0.1854	−0.7654	0.7802
SSTS	0.0414	0.6117	−0.0036	0.1853	−0.9138	0.9276

Table 3.20
The effect of having an exclusion restriction

Estimator	β_1 Baseline		Ex 17		Ex 18	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.0600	0.1722	−0.0391	0.1640
CGMM	0.0293	0.0903	0.0282	0.1073	0.0369	0.0927
PL	−0.0094	0.1256	−0.0032	0.1216	−0.0043	0.1219
INTREG	0.2711	0.2754	0.4757	0.4774	0.3872	0.3897
ULS	0.2673	0.2712	0.4672	0.4689	0.3805	0.3828
IPWLS	0.2738	0.2778	0.4846	0.4862	0.4782	0.4809
SSML	0.1166	0.2177	0.0616	0.2014	−0.0002	0.0490
SSTS	−0.0344	0.5832	−0.0007	0.3699	−0.0031	0.0608
rho	−0.6685		−0.9000		−0.9000	
$corr(r^*, y)$	−0.7097		−0.9178		−0.8353	
$P\{r = 0\}$	0.4934		0.4924		0.4924	
R^2	0.2040		0.2040		0.2032	
pR^2	0.0495		0.0848		0.1612	

Notes:

1. rho is $corr(\varepsilon, v^*)$ for the baseline experiment and is $corr(\varepsilon, v)$ for Experiments 17 and 18

2. pR^2 for Experiment 17 is obtained from a probit model of R^* on X_1 and X_2

3. pR^2 for Experiment 18 is obtained from a probit model of R^* on X_1 , X_2 and Z

Table 3.21
The effect of having an exclusion restriction

Estimator	β_2 Baseline		Ex 17		Ex 18	
	Bias	RMSE	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	0.0666	0.2316	0.0339	0.2201
CGMM	−0.0194	0.1671	−0.0205	0.1672	−0.0387	0.1659
PL	0.0208	0.1985	0.0100	0.1926	0.0003	0.1927
INTREG	−0.2607	0.3023	−0.4737	0.4914	−0.3880	0.4136
ULS	−0.2580	0.2984	−0.4663	0.4839	−0.3822	0.4081
IPWLS	−0.2665	0.3077	−0.4930	0.5096	−0.4888	0.5134
SSML	−0.1046	0.2576	−0.0568	0.2548	−0.0047	0.1565
SSTS	0.0414	0.6117	0.0009	0.3988	−0.0022	0.1648

Table 3.22
The effect of a decrease in the number of observations

Estimator	β_1 Baseline		β_1 Ex 19	
	Bias	RMSE	Bias	RMSE
DGMM	−0.0424	0.1650	−0.0121	0.2556
CGMM	0.0293	0.0903	0.0320	0.1139
PL	−0.0094	0.1256	−0.0114	0.1854
INTREG	0.2711	0.2754	0.2707	0.2785
ULS	0.2673	0.2712	0.2674	0.2751
IPWLS	0.2738	0.2778	0.2724	0.2805
SSML	0.1166	0.2177	0.1530	0.2621
SSTS	−0.0344	0.5832	0.0119	0.7843
Sample size	3000		1500	
$corr(r^*, y)$	−0.7097		−0.7119	
$corr(\varepsilon, \upsilon)$	−0.6685		−0.6698	
$P\{r = 0\}$	0.4934		0.5067	
R^2	0.2040		0.2095	
pR^2	0.2822		0.2852	
pR^2_1	0.0495		0.0521	

Table 3.23
The effect of a decrease in the number of observations

Estimator	β_2 Baseline		β_2 Ex 19	
	Bias	RMSE	Bias	RMSE
DGMM	0.0472	0.2221	−0.0007	0.3327
CGMM	−0.0194	0.1671	−0.0379	0.2347
PL	0.0208	0.1985	0.0015	0.2901
INTREG	−0.2607	0.3023	−0.2789	0.3506
ULS	−0.2580	0.2984	−0.2770	0.3487
IPWLS	−0.2665	0.3077	−0.2827	0.3546
SSML	−0.1046	0.2576	−0.1584	0.3509
SSTS	0.0414	0.6117	0.0005	0.8053

3.B Appendix B: STATA Programs for Some Estimators

Two-Step INRY-GMM Estimator

```
clear
set more off
set mem 100m

*****************
** Main Program **
*****************

capture program drop simugmm
program simugmm, rclass
      version 9
            drop _all
            global numint 32
            setpoint
            calhy
            // GMM estimation
            gen t = .
            replace t = 1 in 1
            replace t = 0 in 2
            replace t = 0 in 3
            replace t = 0 in 4
            nl gmm @ t ystar x1 x2 r, nparameters(4) iterate(1000)
            mat b = e(b)
      end

      ****************************
      ** Set The Evaluation Points **
      ****************************


      capture program drop setpoint
      program setpoint
      /* abscissas and weights are from http://www.efunda.com/math/num_integration/findgausshermite.cfm
*/
            scalar p1 = -7.12581390983
            scalar p2 = -6.40949814928
            scalar p3 = -5.81222594946
            scalar p4 = -5.27555098664
            scalar p5 = -4.77716450334
            scalar p6 = -4.30554795347
            scalar p7 = -3.85375548542
            scalar p8 = -3.41716749282
```

scalar p10 = -2.57724953773
scalar p11 = -2.16949918361
scalar p12 = -1.76765410946
scalar p13 = -1.37037641095
scalar p14 = -0.97650046359
scalar p15 = -0.584978765436
scalar p16 = -0.194840741569
scalar p17 = 0.194840741569
scalar p18 = 0.584978765436
scalar p19 = 0.97650046359
scalar p20 = 1.37037641095
scalar p21 = 1.76765410946
scalar p22 = 2.16949918361
scalar p23 = 2.57724953773
scalar p24 = 2.99249082501
scalar p25 = 3.41716749282
scalar p26 = 3.85375548542
scalar p27 = 4.30554795347
scalar p28 = 4.77716450334
scalar p29 = 5.27555098664
scalar p30 = 5.81222594946
scalar p31 = 6.40949814928
scalar p32 = 7.12581390983

scalar w1 = 0.824566523071
scalar w2 = 0.640950485906
scalar w3 = 0.561749015435
scalar w4 = 0.515037283347
scalar w5 = 0.48357144163
scalar w6 = 0.460786455454
scalar w7 = 0.443553185862
scalar w8 = 0.430163710393
scalar w9 = 0.419597752949
scalar w10 = 0.411206128685
scalar w11 = 0.404557061809
scalar w12 = 0.399354844618
scalar w13 = 0.395393939396
scalar w14 = 0.392531864366
scalar w15 = 0.390672744629
scalar w16 = 0.389757342027
scalar w17 = 0.389757342027
scalar w18 = 0.390672744629
scalar w19 = 0.392531864366
scalar w20 = 0.395393939396
scalar w21 = 0.399354844618
scalar w22 = 0.404557061809
scalar w23 = 0.411206128685

```
scalar w24 = 0.419597752949
scalar w25 = 0.430163710393
scalar w26 = 0.443553185862
scalar w27 = 0.460786455454
scalar w28 = 0.48357144163
scalar w29 = 0.515037283347
scalar w30 = 0.561749015435
scalar w31 = 0.640950485906
scalar w32 = 0.824566523071
end

********************
** Calculating Hy **
********************

capture program drop calhy
program calhy
    tempname mean sd meanr
    quietly sum ystar if r == 1
    scalar 'mean' = r(mean)
    scalar 'sd' = r(sd)
    quietly sum r
    scalar 'meanr' = r(mean)
    forvalues i=1(1)$numint{
        scalar hy'i' = normden(scalar(p'i'),'mean','sd')*'meanr'
    }
end

*****************************
** MIAN GMM PROGRAM **
*****************************

capture program drop nlgmm
program nlgmm
    version 9
    syntax varlist [if], at(name)
    local t : word 1 of 'varlist'
    local y : word 2 of 'varlist'
    local x1 : word 3 of 'varlist'
    local x2 : word 4 of 'varlist'
    local r : word 5 of 'varlist'
    tempname b0 b1 b2 s
    scalar 'b0' = 'at'[1, 1]
    scalar 'b1' = 'at'[1, 2]
    scalar 'b2' = 'at'[1, 3]
    scalar 's' = 'at'[1, 4]
```

```
tempvar g1i g2i g3i g4i
tempvar th xb z
tempname g1 g2 g3 g4
gen double ‘xb’ = ‘b0’ + (‘b1’*x1) + (‘b2’*x2)
gen double ‘z’ = ‘s’*‘y’ -‘xb’

/* Calculate Qybar */
forvalues i=1(1)$numint{
      tempvar Qy‘i’i
}
forvalues i=1(1)$numint{
      tempname Qy‘i’
}

forvalues i=1(1)$numint{
      gen double ‘Qy‘i’i’ = ‘s’*normden(‘s’*scalar(p‘i’) -‘xb’)
      quietly sum ‘Qy‘i’i’
      scalar ‘Qy‘i” = r(mean)
}

/* Calculate Denominator */
tempvar R Rp Rpp bias
gen double ‘R’ = .
      replace ‘R’ = scalar(w1)*(scalar(hy1)/scalar(‘Qy1’))*‘Qy1i’
      forvalues i=2(1)$numint{
            replace ‘R’ = ‘R’ + scalar(w‘i’)*(scalar(hy‘i’)/scalar(‘Qy‘i”))*‘Qy‘i’i’
      }
      gen double ‘Rp’ = .
      replace ‘Rp’ = scalar(w1)*(scalar(hy1)/scalar(‘Qy1’))*‘Qy1i’*(‘s’*scalar(p1)
-‘xb’)
      forvalues i=2(1)$numint{
            replace ‘Rp’ = ‘Rp’ + scalar(w‘i’)*(scalar(hy‘i’)/scalar(‘Qy‘i”))*‘Qy‘i’i’*(‘s’*scalar(p‘i’)
-‘xb’)
      }

      gen double ‘Rpp’ = .
      replace ‘Rpp’ = scalar(w1)*(scalar(hy1)/scalar(‘Qy1’))*(1/‘s’)*‘Qy1i’*(1-
(‘s’*scalar(p1))*(‘s’*scalar(p1) -‘xb’))
      forvalues i=2(1)$numint{
            replace ‘Rpp’ = ‘Rpp’ + scalar(w‘i’)*(scalar(hy‘i’)/scalar(‘Qy‘i”))*(1/‘s’)*‘Qy‘i’i’*(1-
(‘s’*scalar(p‘i’))*(‘s’*scalar(p‘i’) -‘xb’))
      }

gen double ‘bias’ = 1 - ‘R’
// generate the 1st moment
gen double ‘g1i’ = ‘z’ if ‘r’ == 1
```

```
replace ‘g1i’ = - ((‘Rp’)/‘bias’) if ‘r’ == 0
quietly sum ‘g1i’
scalar ‘g1’ = r(mean)
// generate the 2nd moment
gen double ‘g2i’ = ‘z’*‘x1’ if ‘r’ == 1
replace ‘g2i’ = - ((‘Rp’*‘x1’)/‘bias’) if ‘r’ == 0
quietly sum ‘g2i’
scalar ‘g2’ = r(mean)
// generate the 3rd moment
gen double ‘g3i’ = ‘z’*‘x2’ if ‘r’ == 1
replace ‘g3i’ = - ((‘Rp’*‘x2’)/‘bias’) if ‘r’ == 0
quietly sum ‘g3i’
scalar ‘g3’ = r(mean)
// generate the 4th moment
gen double ‘g4i’ = (1/‘s’)-‘y’*‘z’ if ‘r’ == 1
replace ‘g4i’ = - ((‘Rpp’)/‘bias’) if ‘r’ == 0
quietly sum ‘g4i’
scalar ‘g4’ = r(mean)
gen double ‘th’ = ‘g1’ + 1 in 1
replace ‘th’ = ‘g2’ in 2
replace ‘th’ = ‘g3’ in 3
replace ‘th’ = ‘g4’ in 4
replace ‘t’ = ‘th’
end
```

One-Step INRY-GMM Estimator

```
clear
set more off
set mem 100m
******************
** Main Program **
******************

capture program drop simugmm
program simugmm, rclass
      version 9
            drop _all
            gen double id = r
            replace id = . if id == 0
            sort id, stable
            local N = _N
            global numobs 'N'
            count if r==1
            local NY = r(N)
            global numy 'NY'
            // GMM estimation
            gen t = .
            replace t = 1 in 1
            replace t = 0 in 2
            replace t = 0 in 3
            replace t = 0 in 4
            nl gmm @ t ystar x1 x2 r, nparameters(4) iterate(3000)
end

**********************
** MIAN GMM PROGRAM **
**********************

capture program drop nlgmm
program nlgmm
      version 9
      syntax varlist [if], at(name)
      local t : word 1 of 'varlist'
      local y : word 2 of 'varlist'
      local x1 : word 3 of 'varlist'
      local x2 : word 4 of 'varlist'
      local r : word 5 of 'varlist'
      tempname b0 b1 b2 s
      scalar 'b0' = 'at'[1, 1]
      scalar 'b1' = 'at'[1, 2]
```

```
scalar ‘b2’ = ‘at’[1, 3]
scalar ‘s’ = ‘at’[1, 4]
tempvar g1i g2i g3i g4i Qy
tempvar th xb z
tempname g1 g2 g3 g4
        gen double ‘xb’ = ‘b0’ + (‘b1’*‘x1’) + (‘b2’*‘x2’)
        gen double ‘z’ = ‘s’*‘y’ - ‘xb’

        gen double ‘Qy’ = .
        replace ‘Qy’ = (1/$numobs)*‘s’*normden(‘s’*‘y’ - ‘xb’[1]) if ‘r’ == 1
        forvalues i=2(1)$numobs{
                quietly replace ‘Qy’ = ‘Qy’ + (1/$numobs)*‘s’*normden(‘s’*‘y’
- ‘xb’[‘i’])
        }

        tempvar R Rp Rpp bias
        gen double ‘R’ = .
        replace ‘R’ = (1/$numobs)*‘s’*normden(‘s’*‘y’[1] - ‘xb’)*(1/‘Qy’[1]) if
‘r’ == 0
        forvalues i=2(1)$numy{
            replace ‘R’ = ‘R’ + (1/$numobs)*‘s’*normden(‘s’*‘y’[‘i’] - ‘xb’)*(1/‘Qy’[‘i’])
        }

        gen double ‘Rp’ = .
        replace ‘Rp’ = (1/$numobs)*‘s’*(‘s’*‘y’[1] - ‘xb’)*normden(‘s’*‘y’[1] -
‘xb’)*(1/‘Qy’[1]) if ‘r’ == 0
        forvalues i=2(1)$numy{
            replace ‘Rp’ = ‘Rp’ + (1/$numobs)*‘s’*(‘s’*‘y’[‘i’] - ‘xb’)*normden(‘s’*‘y’[‘i’]
- ‘xb’)*(1/‘Qy’[‘i’])
        }

        gen double ‘Rpp’ = .
        replace ‘Rpp’ = (1/$numobs)*normden(‘s’*‘y’[1] - ‘xb’)*(1 - ‘s’*‘y’[1]*(‘s’*‘y’[1]
- ‘xb’))*(1/‘Qy’[1]) if ‘r’ == 0
        forvalues i=2(1)$numy{
            replace ‘Rpp’ = ‘Rpp’ + (1/$numobs)*normden(‘s’*‘y’[‘i’] - ‘xb’)*(1
- ‘s’*‘y’[‘i’]*(‘s’*‘y’[‘i’] - ‘xb’))*(1/‘Qy’[‘i’])
        }
        gen double ‘bias’ = 1 - ‘R’

        // generate the 1st moment
gen double ‘g1i’ = ‘z’ if ‘r’ == 1
replace ‘g1i’ = - ((‘Rp’)/‘bias’) if ‘r’ == 0
quietly sum ‘g1i’
scalar ‘g1’ = r(mean)
// generate the 2nd moment
```

```
gen double ‘g2i’ = ‘z’*‘x1’ if ‘r’ == 1
replace ‘g2i’ = - ((‘Rp’*‘x1’)/‘bias’) if ‘r’ == 0
quietly sum ‘g2i’
scalar ‘g2’ = r(mean)
// generate the 3rd moment
    gen double ‘g3i’ = ‘z’*‘x2’ if ‘r’ == 1
    replace ‘g3i’ = - ((‘Rp’*‘x2’)/‘bias’) if ‘r’ == 0
    quietly sum ‘g3i’
    scalar ‘g3’ = r(mean)
    // generate the 4th moment
    gen double ‘g4i’ = (1/‘s’)-‘y’*‘z’ if ‘r’ == 1
    replace ‘g4i’ = - ((‘Rpp’)/‘bias’) if ‘r’ == 0
    quietly sum ‘g4i’
    scalar ‘g4’ = r(mean)
    gen double ‘th’ = ‘g1’ + 1 in 1
    replace ‘th’ = ‘g2’ in 2
    replace ‘th’ = ‘g3’ in 3
    replace ‘th’ = ‘g4’ in 4
    replace ‘t’ = ‘th’
end
```

PL estimator

```
clear
set more off
set mem 100m

*****************
** Main Program **
*****************

capture program drop simugmm
program simugmm, rclass
      version 9
            drop _all
            local N = _N
            global numobs 'N'
            ml model d0 pml1_d0 (mu: ystar = x1 x2)/lnsigma
            ml maximize, iterate(300)
end

*****************
** Likelihood **
*****************

capture program drop pml1_d0
program pml1_d0
      version 8.1
      args todo b lnf
      tempvar mu lnfj lnfj1 lnfj2 denssim
      tempname lnsigma b1 sigma n
      mleval 'mu' = 'b', eq(1)
      mleval 'lnsigma' = 'b', eq(2) scalar
      quietly{
            scalar 'sigma' = exp('lnsigma')
            gen double 'lnfj1' = ln(normden($ML_y1,'mu','sigma'))
            gen double 'denssim' = normden($ML_y1,'mu'[1],'sigma')
            forvalues i=2(1)$numobs{
                  quietly replace 'denssim' = 'denssim' + normden($ML_y1,'mu'['i'],'sigma')
            }
            gen double 'lnfj2' = ln('denssim'/$numobs)
            gen double 'lnfj' = 'lnfj1' - 'lnfj2'
            replace 'lnfj' = 0 if r == 0
            mlsum 'lnf' = 'lnfj'
      }
end
```

RS GMM Estimator

```
clear
set more off
set mem 100m

*******************
** Main Program **
*******************
capture program drop simugmm
program simugmm, rclass
      version 9
            drop _all
            // calculate the cuttoffs
            centile ystar if r == 1, centile(10 20 30 40 50 60 70 80 90)
            scalar c1 = r(c_1)
            scalar c2 = r(c_2)
            scalar c3 = r(c_3)
            scalar c4 = r(c_4)
            scalar c5 = r(c_5)
            scalar c6 = r(c_6)
            scalar c7 = r(c_7)
            scalar c8 = r(c_8)
            scalar c9 = r(c_9)
            pmcentile

            // GMM estimation
            sort lo, stable
            gen t = .
            replace t = 1 in 1
            replace t = 0 in 2
            replace t = 0 in 3
            replace t = 0 in 4
            replace t = 0 in 5
            replace t = 0 in 6
            replace t = 0 in 7
            replace t = 0 in 8
            replace t = 0 in 9
            replace t = 0 in 10
            replace t = 0 in 11
            replace t = 0 in 12
            replace t = 0 in 13
            replace t = 0 in 14
            replace t = 0 in 15
            replace t = 0 in 16
            replace t = 0 in 17
```

```
replace t = 0 in 18
replace t = 0 in 19
replace t = 0 in 20
replace t = 0 in 21
replace t = 0 in 22
replace t = 0 in 23

nl gmm_miss @ t lo up x1 x2 caty r, nparameters(23) iterate(1000)
mat b = e(b)
return scalar gmmbsQ1 = b[1,11]
return scalar gmmbsQ2 = b[1,12]
return scalar gmmbsQ3 = b[1,13]
return scalar gmmbsQ4 = b[1,14]
return scalar gmmbsQ5 = b[1,15]
return scalar gmmbsQ6 = b[1,16]
return scalar gmmbsQ7 = b[1,17]
return scalar gmmbsQ8 = b[1,18]
return scalar gmmbsQ9 = b[1,19]
return scalar gmmbss = 1/b[1,23]
return scalar gmmbs0 = b[1,20]*return(gmmbss)
return scalar gmmbs1 = b[1,21]*return(gmmbss)
return scalar gmmbs2 = b[1,22]*return(gmmbss)
return scalar rssbs = e(rss)
        return scalar convebs = e(converge)
        return scalar iterbs = e(ic)
end

*********************
** Discretization **
*********************

capture program drop pmcentile
program pmcentile
    tempname c1 c2 c3 c4 c5 c6 c7 c8 c9
    scalar 'c1' = scalar(c1)
    scalar 'c2' = scalar(c2)
    scalar 'c3' = scalar(c3)
    scalar 'c4' = scalar(c4)
    scalar 'c5' = scalar(c5)
    scalar 'c6' = scalar(c6)
    scalar 'c7' = scalar(c7)
    scalar 'c8' = scalar(c8)
    scalar 'c9' = scalar(c9)
    // gen the bounds
    gen double lo = .
    replace lo = 'c9' if ystar >= 'c9'
```

```
replace lo = 'c8' if (ystar >= 'c8' & ystar < 'c9')
replace lo = 'c7' if (ystar >= 'c7' & ystar < 'c8')
replace lo = 'c6' if (ystar >= 'c6' & ystar < 'c7')
replace lo = 'c5' if (ystar >= 'c5' & ystar < 'c6')
replace lo = 'c4' if (ystar >= 'c4' & ystar < 'c5')
replace lo = 'c3' if (ystar >= 'c3' & ystar < 'c4')
replace lo = 'c2' if (ystar >= 'c2' & ystar < 'c3')
replace lo = 'c1' if (ystar >= 'c1' & ystar < 'c2')
gen double up = .
replace up = 'c9' if (ystar >= 'c8' & ystar < 'c9')
replace up = 'c8' if (ystar >= 'c7' & ystar < 'c8')
replace up = 'c7' if (ystar >= 'c6' & ystar < 'c7')
replace up = 'c6' if (ystar >= 'c5' & ystar < 'c6')
replace up = 'c5' if (ystar >= 'c4' & ystar < 'c5')
replace up = 'c4' if (ystar >= 'c3' & ystar < 'c4')
replace up = 'c3' if (ystar >= 'c2' & ystar < 'c3')
replace up = 'c2' if (ystar >= 'c1' & ystar < 'c2')
replace up = 'c1' if ystar < 'c1'
gen double caty = .
replace caty = 1 if up == 'c1'
replace caty = 2 if lo == 'c1'
replace caty = 3 if lo == 'c2'
replace caty = 4 if lo == 'c3'
replace caty = 5 if lo == 'c4'
replace caty = 6 if lo == 'c5'
replace caty = 7 if lo == 'c6'
replace caty = 8 if lo == 'c7'
replace caty = 9 if lo == 'c8'
replace caty = 10 if lo == 'c9'
end

**********
** GMM **
**********

capture program drop nlgmm_miss
program nlgmm_miss
    version 9
    syntax varlist [if], at(name)
    local t : word 1 of 'varlist'
    local y1 : word 2 of 'varlist'
    local y2 : word 3 of 'varlist'
    local x1 : word 4 of 'varlist'
    local x2 : word 5 of 'varlist'
    local y : word 6 of 'varlist'
    local r : word 7 of 'varlist'
```

```
tempname H1 H2 H3 H4 H5 H6 H7 H8 H9 H10
tempname Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9
tempname b0 b1 b2 b3 b4 b5 b6 b7 b8 s
scalar ‘H1’ = ‘at’[1, 1]
scalar ‘H2’ = ‘at’[1, 2]
scalar ‘H3’ = ‘at’[1, 3]
scalar ‘H4’ = ‘at’[1, 4]
scalar ‘H5’ = ‘at’[1, 5]
scalar ‘H6’ = ‘at’[1, 6]
scalar ‘H7’ = ‘at’[1, 7]
scalar ‘H8’ = ‘at’[1, 8]
scalar ‘H9’ = ‘at’[1, 9]
scalar ‘H10’ = ‘at’[1, 10]
scalar ‘Q1’ = ‘at’[1, 11]
scalar ‘Q2’ = ‘at’[1, 12]
scalar ‘Q3’ = ‘at’[1, 13]
scalar ‘Q4’ = ‘at’[1, 14]
scalar ‘Q5’ = ‘at’[1, 15]
scalar ‘Q6’ = ‘at’[1, 16]
scalar ‘Q7’ = ‘at’[1, 17]
scalar ‘Q8’ = ‘at’[1, 18]
scalar ‘Q9’ = ‘at’[1, 19]
scalar ‘b0’ = ‘at’[1, 20]
scalar ‘b1’ = ‘at’[1, 21]
scalar ‘b2’ = ‘at’[1, 22]
scalar ‘s’ = ‘at’[1, 23]
tempvar g1i g2i g3i g4i g5i g6i g7i g8i g9i g10i g11i g12i g13i g14i g15i
g16i g17i g18i g19i g20i g21i g22i g23i
tempvar th xb Q1i Q2i Q3i Q4i Q5i Q6i Q7i Q8i Q9i Q10i z1 z2 phi
dphi ydphi
tempvar ct1 ct2 ct3 ct4 ct5 ct6 ct7 ct8 ct9 R Rp Rpp bias
tempname g1 g2 g3 g4 g5 g6 g7 g8 g9 g10 g11 g12 g13 g14 g15 g16 g17
g18 g19 g20 g21 g22 g23
tempname Q10 H21 H32 H43 H54 H65 H76 H87 H98 H109
tempname c1 c2 c3 c4 c5 c6 c7 c8 c9
scalar ‘c1’ = scalar(c1)
scalar ‘c2’ = scalar(c2)
scalar ‘c3’ = scalar(c3)
scalar ‘c4’ = scalar(c4)
scalar ‘c5’ = scalar(c5)
scalar ‘c6’ = scalar(c6)
scalar ‘c7’ = scalar(c7)
scalar ‘c8’ = scalar(c8)
scalar ‘c9’ = scalar(c9)
gen double ‘xb’ = ‘b0’ + (‘b1’*‘x1’) + (‘b2’*‘x2’)
gen double ‘z1’ = (‘s’*‘y1’-‘xb’)
```

```
gen double ‘z2’ = (‘s’*‘y2’-‘xb’)
gen double ‘ct1’ = (‘s’*‘c1’) - ‘xb’
gen double ‘ct2’ = (‘s’*‘c2’) - ‘xb’
gen double ‘ct3’ = (‘s’*‘c3’) - ‘xb’
gen double ‘ct4’ = (‘s’*‘c4’) - ‘xb’
gen double ‘ct5’ = (‘s’*‘c5’) - ‘xb’
gen double ‘ct6’ = (‘s’*‘c6’) - ‘xb’
gen double ‘ct7’ = (‘s’*‘c7’) - ‘xb’
gen double ‘ct8’ = (‘s’*‘c8’) - ‘xb’
gen double ‘ct9’ = (‘s’*‘c9’) - ‘xb’
gen double ‘Q1i’ = normprob(‘ct1’)
gen double ‘Q2i’ = normprob(‘ct2’) - normprob(‘ct1’) if ‘ct1’ <=0
replace ‘Q2i’ = normprob(-‘ct1’)- normprob(-‘ct2’) if ‘ct1’ > 0
gen double ‘Q3i’ = normprob(‘ct3’) - normprob(‘ct2’) if ‘ct2’ <=0
replace ‘Q3i’ = normprob(-‘ct2’)- normprob(-‘ct3’) if ‘ct2’ > 0
gen double ‘Q4i’ = normprob(‘ct4’) - normprob(‘ct3’) if ‘ct3’ <=0
replace ‘Q4i’ = normprob(-‘ct3’)- normprob(-‘ct4’) if ‘ct3’ > 0
gen double ‘Q5i’ = normprob(‘ct5’) - normprob(‘ct4’) if ‘ct4’ <=0
replace ‘Q5i’ = normprob(-‘ct4’)- normprob(-‘ct5’) if ‘ct4’ > 0
gen double ‘Q6i’ = normprob(‘ct6’) - normprob(‘ct5’) if ‘ct5’ <=0
replace ‘Q6i’ = normprob(-‘ct5’)- normprob(-‘ct6’) if ‘ct5’ > 0
gen double ‘Q7i’ = normprob(‘ct7’) - normprob(‘ct6’) if ‘ct6’ <=0
replace ‘Q7i’ = normprob(-‘ct6’)- normprob(-‘ct7’) if ‘ct6’ > 0
        gen double ‘Q8i’ = normprob(‘ct8’) - normprob(‘ct7’) if ‘ct7’ <=0
        replace ‘Q8i’ = normprob(-‘ct7’)- normprob(-‘ct8’) if ‘ct7’ > 0
        gen double ‘Q9i’ = normprob(‘ct9’) - normprob(‘ct8’) if ‘ct8’ <=0
        replace ‘Q9i’ = normprob(-‘ct8’)- normprob(-‘ct9’) if ‘ct8’ > 0
        gen double ‘Q10i’ = normprob(-‘ct9’)
        scalar ‘Q10’ = 1 - (‘Q1’ + ‘Q2’ + ‘Q3’ + ‘Q4’ + ‘Q5’ + ‘Q6’ + ‘Q7’ +
‘Q8’ + ‘Q9’)
        scalar ‘H21’ = (‘H2’/‘Q2’)-(‘H1’/‘Q1’)
        scalar ‘H32’ = (‘H3’/‘Q3’)-(‘H2’/‘Q2’)
        scalar ‘H43’ = (‘H4’/‘Q4’)-(‘H3’/‘Q3’)
        scalar ‘H54’ = (‘H5’/‘Q5’)-(‘H4’/‘Q4’)
        scalar ‘H65’ = (‘H6’/‘Q6’)-(‘H5’/‘Q5’)
        scalar ‘H76’ = (‘H7’/‘Q7’)-(‘H6’/‘Q6’)
        scalar ‘H87’ = (‘H8’/‘Q8’)-(‘H7’/‘Q7’)
        scalar ‘H98’ = (‘H9’/‘Q9’)-(‘H8’/‘Q8’)
        scalar ‘H109’ = (‘H10’/‘Q10’)-(‘H9’/‘Q9’)
        gen double ‘R’ = normprob(‘ct1’)*‘H21’ + normprob(‘ct2’)*‘H32’ +
normprob(‘ct3’)*‘H43’+ normprob(‘ct4’)*‘H54’+ normprob(‘ct5’)*‘H65’ + norm-
prob(‘ct6’)*‘H76’ + normprob(‘ct7’)*‘H87’ + normprob(‘ct8’)*‘H98’+ norm-
prob(‘ct9’)*‘H109’
        gen double ‘Rp’ = normd(‘ct1’)*‘H21’ + normd(‘ct2’)*‘H32’ + normd(‘ct3’)*‘H43’+
normd(‘ct4’)*‘H54’+ normd(‘ct5’)*‘H65’ + normd(‘ct6’)*‘H76’ + normd(‘ct7’)*‘H87’
+ normd(‘ct8’)*‘H98’+ normd(‘ct9’)*‘H109’
```

```
        gen double 'Rpp' = 'c1'*normd('ct1')*'H21' + 'c2'*normd('ct2')*'H32'
+ 'c3'*normd('ct3')*'H43' + 'c4'*normd('ct4')*'H54' + 'c5'*normd('ct5')*'H65'
+ 'c6'*normd('ct6')*'H76' + 'c7'*normd('ct7')*'H87' + 'c8'*normd('ct8')*'H98'
+ 'c9'*normd('ct9')*'H109'
        gen double 'bias' = 1-(('H10'/'Q10')-'R')

        gen double 'phi' = cond('y1'!=.&'y2'!=., cond('z1'>0,normprob(-'z1')-
normprob(-'z2'), normprob('z2') -normprob('z1')), cond('y2'!=.,normprob('z2'),normprob(-
'z1')))
        gen double 'dphi' = ( cond('y2'!=., normd('z2'), 0) - cond('y1'!=.,
normd('z1'), 0) )
        gen double 'ydphi'=( cond('y2'!=., 'y2'*normd('z2'), 0) - cond('y1'!=.,
'y1'*normd('z1'), 0) )

        // generate the 1st moment
        gen double 'g1i' = cond('y'==1,1,0) if 'r' == 1
        replace 'g1i' = - ('H1'*'Q1i')/('Q1'*'bias') if 'r' == 0
        quietly sum 'g1i'
        scalar 'g1' = r(mean)
        // generate the 2nd moment
        gen double 'g2i' = cond('y'==2,1,0) if 'r' == 1
        replace 'g2i' = - ('H2'*'Q2i')/('Q2'*'bias') if 'r' == 0
        quietly sum 'g2i'
        scalar 'g2' = r(mean)
        // generate the 3rd moment
        gen double 'g3i' = cond('y'==3,1,0) if 'r' == 1
        replace 'g3i' = - ('H3'*'Q3i')/('Q3'*'bias') if 'r' == 0
        quietly sum 'g3i'
        scalar 'g3' = r(mean)
        // generate the 4th moment
        gen double 'g4i' = cond('y'==4,1,0) if 'r' == 1
        replace 'g4i' = - ('H4'*'Q4i')/('Q4'*'bias') if 'r' == 0
        quietly sum 'g4i'
        scalar 'g4' = r(mean)
        // generate the 5th moment
        gen double 'g5i' = cond('y'==5,1,0) if 'r' == 1
        replace 'g5i' = - ('H5'*'Q5i')/('Q5'*'bias') if 'r' == 0
        quietly sum 'g5i'
        scalar 'g5' = r(mean)
        // generate the 6th moment
        gen double 'g6i' = cond('y'==6,1,0) if 'r' == 1
        replace 'g6i' = - ('H6'*'Q6i')/('Q6'*'bias') if 'r' == 0
        quietly sum 'g6i'
        scalar 'g6' = r(mean) .
        // generate the 7th moment
        gen double 'g7i' = cond('y'==7,1,0) if 'r' == 1
```

```
replace ‘g7i’ = - (‘H7’*‘Q7i’)/(‘Q7’*‘bias’) if ‘r’ == 0
quietly sum ‘g7i’
scalar ‘g7’ = r(mean)
// generate the 8th moment
gen double ‘g8i’ = cond(‘y’==8,1,0) if ‘r’ == 1
replace ‘g8i’ = - (‘H8’*‘Q8i’)/(‘Q8’*‘bias’) if ‘r’ == 0
quietly sum ‘g8i’
scalar ‘g8’ = r(mean)
// generate the 9th moment
gen double ‘g9i’ = cond(‘y’==9,1,0) if ‘r’ == 1
replace ‘g9i’ = - (‘H9’*‘Q9i’)/(‘Q9’*‘bias’) if ‘r’ == 0
quietly sum ‘g9i’
scalar ‘g9’ = r(mean)
// generate the 10th moment
gen double ‘g10i’ = cond(‘y’==10,1,0) if ‘r’ == 1
replace ‘g10i’ = - (‘H10’*‘Q10i’)/(‘Q10’*‘bias’) if ‘r’ == 0
quietly sum ‘g10i’
scalar ‘g10’ = r(mean)
// generate the 11th moment
gen double ‘g11i’ = ‘Q1’ - ‘Q1i’
quietly sum ‘g11i’
scalar ‘g11’ = r(mean)
// generate the 12th moment
gen double ‘g12i’ = ‘Q2’ - ‘Q2i’
quietly sum ‘g12i’
scalar ‘g12’ = r(mean)
// generate the 13th moment
gen double ‘g13i’ = ‘Q3’ - ‘Q3i’
quietly sum ‘g13i’
scalar ‘g13’ = r(mean)
// generate the 14th moment
gen double ‘g14i’ = ‘Q4’ - ‘Q4i’
quietly sum ‘g14i’
scalar ‘g14’ = r(mean)
// generate the 15th moment
gen double ‘g15i’ = ‘Q5’ - ‘Q5i’
quietly sum ‘g15i’
scalar ‘g15’ = r(mean)
// generate the 16th moment
gen double ‘g16i’ = ‘Q6’ - ‘Q6i’
quietly sum ‘g16i’
scalar ‘g16’ = r(mean)
// generate the 17th moment
gen double ‘g17i’ = ‘Q7’ - ‘Q7i’
quietly sum ‘g17i’
scalar ‘g17’ = r(mean)
```

```
// generate the 18th moment
gen double ‘g18i’ = ‘Q8’ - ‘Q8i’
quietly sum ‘g18i’
scalar ‘g18’ = r(mean)
// generate the 19th moment
gen double ‘g19i’ = ‘Q9’ - ‘Q9i’
quietly sum ‘g19i’
scalar ‘g19’ = r(mean)
// generate the 20th moment
gen double ‘g20i’ = -‘dphi’/‘phi’ if ‘r’ == 1
replace ‘g20i’ = - (‘Rp’/‘bias’) if ‘r’ == 0
quietly sum ‘g20i’
scalar ‘g20’ = r(mean)
// generate the 21st moment
gen double ‘g21i’ = (‘x1’*(-‘dphi’))/‘phi’ if ‘r’ == 1
replace ‘g21i’ = - ((‘Rp’*‘x1’)/‘bias’) if ‘r’ == 0
quietly sum ‘g21i’
scalar ‘g21’ = r(mean)
// generate the 22th moment
gen double ‘g22i’ = (‘x2’*(-‘dphi’))/‘phi’ if ‘r’ == 1
replace ‘g22i’ = - ((‘Rp’*‘x2’)/‘bias’) if ‘r’ == 0
quietly sum ‘g22i’
scalar ‘g22’ = r(mean)
// generate the 23th moment
gen double ‘g23i’ = ‘ydphi’/‘phi’ if ‘r’ == 1
replace ‘g23i’ = -(- ‘Rpp’)/‘bias’ if ‘r’ == 0
quietly sum ‘g23i’
scalar ‘g23’ = r(mean)
gen double ‘th’ = ‘g1’ + 1 in 1
replace ‘th’ = ‘g2’ in 2
replace ‘th’ = ‘g3’ in 3
replace ‘th’ = ‘g4’ in 4
replace ‘th’ = ‘g5’ in 5
replace ‘th’ = ‘g6’ in 6
replace ‘th’ = ‘g7’ in 7
replace ‘th’ = ‘g8’ in 8
replace ‘th’ = ‘g9’ in 9
replace ‘th’ = ‘g10’ in 10
replace ‘th’ = ‘g11’ in 11
replace ‘th’ = ‘g12’ in 12
replace ‘th’ = ‘g13’ in 13
replace ‘th’ = ‘g14’ in 14
replace ‘th’ = ‘g15’ in 15
replace ‘th’ = ‘g16’ in 16
replace ‘th’ = ‘g17’ in 17
replace ‘th’ = ‘g18’ in 18
```

```
        replace 'th' = 'g19' in 19
        replace 'th' = 'g20' in 20
        replace 'th' = 'g21' in 21
        replace 'th' = 'g22' in 22
        replace 'th' = 'g23' in 23
            replace 't' = 'th'
    end
```

Chapter 4
A Comparative Analysis of Wage Rates in the LFS

4.1 Introduction

This chapter focuses on comparing the estimators employed in the Monte Carlo experiments of Chapter 3 using real data. The estimators are applied to analyse the UK wage distribution. The approach of this empirical exercise is based on Skinner et al (2002) which develop a method of estimating the distribution of the Labour Force Survey (LFS) hourly wage rate variable for the Office for National Statistics (ONS).

Since 1992, the LFS has been a quarterly survey. In each quarter, it draws approximately 12,000 households as the first wave sample from the Postcode Address File (PAF) using a stratified sampling. All adults in the selected households are interviewed and retained for five successive quarters. Thus, in any given quarter, the LFS is comprised of five sample waves with a total sample size of around 60,000 households (Werner, 2006). The questions involving the measures of the wage rate are only asked in the first and fifth waves. Thus, only a portion of the individuals in the LFS provides information on wage rates in each quarter.

There are two measures of the wage rate in the LFS. One measure derives an estimate of the wage rate indirectly from several related questions on earnings and working hours and it is referred to as the *derived variable* by Skinner et al. Prior to March 1999, it is

the only available measure of the wage rate. The other measure, referred to as the *direct variable*, is introduced in March 1999. To obtain this variable, each individual is first asked whether he or she is paid a fixed hourly rate. Those who answer "yes" to this question are then asked to provide their (basic) hourly rate. This rate is subsequently recorded as the direct variable.

Skinner et al study the sources of measurement error in the LFS's measures of the wage rate and conclude that the derived variable suffers from these sources more than the direct variable. They also point out that similar findings have been reported for the CPS and PSID. Therefore, it is preferable to use the direct variable, rather than the derived variable, in a study of the wage rate and its distribution.

However, there are two sampling issues which must be addressed prior to use of the direct variable. The first issue is that the LFS is not a random sample, but is a stratified sample. It is clustered by households and there are some unit nonresponses. The second issue is that the direct variable is not observed for a large portion of individuals in the LFS. It is also likely that the subset of individuals who have reported the direct variable is a nonrandom subsample. These two issues must be resolved to ensure the validity of any subsequent inference.

Evidence that the subsample of interest is nonrandom is shown in Table 4.1, which is presented in the appendix to this chapter along with other related tables. It reports the summary statistics of the derived variable for all individuals and for the subsample for whom the direct variable is observed. One can see that the distribution of the derived variable in the full sample is different to that in the subsample. Individuals in the subsample

are, overall, paid at a lower wage rate. For example, only 5% of people in the subsample earn higher than £10.54 per hour whereas 25% of employees in the full sample are paid more than £9.97 per hour. This is because jobs which pay a fixed hourly rate are typically lower paid ones. Thus, the table indicates that individuals are, at least, selected into the subsample in accordance with their wage rate.

The first sampling issue is usually dealt with by using sampling weights provided with the LFS. We show below that this issue seems to cause no problem in this study and can therefore be left on one side. It is the second sampling issue where our interest is focused since it can be handled in many ways depending on assumptions one is willing to impose on the missing data mechanism (MDM). Skinner et al use an imputation method, which asserts that the MDM is MAR, to address this issue. Beissel-Durrant and Skinner (2003) extend further the imputation technique in Skinner et al. In this chapter, we use the techniques elaborated in Chapter 3 to deal with this sampling issue. The main objective is to study, in an empirical context, the sensitivity of the outcomes of interest with respect to different assumptions on the MDM and estimation procedures.

In what follows, we are interested in estimating (i) coefficient parameters of the linear conditional mean function of the wage rate and (ii) a point in the distribution of the wage rate: the proportion below the National Minimum Wage (NMW). The former is discussed in the next section while the latter is examined in Section 4.3. All estimation procedures from Chapter 3 are used in the discussion. In addition to these procedures, an imputation technique called Multiple Imputation by Chained Equations (MICE) is also described and employed in the investigation. Then, the conclusion is put forward in the last section.

4.2 Estimation of the coefficient parameters of the conditional mean function

Skinner et al use the June-August 1999 quarter of the LFS because its interest is on the effect of the NMW, which is introduced in April 1999, on the bottom of the UK wage distribution. We use the same data set for comparison purposes. The linear model,

$$y = x'\beta + \varepsilon, \tag{4.101}$$

where β is a vector of unknown parameters, will be fitted to the data using different estimation techniques.

Although the model specification adopted is also based on that of Skinner et al, fewer explanatory variables are used for simplicity. The log of the direct variable is taken as Y or the dependent variable. The log of the derived variable is considered as the dependent variable plus measurement error and is denoted as Z. It will be included as an extra covariate in some estimation procedures. In accordance with Stuttard and Jenkins (2001), the values of the direct and derived variables reported by non-spouse proxy respondents are adjusted to reduce a systematic measurement error. The main set of explanatory variables in X are years of education, a quadratic in experience and dummy variables for female respondents, for marital status, for workplaces with more than 25 employees, for full- or part-time status and for residences in London or South East England.

Below, the parameter vector β is firstly estimated by the unweighted Least Squares estimator. Next, the estimations and discussions are based, in turn, on the Inverse Probability Weighted Least Squares, Sample Selection model, RS GMM, INRY-GMM, Pseudo-likelihood and Multiple Imputation estimators. All resultant estimates are then compared

to see the effect of changes in assumptions imposed on the MDM and in estimation procedures.

4.2.1 Unweighted Least Squares Estimators

Table 4.2 shows results from a number of applications of the unweighted Least Squares (ULS) estimator on individuals whose values of all variables are recorded or, for brevity, on *complete cases*. The first column of Table 4.2 is estimates from the ULS regression without any kind of adjustment, which is referred to as the ULS1 estimator. Based on these estimates, we identify six influential observations such as outliers by plotting the residuals from the ULS1 estimator against its fitted values and plotting the absolute values of standardized residuals against the leverage values. Estimates from the ULS estimator after deleting these six influential observations, or the ULS2 estimator, are presented in the second column of Table 4.2. It is clear from comparing the results of the ULS1 and ULS2 estimators that deleting these observations has almost no effect on the coefficient estimates. This implies that the results are robust against the influential observations and, as a result, we will not exclude these sampled units from our investigation.

Furthermore, we use the sampling weights given with the LFS to weight the ULS estimator. There are two types of weights supplied, pwt03 and piwt03. Even though both types of weights are for dealing with sampling design and unit non-responses, i.e., the first sampling issue, they are constructed with two different concerns. The first type of weights is for individual. It compensates for non-response and grosses to population estimates. The

second type of weights is for income data. It operates such that the weight of a subgroup corresponds to that subgroup's size in the population (Crockett, 2007).

The third and fourth columns of Table 4.2 illustrate results from using piwt03 and pwt03 to weight the ULS estimator. The corresponding estimators are labelled as the ULS3 and ULS4 estimators respectively. Relative to the ULS1 estimator, the use of both types of sampling weights has only minimal impact on the estimates. As noted above, this means that the first sampling issue does not cause any serious problem in terms of inference in this context. Thus, we do not use these weights in the following analyses and treat the LFS as if it is a proper random sample from the population.

4.2.2 Inverse Probability Weighted Least Squares Estimators

Table 4.3 reports results from various Inverse Probability Weighted Least Squares (IPWLS) estimations whose Inverse Probability (IP) weights are computed using different probit models. Consider first the results of the Probit-IPWLS1 estimator in the second column where the covariates in X are used in both the probit and IPWLS regression models. By comparing the Probit-IPWLS1 estimator to the ULS3 and ULS4 estimators, it is apparent that the impact of IP weights on the resultant estimates is both sizeable and dissimilar to those of pwt03 and piwt03. This is expected because they are for different sampling issues, that is, IP weights should deal with the item-nonresponses whereas pwt03 and piwt03 should handle the stratified sampling scheme and unit-nonresponses.

Even though estimates of the Probit-IPWLS1 estimator have the same sign as those of the ULS1 estimator, the coefficient estimate of married is no longer significant[9]. To explain this circumstance, observe the sizes of the four largest IP weights which are also reported in Table 4.3. For the Probit-IPWLS1 estimator, the two largest IP weights are markedly bigger than the other two. Also, the marital status of the two observations, to which these two IP weights are assigned, happens to be single. It is therefore possible that overweighting these observations has contributed to the married variable becoming insignificant.

This is indeed the case as indicated by results of the Probit-IPWLS2 estimator. In this case, the two problematic observations are dropped before the estimation procedures are applied. As a result, all coefficient estimates of the Probit-IPWLS2 estimator are significant at 5%.

The source of this overweighting problem is not from the inclusion of these two observations but is from the specification of the probit model assigning weights. We demonstrate this point by using the log of the derived variable, or Z, as an extra explanatory variable in the first-step probit estimation of the Probit-IPWLS3 estimator. Without dropping any observation, one can see that all coefficient estimates of the Probit-IPWLS3 estimator are significant at 5%. Also the four largest IP weights in this case are smaller than those of the Probit-IPWLS1 estimator where Z is not employed in the probit estimation.

---

[9] We also attempt to weight the first-step probit model with pwt03 (piwt03) and to weight the IPWLS regression model with the resultant IP weights multiplied by pwt03 (piwt03). However, these variations do not change the inference, i.e., the dummy variable for married is still the only insignificant variable.

It is more appropriate for the specification of the probit model to include the derived variable. Note that a main assumption underpinning the IPWLS estimation is MAR. For the estimation of the Probit-IPWLS1 and Probit-IPWLS2 estimators, it is assumed that MAR holds conditioning only on the explanatory variables in X, that is

$$\mathcal{P}\{R = 1|Y = y, X = x\} = \mathcal{P}\{R = 1|X = x\}. \tag{4.102}$$

On the other hand, the Probit-IPWLS3 estimator assumes a different MAR assumption, which is

$$\mathcal{P}\{R = 1|Y = y, X = x, Z = z\} = \mathcal{P}\{R = 1|X = x, Z = z\}. \tag{4.103}$$

Since Table 4.1 indicates that the MDM is strongly related to the dependent variable, including Z which can be thought of as the log of the direct variable plus a measurement error should provide more information and should make (4.103) more plausible than (4.102)[10]. In addition, this circumstance is similar to that of Experiment 16 in Chapter 3. It is noted there that, if the MDM depends on Y, using variables such as Z in the selection equation can (i) allow for correlation between R and Y under assumption MAR and (ii) improve the performance of the IPWLS estimator.

The percent correctly predicted is shown in Table 4.3 as a measure of goodness of fit of a binary choice model. It suggests that the probit model with Z fits the data better than the one without it. Nevertheless, note that both probit models do not pass the LM test for normality[11].

---

[10] Adding Z should not cause any estimation problem as long as the measurement error is independent of the disturbance term of the selection equation.

[11] The critical values of a chi-squared distribution, with degree of freedom equals to 2, are 4.61(10%), 5.99(5%) and 9.21(1%).

Other distributional specifications of binary choice model are then explored in Tables 4.4, 4.5 and 4.6. For each distributional specification, we consider both selection models with and without the extra covariate Z. In Table 4.4, the normal distribution is replaced by the logit and complementary log-log distributions[12]. The IPWLS estimations based on logit specification are called the Logit-IPWLS1 and Logit-IPWLS2 estimators while those with the complementary log-log specification are labelled Cloglog-IPWLS1 and Cloglog-IPWLS2 estimators. The log of the derived variable is used only in the first-step binary response model estimation of the Logit-IPWLS2 and Cloglog-IPWLS2 estimators.

Although the logit distribution is symmetric, it is chosen because it has fatter tails than the normal distribution. On the other hand, the complementary log-log distribution is asymmetric and is skewed to the right. The four largest IP weights from all logit and complementary log-log models in Table 4.4 are smaller than those from the probit models in Table 4.3. Even though the coefficient of married in the Logit-IPWLS1 estimator is insignificant at 5% as in the case of the Probit-IPWLS1 estimator, it is significant at 10% which may due to smaller IP weights assigned. However, all coefficient estimates of the Logit-IPWLS2 estimator are significant even at 5%. On the basis of the percent correctly predicted, the logit model with Z of the Logit-IPWLS2 estimator fits with the data better than other binary choice models in Table 4.4 and as well as the probit model with Z in Table 4.3.

The scobit model is examined in Table 4.5 as another alternative specification. In the table, the Scobit-IPWLS1 estimator is the IPWLS whose IP weights are estimated by

---

[12] The index function under the complementary log-log distribution is $F(x'\beta) = 1 - \exp\left(-\exp(x'\beta)\right)$.

the scobit model without Z. Similarly, Scobit-IPWLS2 estimator is the IPWLS estimator whose corresponding scobit model includes Z.

The scobit distribution is asymmetric and is skewed to the left. The index function of the binary response model in this case is

$$F(x'\beta) = \frac{1}{(1+\exp(x'\beta))^{\alpha}},$$

where α is a parameter to be estimated and is greater than zero. It is of interest because the logit distribution is its special case whenever α is unit. We also reproduce the results of Logit-IPWLS1 and Logit-IPWLS2 estimators in Table 4.5 for comparison purposes.

Since the logit model is nested in the scobit model, one can conduct the LR test between the two models. Two such tests are reported in Table 4.5 and the null hypothesis that $\alpha = 1$ is rejected whether or not Z is used. The percent correctly predicted also suggests that the scobit model with Z is better than the one without it. Thus, one should choose the Scobit-IPWLS2 estimator over other IPWLS estimators in Tables 4.4 and 4.5. The results of this estimator assert interestingly that not only married but also size25 are insignificant after corrected for missing data.

All of the specifications for the binary response model so far assume that there is no heteroskedasticity problem in the selection equation. In table 4.6, we analyse various variations of the heteroskedastic probit (or hetprob) model which allows the variance to be a function of fully observed variables. This model extends the probit model by modelling explicitly the variance of the disturbance term as a squared exponential function whose argument is a linear function of a set of variables. As a consequence, we can also use the LR test to test this model against the probit model.

Table 4.6 shows results from using different sets of variables in both selection and variance equations. These specifications are chosen to see especially the effect of allowing the variance to vary with Z. The Hetprob-IPWLS1 estimator, which is the benchmark setting, employs the covariates in X for both selection and variance equations. The setting of the Hetprob-IPWLS2 estimator is similar to that of the Hetprob-IPWLS1 but it also includes Z as an additional covariate for both equations. The Hetprob-IPWLS3 estimator uses both X and Z in its variance equation but uses only X in the selection equation. Lastly, the Hetprob-IPWLS4 estimator uses X in the selection equation and uses only Z in its variance equation.

All of the LR test statistics presented in Table 4.6 indicate that the probit specification is rejected. The best-fitting hetprob model in this table, in accordance with the percent correctly predicted, is the one associated with the Hetprob-IPWLS2 estimator. The results for this estimator again suggest that the coefficients on both married and size25 are not significant.

It is clear from Tables 4.3 to 4.6 that IPWLS estimates are quite sensitive to a change in the specification of the selection model. The term specification here includes both the distributional assumption imposed on the disturbance of the selection equation and the conditioning variables used in this equation to verify MAR assumption.

The two best-fitting binary choice models on the basis of different test statistics and of the percent correctly predicted are those associated with the Scobit-IPWLS2 and Hetprob-IPWLS2 estimators. All slope coefficient estimates from both models display the same sign to those from the ULS1 estimator. Although these estimators are based on different

distributional assumptions, the coefficient estimates of married and size25 from both of them are not significant. This inference is intriguing and is quite a departure from that implied by the ULS1 estimator.

A possible extension to the study of IPWLS estimation presented in this section would be to use a semi- or nonparametric methods to estimate IP weights. An interesting choice of such methods would be an estimator for distribution free heteroskedastic binary response model developed by Khan (2006). The virtue of this estimator is that, while heteroskedasticity is allowed and no distributional specification is assumed, it is also possible to jointly estimate the regression coefficients and the choice probabilities. Hence, by implementing this estimation procedure, one can obtain IP weights to use in the IPWLS estimation under very weak assumptions.

4.2.3 Sample Selection Model Estimators

In this section, the LFS data is analysed by the Sample Selection (SS) model estimator using both two-step and partial maximum likelihood estimation. As before, the former is generally referred to as the SSTS estimator and the latter is called the SSML estimator.

One of the most common applications of the SS model is the estimation of a wage offer equation for people of working age. In that context, standard estimation techniques such as ULS estimator are inappropriate because, although the target population is all individuals of working age, we can only observe wage offers for those who choose to participate in the labour market. The observed sample is thus a nonrandom subsample because people are self-selected into the labour force. However, the target population of our study

is all individuals who are participating in the labour force at the time of the survey, i.e., the observed subsample in the aforementioned context of wage offer equation. The need for the SS model arises in our context because values of the direct variable are missing in a large proportion of our target population and we suspect that the missing data or selection mechanism is endogenous, namely,

$$\mathcal{P}\{R = 1|Y = y, X = x\} \neq \mathcal{P}\{R = 1|X = x\} \tag{4.104}$$

The second and third columns of Table 4.7 present results from the SSML1 and SSTS1 estimators where the covariates in X are used in both structural and selection equations, that is, there is no exclusion restriction. It can be seen from the table that if a variable is statistically significant in both models, its estimate will have the same sign and will be of similar magnitude. These signs also conform to those from the ULS1 estimator. Moreover, both models agree that the dummy variables for being married and for residing in London or South East England are insignificant.

However, both models disagree on the effects of years of education and of having part-time job as the main job. The SSTS1's estimate of years of education is insignificant with positive sign whereas that of SSML1 estimator is statistically significant with negative sign. These results are unreasonable since, logically, one would expect the effect of time spent on education to be both significant and positive as in the case of the ULS1 estimator.

While the coefficient estimate of the dummy variable for part-time job from the SSTS1 estimator is not statistically significant, the SSML1's estimate is significant. Both estimates also have positive sign which is implausible since, on average, individuals with

part-time jobs as their main jobs should earn less than individuals who have secured full-time jobs.

As pointed out by the Monte Carlo investigation in Chapter 3, these different and irrational results from both estimators may be a consequence of either having no exclusion restriction or low explanatory power of the covariates in the selection equation. There, we have learnt that, in spite of all underlying assumptions being correct, the SS model does not perform well if there is no exclusion restriction or the MDM is not explained well by the variation in the covariates. However, the pseudo R-squared, pR^2, of the probit model associated with the SSTS1 estimator is 0.1107 which is relatively high[13]. Thus, the poor results should be a result of having no exclusion restriction.

Estimates produced by the SSML2 and SSTS2 estimators, which use Z as an exclusion restriction, are reported in the fourth and fifth columns of Table 4.7. The pR^2 of the probit model associated with the SSTS2 estimator is now 0.1462. Although the derived variable suffers from the measurement error problem, it is still a measure of the wage rate and, consequently, should be related to the disturbances of both equations. Thus, one ought not to use Z as an exclusion restriction because of the endogeneity problem. However, we use it to see the sensitivity of the estimates from both estimators with respect to the incorporation of Z. One should therefore examine these results cautiously.

Although the changes in the results of both estimators due to the inclusion of the derived variable are obvious, these are more apparent in the estimates of the SSTS2 estimator.

---

[13] For example, pR^2 of Experiments 6, 17 and 18 in Chapter 3 are 0.1295, 0.0843 and 0.1612, which are considered as high. At these levels of pR^2, the SS model estimators show small mean bias in the simulation studies.

The disagreement of both estimators is also more evident. For instance, although the coefficients of years of education are now negative in both estimators, the estimate from the SSML2 estimator is not statistically different from zero. The SSTS2 estimator also infers that the quadratic in experience is insignificant which is contradictory to the results from the SSML2 estimator. The dummy variables for London and South East England and for part-time job become significant in the SSTS2 estimator. However, both dummy variables remain significant in the SSML2 estimator.

As expected, the prediction from both estimators has become even more implausible when Z is included as an exclusion restriction. It is also clear that results from the two-step estimation are extremely sensitive to the inclusion. Nevertheless, there are some results which remain the same through all deliberated changes in estimation method and in specification[14]. One of such results is that the marginal effect of married on wage rate is not significant.

Another unvarying result is that the subsample of interest is indeed a selected sample. For the two SSTS estimators, this result is implied through the significance of the coefficient of the inverse Mills ratio. For the SSML estimators, both LR tests reject the joint likelihood of an independent probit model and a normal regression model for the likelihood of the sample selection model. The correlation coefficients of the two disturbances, i.e., rho are also shown to be very high.

---

[14] The sampling weights, pwt03 and piwt03, are also employed to weight both SSML estimators. These results are however not significantly different from those shown in Table 7. Thus we have decided not to show them in the table.

To sum up, using these sample selection models as a base for inference will generally lead us to very different conclusions from those implied by the ULS and IPWLS estimators. This is likely to be a consequence of having no valid exclusion restriction as suggested by the Monte Carlo investigation in Chapter 3. Another possibility is that the normality assumption does not hold in this empirical context. An obvious extension to the study would therefore be to use the semiparametric sample selection model to analyse the data set instead.

4.2.4 GMM and Pseudolikelihood Estimators

This section examines the empirical application of the RS GMM, one-step INRY-GMM, two-step INRY-GMM and Pseudolikelihood (PL) estimators used in the Monte Carlo experiments. As before, they are denoted as DGMM, CGMM, CGMM1 and PL estimators respectively. All of these estimators impose on the MDM that,

$$\mathcal{P}\{R = 1|Y = y, X = x\} = \mathcal{P}\{R = 1|Y = y\}, \tag{4.105}$$

which is a special case of Assumption NMAR of the SS model estimators in (4.104).

Table 4.8 shows results from the DGMM and interval regression (INTREG) estimators with 13 and 19 groups. These four estimators will be referred to as the 13DGMM, 19DGMM, 13INTREG and 19INTREG estimators, respectively. Note that the INTREG estimators use data from the complete cases only and are shown here because the DGMM estimators can be regarded as correcting the missing-data bias in the INTREG estimators. The numbers of groups for these estimators are chosen to ensure that the loss of informa-

tion from the discretisation is minimal and that each group has roughly equal number of observations. However, the first group is deliberately selected so that it consists of values below the NMW. Thus, the estimate of population share of the first group, or Q_1, is also the estimate of the proportion of wage rates below the NMW.

Each coefficient estimate of all estimators in Table 4.8 is significant and is of the same sign as that from the ULS1 estimator. The change in number of groups, namely, from 13 to 19 groups, has negligible effect on the resultant estimates. This means that the loss in information from the discretisation is indeed small. Also, estimates from the two DGMM estimators are clearly different from those of the 13INTREG and 19INTREG estimators, implying that the bias-correcting scheme in the RS GMM estimation has a significant effect. Moreover, both 13DGMM and 19DGMM estimators yield the same estimate of the proportion of wage rate below the NMW, which is 0.05.

Results from the CGMM, CGMM1 and PL estimator are reported in Table 4.9. The results of the 19DGMM estimator are also reproduced in the table for comparison purposes. Since the CGMM estimator is a two-step estimator, its standard errors reported in the table are computed by the bootstrap procedures to take account of the first-step estimation.

With one exception, all estimates from these three estimators are of the same sign as the ULS1 estimator. Such estimates are also significant at 5% and are of similar magnitude to those of the 19DGMM estimator; the latter is anticipated as all of these estimators impose (4.105) on the MDM. The only exception is the coefficient estimate of the dummy variable for female from the CGMM and CGMM1 estimators. Although this particular coefficient estimate of the CGMM estimator is not significant, it is almost identical to that

from the 19DGMM estimator. The same estimate from the CGMM1 estimator however is insignificant, of the opposite sign and relatively closer to zero. This indicates that the choice of estimation procedure used has an important impact on the outcomes.

It should also be noted that the estimates from all GMM and PL estimators in this section are somewhat sensitive to the set of initial values used to start the maximisation process. This suggests that the objective function may be flat near the true values. In a future study, one could try to solve this problem by using an alternative algorithm such as a genetic algorithm.

Even though not every estimate from these GMM and PL estimators is significant, their results are more stable and reasonable than those from the SS model estimators. Thus, a main virtue of the estimators in this section is that, without requiring any exclusion restriction, they can give fairly sensible estimates while allowing the MDM to depend explicitly on Y [15].

4.2.5 Multiple Imputation

Skinner et al (2002) and Beissel-Durrant and Skinner (2003) use an imputation technique called donor imputation to deal with the missing data in the direct variable. This technique is attractive in this context because it can recreate the spike at the NMW, which is observed in the direct variable, in the imputed wage rate data. There are many ways of choosing a donor for a nonrespondent unit from all respondent units. Those which are examined by

---

[15] It must be noted that the MDM of the SS model is more generic than that of the DGMM, CGMM, PL estimators because it allows both Y and X to determine the probability of being missing.

Skinner et al and Beissel-Durrant and Skinner are fractional imputation, nearest neighbour imputation and predictive mean matching.

However, we choose to present another imputation technique in this study. This method is referred to as Multiple Imputation by Chained Equations or MICE in the imputation literature; see, for example, van Buuren, Boshuizen and Knook (1999). It is implemented in Stata by a user-written command called ICE which is developed in Royston (2004), Royston (2005a) and Royston (2005b). The availability of ICE is convenient and is a main reason why we prefer to use this imputation technique.

ICE creates multiple imputed data sets from a data set provided. The number of these imputed data sets can be specified prior to the start of the estimation procedure. In each imputed data set, ICE starts the process of imputation by filling in missing values of incomplete variables with randomly selected observed values. Then, for each incomplete variable in turn, the filled-in values are replaced by imputed values calculated from the current "completed" version of the data set. A cycle is completed whenever this process is repeated for all incomplete variables and ten cycles are usually required for an imputed data set.

The calculation of the imputed values in ICE involves (i) regressing the incomplete variable under consideration on all other "completed" variables, (ii) drawing values of parameters from posterior distribution based on the aforementioned imputation regression and (iii) computing the imputed values from the drawn values of parameters. The posterior distribution in the second step is assumed to be multivariate normal. This can be relaxed using a bootstrap estimation. Moreover, ICE allows the imputed values in the third step

to be calculated by predictive mean matching which is also used in the donor imputation of Skinner et al (2002). This similarity to the donor imputation is another reason why we choose this imputation method.

Since the outcomes of ICE's procedure are multiple imputed data sets, one must apply a suitable averaging tool on these data sets to obtain the final parameter estimates and the associated standard errors. One of such tools is provided by the developer of ICE as a command in Stata called MICOMBINE and it is what we use here.

Table 4.10 shows results from four different multiple imputations. The number of imputed data sets is fixed at ten across these imputations. Also, since there is only one variable to be imputed, the number of cycles is automatically set to one by the program. Like other imputation methods, MICE is valid under MAR. For the MI1 and MI2 estimators, the MAR assumption in (4.102) is imposed on the MDM. For the MI3 and MI4 estimators, the MDM is assumed to follow MAR assumption in (4.103), i.e., the log of the derived variable is used as a extra conditioning variable. While the MI1 and MI3 estimators impute missing values using directly predictive values from the imputation regression, the MI2 and MI4 estimators use the predictive mean matching. All four estimators employ a bootstrap estimator to avoid making the multivariate normality assumption.

It can be seen from Table 4.10 that estimates from all multiple imputation estimators are significant and that their signs are the same as those from the ULS1 estimator. The estimates from the MI1 and MI2 estimators are very similar to the ULS1 estimator even though this ULS estimator is based only on the complete cases. This is because the MAR assumption in (4.102), which underpins the MI1 and MI2 estimators, implies that $\mathcal{P}\{Y =$

$y|X = x, R = 1\} = \mathcal{P}\{Y = y|X = x, R = 0\} = \mathcal{P}\{Y = y|X = x\}$. In other words, (4.102) implies that one can make the inference about any feature of $\mathcal{P}\{Y = y|X = x\}$ directly from $\mathcal{P}\{Y = y|X = x, R = 1\}$. This is precisely what the ULS1 estimator does; it estimates the conditional mean of Y given X, i.e., $E[y|X = x]$ from $E[y|X = x, R = 1]$. Since the MI1, MI2 and ULS1 estimators impose the similar assumption on the MDM, one should expect to see these estimators giving similar results.

It is therefore not a surprise to see that the MI3 and MI4 estimators give different results to those from the MI1, MI2 and ULS1 estimators because they impose (4.103) on the MDM. Moreover, by comparing the MI1 and MI3 estimators to the MI2 and MI4 estimators respectively, one can see that using the predictive mean matching leads to a slight increase in absolute terms of the resultant estimates. This impact is however very small relative to that of using (4.103) instead of (4.102) or, to put it differently, of including the log of the derived variable as an additional conditioning variable in the MAR Assumption.

Notice that although one can run an imputation regression of Y on X and Z, it is illogical to implement a ULS estimator of this specification. An obvious reason is that, for the imputation regression, the objective is to impute values of Y but not to make inference from the resultant coefficient estimates. However, the goal of applying the ULS estimator is to make the inference and, consequently, including an endogenous variable such as Z will bias the results.

4.2.6 Comparison

In this subsection, we examine the effects on the estimates for β of imposing different assumptions on the MDM and using various estimation procedures. For this purpose, results of certain estimators from previous subsections are reproduced in Table 4.11. The estimates of the constant term are omitted. As above, the ULS1 estimator is used as the benchmark model. For IPWLS estimation, the Scobit-IPWLS2 and Hetprob-IPWLS2 estimators are chosen because their binary response models fit with the data better than others. Since these two IPWLS estimators assume the MAR assumption in (4.103), we also include the Probit-IPWLS1 and Logit-IPWLS2 estimators, which maintain the MAR assumption in (4.102) for comparison. Likewise, the MI2 and MI4 estimators are selected to represent the multiple imputation method because they assert two different MAR assumptions. For the SS model, the SSML1 and SSTS1 estimators are chosen since Z should not be employed as an exclusion restriction. Lastly, the 19DGMM, CGMM, CGMM1 and PL estimators represent the group of estimation procedures which imposes (4.105) on the MDM.

Coefficient estimates of each estimator are of the same sign as those of the ULS1 estimator. Exceptions are the SSML1, SSTS1 and CGMM1 estimators. The SSTS1 and CGMM1 estimators give contradictory sign to the estimate of part-time and female respectively. The SSML1 estimator assigns the opposite sign to two variables which are part-time and years of education.

The first row of Table 4.11 shows results from four estimators which restrict the MDM to be (4.102). As noted before, the estimates from the ULS1 and the MI2 estima-

tors are very similar since both estimators assume (4.102). This conclusion is however not applicable to the estimates of the Probit-IPWLS1 and Logit-IPWLS1 estimators. Even though both estimators maintain (4.102), their coefficient estimates of married are not significant at 5% and a few other estimates are also relatively distinct from those of the ULS1 estimator. This indicates that the estimation procedure adopted indeed has influence on the outcome under consideration.

Moreover, all estimators which are in the second row of the table impose (4.103) on the MDM. In other words, these estimators use Z as an extra conditioning variable in the MAR assumption. For MICE, this leads to an increase, in absolute terms, of almost all coefficient estimates of the MI4 estimator in comparison to both the MI2 and ULS1 estimators. This demonstrates that the resultant estimates are certainly sensitive to the assumption imposed on the MDM.

However, the impacts of this change on the IPWLS estimation are not as uniform as those on MICE. Even though most of the resultant estimates from the Scobit-IPWLS2 and Hetprob-IPWLS2 increase relative to those of the benchmark estimator, married and size of the work place are no longer statistically significant. The estimates of these two IPWLS estimators for certain variables are also fairly different. In fact, this is also true if we compare the Probit-IPWLS1 estimator to the Logit-IPWLS1 estimator. All of these factors suggest that, as an estimation method, the IPWLS estimation is noticeably sensitive to changes in MAR assumption and the specification of the selection equation.

The third row of the table reports results from the SS model estimators whose MDM is assumed to be (4.104) while the final row presents the results from the GMM and PL

estimator which are based on (4.105). As can be seen, the estimates from both SS model estimators are dramatically distinct from other estimators in Table 4.11. For example, the SSML1 estimator implies that the effect of years of education on wage rate is significantly negative. It is likely that such an irrational implication is a consequence of having no exclusion restriction.

In comparison to the ULS1 estimator, maintaining (4.105) increases moderately, in absolute terms, almost all coefficient estimates of the GMM and PL estimators; especially the marginal effect of having a part-time job as the main job. Some estimates of these estimators are also of similar size to those of the MI4 estimator. This may due to the fact that both (4.105) and (4.103) permit the MDM to be related to Y. While (4.105) asserts this relationship explicitly, (4.103) allows the MDM to vary with Y through the observable variable Z.

It is impossible to assert which estimator is the most reliable because the true DGP and MDM are unknown. However, it is clear from the results in the table that the outcomes of interest are sensitive to both assumptions on the MDM and estimation procedures employed.

4.3 Estimation of the proportion of UK wage below the NMW

A main objective of Skinner et al (2002) is to estimate the proportion of wage rate below the NMW. The June-August 1999 quarter of the LFS is collected when the NMW is £3.60 per hour for people whose age is over 22 years old. If the direct variable was fully observed,

the straightforward estimator for the proportion of interest would be

$$\widehat{\mathcal{P}}\{y < \log(3.6)\} = \sum_{n=1}^{N} \frac{I[y_n < \log(3.6)]}{N}, \tag{4.106}$$

where $I[\cdot]$ is the indicator function. Due to the missing data, we cannot however use (4.106) to estimate $\mathcal{P}\{y < \log(3.6)\}$ directly from the data. We can only estimate $\mathcal{P}\{y < \log(3.6)|r = 1\}$ from the complete cases and, as shown in Table 4.12, the estimates of this proportion are 0.0358 for all individuals and 0.0189 for $22+$ age group. These estimates are not consistent for $\mathcal{P}\{y < \log(3.6)\}$ in both age groups unless $\mathcal{P}\{y < \log(3.6)|r = 1\} = \mathcal{P}\{y < \log(3.6)|r = 0\}$.

Since the NMW is for people aged $22+$, we should logically strict our attention on this subsample and discard observations whose age is below or is equal to 22 years old. Nonetheless, we decide not to drop these observations. This is because, firstly, we already lose approximately 74% of the sample to missing data. If we also exclude these observations, the number of complete cases will reduce from 4495 to 4046. Secondly, we are not only interested in the proportion below the NMW but also in the effects of covariates on wage rate as shown in the previous section. Thirdly, the probability of the direct variable being missing should not depend on whether or not your age is over 22 years old. Hence, the decision to include these observations should not bias the estimation. In what follows, we will report the proportion estimates for all-age group and for $22+$ age group. The latter will be computed using the same set of $\widehat{\beta}$ but only with covariates from observations whose age is more than 22 years old.

To see the effect of using the derived variable instead of the direct variable on the estimation, suppose for a moment that our interest is on $\mathcal{P}\{y < \log(3.6)|r = 1\}$. Table

4.12 shows that, by replacing Y with log of the derived variable, the estimates of this proportion are 0.1321 for all individuals and 0.1154 for $22+$ age group. By comparing these estimates to the supposedly consistent estimates from the direct variable, the measurement error in the derived variable clearly causes an upward bias in the estimation. This point is also stressed in Skinner et al. Table 4.12 also reports the estimates of $\mathcal{P}\{y < \log(3.6)\}$ using the derived variable and the mixture of the direct and derived variables. In the mixture data, missing values of the direct variable are replaced by the corresponding observed values from the derived variable. The resultant estimates based on the mixture data are smaller than those from the derived variable. This indicates again that the measurement error causes overestimation of the proportion.

To understand why the derived variable has such an effect, we consider, in Table 4.13, how well the derived variable approximates the direct variable in the subsample. The values of these two variables are divided into three groups: (i) less than 3.6, (ii) at 3.6 and (iii) greater than 3.6. For values below 3.6, the percent correctly predicted of the derived variable is 70.19%, which is relatively high. This measure is even higher in the group of values above the NMW. However, the percent correctly predicted for values at the NMW is only 13.73%.

According to the direct variable, there are 408 employees whose wage rate is at the NMW resulting in a large spike in its distribution. The problem is that the derived variable puts only 56 of these employees correctly into the second group but puts the rest of them relatively equally into the other two groups. Thus, the spike is effectively smoothed in the derived variable and this leads to higher proportion estimates of the first and third groups.

Notice that the disappearance of the spike feature in the data is a separate problem from the two sampling issues previously described. This means that there are, at least, three issues which should be resolved in estimating the proportion under the NMW. Nevertheless, it has been indicated that sampling design and unit non-responses may not cause any estimation problem in this context. Thus, we should be able to safely eliminate this sampling issue from our consideration and focus on the other two.

Another difficulty is that the size of the proportion of interest is usually small. This makes its estimate very sensitive to the error of the estimation procedure. For instance, notice from Table 4.13 that the derived variable incorrectly places 302 individuals whose wage rates are supposed to be greater than 3.6 into the first group. This number of employees is small in comparison to the size of the third group. Therefore, this mistake has only minimal effect on its proportion estimate. However, it is markedly large relative to the first group. As a result, this error incorrectly increases the proportion estimate of the first group.

Let Y^I denote the imputed value of wage rate and let $\widetilde{Y} = Y$ if $R = 1$ and $\widetilde{Y} = Y^I$ if $R = 0$. Because of the missing data problem, Skinner et al suggests the following estimator for the proportion of interest:

$$\sum_{n=1}^{N} \frac{I[\widetilde{y}_n < \log(3.6)]}{N}. \tag{4.107}$$

Apparently, (4.107) is an adaptation of (4.106) where the missing values are replaced by the imputed values. Another possible estimator if one is willing to also estimate IP weights

is that

$$\frac{\sum_{n=1}^{N} r_n \cdot \widehat{p}_n^{-1} \cdot I[y_n < \log(3.6)]}{\sum_{m=1}^{N} r_m \cdot \widehat{p}_m^{-1}}, \tag{4.108}$$

where $\widehat{p}^{-1}$ is the IP weight or the inverse of probability of being observed estimated by a binary choice model.

Skinner et al chooses to estimate the proportion using (4.107). They use the donor imputation to calculate Y^I from the observed values. They also experiment with several imputation models but their chosen specification yields the estimate of 0.0153 or 1.53% for the $22+$ age group. For convenience, this particular method of calculating Y^I will be referred to as Skinner's Imputation.

Notice that the two estimation problems described are addressed in Skinner's Imputation. The issue of the spike at the NMW is dealt with using the donor imputation because this particular imputation technique can recreate the spike feature in the imputed data set as discussed in Section 4.2.5. The missing data problem is also overcome by maintaining the MAR assumption in (4.103).

In this study, we will use both (4.107) and (4.108). For (4.107), we mimic the procedure of Skinner's Imputation by fixing the number of imputed data set at one and using MICE with predictive mean matching to compute Y^I. For (4.108), the IP weights are computed by the binary response models associated with the Scobit-IPWLS2 and Hetprob-IPWLS2 estimators because they perform better than other specifications.

There are two other methods which will be used to estimate the proportion of wage rate under the NMW. First, since all the GMM and PL estimators already assume that the distribution of Y given X is normal, one can calculate the proportion of interest using the

standard normal CDF and the resultant parameter estimates $\widehat{\beta}$. That is, the proportion of interest can be estimated by

$$\Phi\left(\frac{\log(3.6) - x'\widehat{\beta}}{\widehat{\sigma}_\varepsilon}\right), \tag{4.109}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution. Due to the dependence of (4.109) on X, we must compute it using interesting values of X. Two natural approaches are (i) using the sample averages of explanatory variables to replace X and (ii) evaluating (4.109) at each observation and calculating the sample average of the outcomes. These two approaches will be referred to as CDF and ACDF respectively. In fact, (4.109) can be applied to other estimators considered if ε is assumed to be independently and normally distributed. Unlike the GMM and PL estimators, some of these estimators do not parameterise σ_ε and its estimate must therefore be computed from the residuals, $\widehat{\varepsilon}$. For comparison purposes, such estimates will be shown with those from the GMM and PL estimators.

The second alternative method is to use the parameter estimates, $\widehat{\beta}$, to impute the missing values of Y and to apply (4.107). This method is simple and can be applied to all estimators analysed. However, it is well known that using the conditional mean, i.e., $x'\widehat{\beta}$ to impute missing values can distort the inference; especially when the tails of the distribution are of interest. Little and Rubin (2002, p.65) recommends therefore drawing an imputed value from the predictive distribution of the missing values instead of the conditional mean. In practice, this means the imputed values are calculated from

$$y^I = x'\widehat{\beta} + \varepsilon^I, \tag{4.110}$$

where the distribution of ε^I is pre-specified with mean zero and variance $\widehat{\sigma}_\varepsilon$. The variance of the residuals $\widehat{\varepsilon}$ can again be used as $\widehat{\sigma}_\varepsilon$ for the estimators that do not estimate σ_ε along

with β. Little and Rubin refer to this method of imputation as Conditional Draw. We will impose in this study that ε^I is normally distributed so that the resultant estimates are comparable to those from (4.109).

Although the Conditional Draw in (4.110) is a type of imputation, it is different from MICE and Skinner's imputation. The nature of this difference is the same as that between an imputation regression and a standard LS regression. That is, the Conditional Draw method in this study uses the estimates of β, which are of interest in themselves, as the basis of the imputation. In contradistinction, the coefficient estimates from Skinner's imputation are ignored and may even offer no meaningful interpretation. They are only exploited to compute the imputed values. For example, an endogenous variable like Z can appear in the imputation regression of Skinner's imputation but cannot be an explanatory variable in (4.110). Another dissimilarity is that the Conditional Draw uses (4.110) to calculate Y^I whereas Skinner's imputation uses the predictive mean matching.

An advantage of using the Conditional Draw, CDF and ACDF is that the missing data problem can be solved not only by MAR but also by other assumptions such as NMAR or (4.105). This depends on what kind of assumption is assumed in estimating β. However, a drawback of using these methods is that, unlike Skinner's imputation, they cannot take account of the spike in the distribution of wage rate at the NMW.

4.3.1 Comparison

Tables 4.14 and 4.15 show results from all methods discussed. In all cases considered, the estimates for 22+ age group are smaller than those for all individuals. This is because

the majority of people whose age is younger than 22 years old are low-paid employees. Thus, excluding them must result in lower estimate of the proportion of wage rate under the NMW.

The results from different Skinner's imputations and IP weighted estimators are presented in Table 4.14. The S-IMP1 and S-IMP2 estimators are Skinner's imputations under MAR assumption in (4.102). On the other hand, the S-IMP3 and S-IMP4 estimators impose MAR assumption in (4.103) on the MDM. The predictive mean matching is employed only in the S-IMP2 and S-IMP4 estimators. Thus, only the S-IMP4 estimator has exactly the same specification as the donor imputation of Skinner et al. These various settings of the Skinner's imputation are chosen in order to study the effects of switching between the two MAR assumptions and of using the predictive mean matching.

As can be seen from Table 4.14, the estimate for 22+ age group from the S-IMP4 estimator is 0.0151 which is very close to the estimate of 0.0153 from Skinner et al. This result is very encouraging because it means that this imputation estimator, as it is intended to, approximates well the procedure of donor imputation in Skinner et al. Moreover, it is evident from the results that replacing (4.102) with (4.103) reduces the proportion estimates for both groups. The effect of using the predictive mean matching on the estimates is even more sizeable. Under (4.102), using the predictive mean matching reduces dramatically the proportion estimate for 22+ age group from 0.0726 to 0.0213. Skinner's imputations with the predictive mean matching also produce estimates that are closer to those from the IP weighted estimators which are very low. For 22+ age group, both IP weighted

estimators predict that the proportion of interest is even less than one percent of the working population.

For each estimator considered in Table 4.15, the order of the proportion estimates from the Conditional Draw, CDF and ACDF are such that ACDF > Conditional Draw > CDF. The only exception is the estimates from the SSML1 and SSTS1 estimators. The sizes of these estimates are unrealistically large for the proportion of wage rates below the NMW. As before, this could be a consequence of not having exclusion restrictions and we therefore exclude these estimates from our consideration.

Furthermore, it is clear from the results in Tables 4.14 and 4.15 that different estimation methods can lead to dissimilar resultant estimates of the proportion of interest. Moreover, the results of ACDF that are based on the 13DGMM and 19DGMM estimators are approximately the same, namely, 0.05. This is exactly the same as the estimates for population share of the first group, Q_1, from these two estimators in Section 4.2.4. The reason for such identical results is that the way in which both estimators calculate $\widehat{Q}_1$ is the same as how ACDF compute its results.

In Table 4.15, the estimators which permit the MDM to be related to Y tend to give smaller estimates than the other estimators. For instance, the proportion estimates from the 13DGMM and 19DGMM estimators are smaller than those from the 13INTREG and 19INTREG estimators. Also, the MI3 and MI4 estimators produce the estimates for the proportion below the NMW which are smaller than those from the MI1 and MI2 estimators. However, this statement is not true for the Hetprob-IPWLS2 estimator. If one uses the ACDF or Conditional Draw methods, Hetprob-IPWLS2's estimates are then as large as

those from the ULS1 estimator. This anomaly may be due to the fact that the IPWLS estimation tends to be considerably sensitive to a change in the specification of the selection equation.

Moreover, the estimators which assume (4.105) such as the CGMM, CGMM1 and PL estimators seem to give lower estimates than the estimators which are based on (4.103) such as the MI4 and Scobit-IPWLS2 estimators. Notice that (4.105) deliberately permit the MDM to depend on Y while (4.103) allows this dependency indirectly via the inclusion of Z. This may indicate that the degree to which MDM is allowed to vary with Y can also affect the outcome.

Skinner et al are fairly confident that their proposed estimator yields improved estimates in comparison to other methodologies previously used by the ONS. If this is true then 0.0153 might be used as the benchmark of a reasonable estimate for the proportion of UK wage rates below the NMW. Since only the estimators that allow the MDM to be related to Y such as the MI4, CGMM, CGMM1, PL and 19DGMM estimators can produce such low estimates, one may conclude that they are the appropriate estimators for this empirical context. Table 4.15 also suggests that the estimate of the proportion of interest is relatively sensitive to the computation method used. For example, based on the CGMM estimator, the estimate for 22+ age group can be 0.0180, 0.0312 or 0.0489 depending on which method is used.

4.4 Summary

This study shows that using both different assumptions on the MDM and estimation procedures certainly have considerable effect on the estimation of (i) coefficient parameters of the conditional mean function of wage rate and (ii) the proportion of the UK wage rate below the NMW. It confirms that the SS model does not produce plausible results when there is no exclusion restriction. Also, the IPWLS estimation seems to be fairly sensitive to the changes in the specification of the selection equation.

In the empirical application under consideration, the MDM is likely to depend on the dependent variable. As a result, the availability of a variable such as Z proved to be important for any estimation procedure which is based on the MAR assumption. Without such a variable, results from these estimators can be as unreliable as those of the procedures which are based only on the complete cases. The estimators of interest, namely, the GMM and PL estimators seem to be appropriate for this empirical context. This is because they permit the MDM to vary with the dependent variable requiring neither an exclusion restriction nor a variable such as Z. We however do not claim that these estimators are perfect for the application. If a variable like Z can be acquired then an imputation method such as MICE may be more attractive since it does not specify the distribution of the dependent variable given the covariates.

As noted above, replacing the parametric binary response and SS models with their semi- or nonparametric counterparts is an interesting extension to this study. One should also attempt to analyse the data by the INRYX-GMM estimator which allows the MDM to be unspecified and to depend on both the dependent and explanatory variables. Moreover,

estimators such as the GMM or PL estimators may be jointly used with the predictive mean matching to compute the imputed values. This would lead to a new type of Skinner's imputations which depend on (4.105) rather than the MAR assumption.

4.A Appendix A: Tables of Results from the Empirical Applications

Table 4.1
Summary measures for distribution of derived variable

Summary Measure	All	Subsample with $r = 1$
Mean	8.75	5.57
Standard Deviation	5.70	3.07
1st percentile	1.63	1.67
5th percentile	3.01	2.81
10th percentile	3.60	3.25
25th percentile	4.64	3.85
50th percentile	6.67	4.80
75th percentile	9.97	6.49
90th percentile	14.40	8.60
95th percentile	17.80	10.54
99th percentile	29.26	16.00

Notes:
1. The weight piwt03 is used in computing these statistics to adjust for the sampling scheme and unit-nonresponses

Table 4.2
Complete Case Analysis using LS regression

	ULS1	ULS2	ULS3	ULS4
Years of Education	0.045(0.003)	0.045(0.003)	0.043(0.004)	0.044(0.004)
Experience/10	0.165(0.012)	0.168(0.012)	0.168(0.013)	0.166(0.013)
Experience^2/100	−0.026(0.002)	−0.027(0.002)	−0.027(0.002)	−0.027(0.002)
Married	0.047(0.010)	0.046(0.009)	0.047(0.010)	0.046(0.010)
Female	−0.161(0.010)	−0.158(0.010)	−0.155(0.011)	−0.156(0.012)
Size25	0.099(0.009)	0.098(0.009)	0.098(0.009)	0.098(0.009)
Part-time	−0.117(0.011)	−0.119(0.010)	−0.120(0.012)	−0.121(0.012)
London and SE	0.098(0.011)	0.094(0.010)	0.099(0.012)	0.100(0.012)
Constant	0.940(0.037)	0.936(0.036)	0.949(0.054)	0.949(0.055)
Log Likelihood	−726.56	−542.15		
Sample Size	4495	4489	4453	4495

Notes:

1. The ULS1 estimator is based on all complete cases; the ULS2 estimator is based on the data set without outliers; the ULS3 and ULS4 estimators are weighted by piwt03 and pwt03, respectively.

2. piwt03 assigns zero to 42 observations, hence the dedection in the sample size

Table 4.3
IPWLS Estimation with Probit model

	ULS1	Probit-IPWLS1	Probit-IPWLS2	Prbit-IPWLS3
Years of Education	0.045(0.003)	0.048(0.003)	0.042(0.012)	0.044(0.004)
Experience/10	0.165(0.012)	0.365(0.068)	0.207(0.044)	0.259(0.054)
Experience^2/100	−0.026(0.002)	−0.064(0.012)	−0.036(0.008)	−0.049(0.010)
Married	0.047(0.010)	**0.014(0.034)**	0.066(0.031)	0.074(0.037)
Female	−0.161(0.010)	−0.131(0.029)	−0.137(0.032)	−0.161(0.039)
Size25	0.099(0.009)	0.304(0.082)	0.079(0.026)	0.135(0.065)
Part-time	−0.117(0.011)	−0.090(0.030)	−0.140(0.027)	−0.128(0.034)
London and SE	0.098(0.011)	0.155(0.039)	0.151(0.039)	0.181(0.044)
Constant	0.940(0.037)	0.543(0.123)	0.928(0.157)	0.922(0.084)
Log Likelihood	−726.56	−208.72	−1754.28	−1981.66
Sample Size	4495	4495	4493	4495
4 Largest IP Weights		125.75	118.31	98.85
		223.16	118.37	151.01
		4949.95	143.72	589.17
		17840.73	272.72	2523.48
LM test for non-normality		14.33		116.78
Percent Correctly Predicted for $r = 1$		0.19		0.24
Percent Correctly Predicted for $r = 0$		0.95		0.94

Notes:

1. The Probit-IPWLS1 estimator uses only X in both regression and probit models; the Probit-IPWLS2 estimator uses the same models and data set but without the two problematic observations; the Probit-IPWLS3 estimator uses both X and Z in the probit model.

2. Any highlighted coefficient estimate is insignificant at 5 percent

Table 4.4
IPWLS Estimation with Logit and Complementary log-log model

	ULS1	Logit-IPWLS1	Logit-IPWLS2	Cloglog-IPWLS1	Cloglog-IPWLS2
Years of Education	0.045(0.003)	0.043(0.006)	0.050(0.007)	0.042(0.006)	0.049(0.006)
Experience/10	0.165(0.012)	0.228(0.038)	0.232(0.036)	0.197(0.025)	0.211(0.025)
Experience^2/100	−0.026(0.002)	−0.040(0.007)	−0.042(0.007)	−0.034(0.004)	−0.037(0.004)
Married	0.047(0.010)	**0.045 (0.025)**	0.065(0.027)	0.047(0.018)	0.053(0.019)
Female	−0.161(0.010)	−0.144(0.024)	−0.174(0.030)	−0.154(0.019)	−0.171(0.021)
Size25	0.099(0.009)	0.140(0.040)	0.118(0.039)	0.113(0.022)	0.114(0.024)
Part-time	−0.117(0.011)	−0.118(0.024)	−0.113(0.028)	−0.116(0.019)	−0.110(0.021)
London and SE	0.098(0.011)	0.134(0.028)	0.154(0.033)	0.121(0.021)	0.132(0.023)
Constant	0.940(0.037)	0.869(0.090)	0.875(0.089)	0.932(0.077)	0.887(0.074)
Log Likelihood	−726.56	−1484.14	−2023.95	−1380.39	−1667.94
Sample Size	4495	4495	4495	4495	4495
4 Largest IP Weights		78.39	61.68	46.98	38.02
		121.14	82.29	60.46	47.62
		655.01	206.10	231.24	136.96
		1158.24	432.94	343.33	223.79
Percent Correctly Predicted for $r = 1$	0.21	0.25	0.18	0.21	
Percent Correctly Predicted for $r = 0$	0.94	0.93	0.95	0.94	

Notes:

1. The Logit-IPWLS1 and Cloglog-IPWLS1 estimators use only X in both regression and selection models; the Logit-IPWLS2 and Cloglog-IPWLS2 estimators include also Z in their selection models

Table 4.5
IPWLS Estimation with Scobit Model

	ULS1	Logit-IPWLS1	Scobit-IPWLS1	Logit-IPWLS2	Scobit-IPWLS2
Years of Education	0.045(0.003)	0.043(0.006)	0.044(0.006)	0.050(0.007)	0.042(0.010)
Experience/10	0.165(0.012)	0.228(0.038)	0.273(0.055)	0.232(0.036)	0.330(0.095)
Experience^2/100	−0.026(0.002)	−0.040(0.007)	−0.048(0.010)	−0.042(0.007)	−0.072(0.019)
Married	0.047(0.010)	**0.045 (0.025)**	**0.038 (0.031)**	0.065(0.027)	**0.128 (0.099)**
Female	−0.161(0.010)	−0.144(0.024)	−0.137(0.028)	−0.174(0.030)	−0.199(0.087)
Size25	0.099(0.009)	0.140(0.040)	0.180(0.062)	0.118(0.039)	**0.023 (0.125)**
Part-time	−0.117(0.011)	−0.118(0.024)	−0.115(0.029)	−0.113(0.028)	−0.188(0.076)
London and SE	0.098(0.011)	0.134(0.028)	0.145(0.035)	0.154(0.033)	0.292(0.106)
Constant	0.940(0.037)	0.869(0.090)	0.770(0.107)	0.875(0.089)	1.107(0.247)
Log Likelihood	−726.56	−1484.14	−1401.88	−2023.95	−3047.02
Sample Size	4495	4495	4495	4495	4495
4 Largest IP Weights		78.39	110.55	61.68	448.50
		121.14	182.49	82.29	479.22
		655.01	1411.03	206.10	484.27
		1158.24	2982.69	432.94	2644.74
Percent Correctly Predicted for $r = 1$	0.21	0.20	0.25	0.25	
Percent Correctly Predicted for $r = 0$	0.94	0.94	0.93	0.92	
LR test of alpha = 1 (chi2(1))		16.40		236.57	

Notes:

1. The Scobit-IPWLS1 estimator uses only X in both regression and selection models; the Scobit-IPWLS2 estimator includes also Z in its selection model

Table 4.6
IPWLS Estimation with Hetprob Model

	ULS1	Hetprob-IPWLS1	Hetprob-IPWLS2	Hetprob-IPWLS3	Hetprob-IPWLS4
Years of Education	0.045(0.003)	0.058(0.010)	0.142(0.016)	0.171(0.027)	**-0.003 (0.014)**
Experience/10	0.165(0.012)	0.234(0.033)	0.411(0.084)	0.225(0.089)	0.769(0.121)
Experience^2/100	−0.026(0.002)	−0.038(0.005)	−0.079(0.014)	−0.078(0.019)	−0.203(0.022)
Married	0.047(0.010)	**0.036 (0.027)**	**0.076 (0.100)**	−0.324(0.069)	−0.265(0.124)
Female	−0.161(0.010)	−0.166(0.030)	−0.342(0.090)	−0.691(0.145	−0.794(0.122)
Size25	0.099(0.009)	0.091(0.026)	**0.003 (0.115)**	**-0.074 (0.112)**	−0.375(0.077)
Part-time	−0.117(0.011)	−0.123(0.023)	−0.201(0.086)	0.433(0.128)	**0.114 (0.318)**
London and SE	0.098(0.011)	0.112(0.029)	0.238(0.105)	0.407(0.157)	0.242(0.109)
Constant	0.940(0.037)	0.724(0.140)	**-0.070(0.219)**	0.550(0.389)	2.822(0.280)
Log Likelihood	−726.56	−1663.08	−2627.31	?898.04	−190.80
Sample Size	4495	4495	4495	4495	4495
3 Largest IP Weights		83.33	2003.52	17000.79	14259.65
		91.04	2170.64	47473.08	77797.92
		180.26	2449.80	292915.20	185664.6
Percent Correctly Predicted for $r = 1$	0.18	0.25	0.10	0.05	
Percent Correctly Predicted for $r = 0$	0.94	0.93	0.97	0.98	
LR test of lnsigma2 = 0	149.15	403.90	910.39	782.38	

Notes:

1. The Hetprob-IPWLS1 estimator does not use Z; the Hetprob-IPWLS2 estimator includes Z in both selection and variance equations; the Hetprob-IPWLS3 estimator uses Z in its variance equation; Hetprob-IPWLS4's variance equation is a function of Z only

Table 4.7
Sample Selection Models

	ULS1	SSML1	SSTS1	SSML2	SSTS2
Years of Education	0.045(0.003)	−0.020(0.004)	**0.003 (0.021)**	**-0.004 (0.003)**	−0.076(0.011)
Experience/10	0.165(0.012)	0.066(0.016)	0.099(0.036)	0.104(0.014)	**0.0002 (0.0393)**
Experience^2/100	−0.026(0.002)	−0.013(0.003)	−0.017(0.005)	−0.019(0.003)	**-0.0044 (0.0072)**
Married	0.047(0.010)	**0.008 (0.012)**	**0.018 (0.019)**	**0.019 (0.011)**	**-0.031 (0.030)**
Female	−0.161(0.010)	−0.198(0.012)	−0.207(0.026)	−0.189(0.011)	−0.292(0.031)
Size25	0.099(0.009)	0.074(0.011)	0.086(0.012)	0.091(0.010)	0.078(0.027)
Part-time	−0.117(0.011)	0.104(0.014)	**0.049 (0.084)**	0.602(0.013)	0.384(0.045)
London and SE	0.098(0.011)	**-0.019 (0.013)**	**0.015 (0.043)**	**-0.003 (0.011)**	−0.160(0.034)
Constant	0.940(0.037)	1.344(0.045)	1.128(0.106)	1.170(0.040)	1.320(0.108)
Log Likelihood	−726.56	−9339.89		−8299.87	
Sample Size	4495	4495/17085	4495/17085	4495/17085	4495/17085
LR-test for rho=0		284.90		1665.87	
Inverse Mill ratio			0.335(0.167)		1.125(0.067)
rho	0.922(0.006)	0.836		0.940(0.003)	1.169
pR^2		0.1107			0.1462

Notes:

1. The SSML1 and SSTS1 estimators use only X in both structural and selection equations; the SSML2 and SSTS2 estimators also use Z as an exclusion restriction

2. pR^2 is obtained from a probit model of R^* on X for the SSTS1 estimator and on X and Z for the SSTS2 estimator

Table 4.8
RSGMM and Interval Regression Estimator

	ULS1	13DGMM	13INTREG	19DGMM	19INTREG
Years of Education	0.045(0.003)	0.075(0.011)	0.033(0.002)	0.076(0.011)	0.034(0.002)
Experience/10	0.165(0.012)	0.226(0.050)	0.149(0.011)	0.226(0.050)	0.148(0.011)
Experience^2/100	−0.026(0.002)	−0.038(0.009)	−0.025(0.002)	−0.038(0.009)	−0.025(0.002)
Married	0.047(0.010)	0.062(0.029)	0.044(0.008)	0.063(0.029)	0.045(0.008)
Female	−0.161(0.010)	−0.040(0.036)	−0.153(0.009)	−0.037(0.036)	−0.151(0.009)
Size25	0.099(0.009)	0.087(0.029)	0.088(0.008)	0.089(0.029)	0.090(0.008)
Part-time	−0.117(0.011)	−0.318(0.040)	−0.121(0.009)	−0.322(0.040)	−0.120(0.009)
London and SE	0.098(0.011)	0.161(0.030)	0.087(0.009)	0.165(0.031)	0.089(0.009)
Constant	0.940(0.037)	0.776(0.250)	1.100(0.033)	0.751(0.243)	1.084(0.033)
Log Likelihood	−726.56		−11500.42		−13105.56
Sample Size	4495	17085	4495	17085	4495
1/sigma		2.947(0.101)		2.948(0.097)	
sigma		0.339	0.243(0.003)	0.339	0.245(0.003)
Q1		0.050(0.003)		0.050(0.004)	
Q2		0.010(0.0004)		0.007(0.0002)	
Q3		0.020(0.001)		0.011(0.0004)	
Q4		0.026(0.001)		0.012(0.0004)	
Q5		0.023(0.001)		0.009(0.0003)	
Q6		0.017(0.001)		0.011(0.0004)	
Q7		0.042(0.002)		0.018(0.001)	
Q8		0.039(0.002)		0.011(0.0004)	
Q9		0.059(0.003)		0.022(0.001)	
Q10		0.063(0.003)		0.024(0.001)	

Notes:

1. The 13DGMM and 19DGMM estimators are RSGMM estimators with 13 and 19 groups; The 13INTREG and 19INTREG estimators are interval regressions with 13 and 19 groups

2. The standard errors for both DGMM estimators are for beta/sigma

Table 4.9
CGMM, CGMM1, and Pseudo Likelihood Estimators

	ULS1	19DGMM	CGMM	CGMM1	PL
Years of Education	0.045(0.003)	0.076(0.011)	0.088(0.013)	0.076(0.011)	0.089(0.004)
Experience/10	0.165(0.012)	0.226(0.050)	0.229(0.039)	0.150(0.039)	0.283(0.021)
Experience^2/100	−0.026(0.002)	−0.038(0.009)	−0.036(0.007)	−0.022(0.009)	−0.047(0.004)
Married	0.047(0.010)	0.063(0.029)	0.069(0.014)	0.063(0.010)	0.071(0.013)
Female	−0.161(0.010)	−0.037(0.036)	**-0.036(0.023)**	**0.001(0.022)**	−0.054(0.014)
Size25	0.099(0.009)	0.089(0.029)	0.092(0.015)	0.058(0.015)	0.119(0.013)
Part-time	−0.117(0.011)	−0.322(0.040)	−0.329(0.062)	−0.288(0.014)	−0.390(0.023)
London and SE	0.098(0.011)	0.165(0.031)	0.181(0.029)	0.152(0.010)	0.197(0.015)
Constant	0.940(0.037)	0.751(0.243)	0.632(0.123)	1.082(0.138)	0.584(0.065)
Log Likelihood	−726.56				
Sample Size	4495		17085	17085	4495
ln(sigma)			−0.986(0.112)	−0.823(0.208)	−0.908(0.028)
sigma	0.339	0.373	0.439	0.403	

Table 4.10
Multiple Imputation

	ULS1	MI1	MI2	MI3	MI4
Years of Education	0.045(0.003)	0.045(0.004)	0.046(0.005)	0.068(0.004)	0.075(0.003)
Experience/10	0.165(0.012)	0.162(0.014)	0.162(0.019)	0.244(0.013)	0.263(0.022)
Experience^2/100	−0.026(0.002)	−0.026(0.003)	−0.026(0.004)	−0.039(0.002)	−0.042(0.004)
Married	0.047(0.010)	0.054(0.009)	0.055(0.016)	0.070(0.009)	0.073(0.014)
Female	−0.161(0.010)	−0.166(0.007)	−0.171(0.014)	−0.146(0.008)	−0.163(0.012)
Size25	0.099(0.009)	0.098(0.009)	0.101(0.011)	0.124(0.007)	0.133(0.009)
Part-time	−0.117(0.011)	−0.113(0.010)	−0.113(0.012)	−0.185(0.010)	−0.197(0.013)
London and SE	0.098(0.011)	0.097(0.011)	0.105(0.023)	0.137(0.010)	0.138(0.009)
Constant	0.940(0.037)	0.931(0.056)	0.925(0.075)	0.639(0.054)	0.558(0.048)
Log Likelihood	−726.56				
Sample Size	4495	17085	17085	17085	17085

Notes:

1. The MI1 and MI2 estimators use only X in the imputation regression; the MI3 and MI4 estimators use both X and Z in the imputation regression; Only the MI2 and MI4 estimators use the predictive mean matching

Table 4.11
Comparison

	ULS1	Probit-IPWLS1	Logit-IPWLS1	MI2
Years of Education	0.045(0.003)	0.048(0.003)	0.043(0.006)	0.046(0.005)
Experience/10	0.165(0.012)	0.365(0.068)	0.228(0.038)	0.162(0.019)
Experience^2/100	−0.026(0.002)	−0.064(0.012)	−0.040(0.007)	−0.026(0.004)
Married	0.047(0.010)	**0.014(0.034)**	**0.045 (0.025)**	0.055(0.016)
Female	−0.161(0.010	−0.131(0.029)	−0.144(0.024)	−0.171(0.014)
Size25	0.099(0.009)	0.304(0.082)	0.140(0.040)	0.101(0.011)
Part-time	−0.117(0.011)	−0.090(0.030)	−0.118(0.024)	−0.113(0.012)
London and SE	0.098(0.011)	0.155(0.039)	0.134(0.028)	0.105(0.023)

	Hetprob-IPWLS2	Scobit-IPWLS2	MI4
Years of Education	0.142(0.016)	0.042(0.010)	0.075(0.003)
Experience/10	0.411(0.084)	0.330(0.095)	0.263(0.022)
Experience^2/100	−0.079(0.014)	−0.072(0.019)	−0.042(0.004)
Married	**0.076 (0.100)**	**0.128 (0.099)**	0.073(0.014)
Female	−0.342(0.090)	−0.199(0.087)	−0.163(0.012)
Size25	**0.003 (0.115)**	**0.023 (0.125)**	0.133(0.009)
Part-time	−0.201(0.086)	−0.188(0.076)	−0.197(0.013)
London and SE	0.238(0.105)	0.292(0.106)	0.138(0.009)

	SSTS1	SSML1
Years of Education	**0.003 (0.021)**	−0.020(0.004)
Experience/10	0.099(0.036)	0.066(0.016)
Experience^2/100	−0.017(0.005)	−0.013(0.003)
Married	**0.018 (0.019)**	**0.008 (0.012)**
Female	−0.207(0.026)	−0.198(0.012)
Size25	0.086(0.012)	0.074(0.011)
Part-time	**0.049 (0.084)**	0.104(0.014)
London and SE	**0.015 (0.043)**	**-0.019 (0.013)**

	19DGMM	CGMM	CGMM1	PL
Years of Education	0.076(0.011)	0.088(0.013)	0.076(0.011)	0.089(0.004)
Experience/10	0.226(0.050)	0.229(0.039)	0.150(0.039)	0.283(0.021)
Experience^2/100	−0.038(0.009)	−0.036(0.007)	−0.022(0.009)	−0.047(0.004)
Married	0.063(0.029)	0.069(0.014)	0.063(0.010)	0.071(0.013)
Female	−0.037(0.036)	**-0.036(0.023)**	**0.001(0.022)**	−0.054(0.014)
Size25	0.089(0.029)	0.092(0.015)	0.058(0.015)	0.119(0.013)
Part-time	−0.322(0.040)	−0.329(0.062)	−0.288(0.014)	−0.390(0.023)
London and SE	0.165(0.031)	0.181(0.029)	0.152(0.010)	0.197(0.015)

Table 4.12
Estimation of Proportion of Wage Rate Under the NWM Using the Direct and Derived Variables

	Sample	All	Age>22
Direct variable	subsample	0.0358	0.0189
Derived variable	subsample	0.1321	0.1154
Derived variable	full sample	0.0795	0.0660
Mixture of both variables	full sample	0.0541	0.0413

Notes:
1. For the mixture of both variables, missing values of the direct variable are replaced by the corresponding values from the derived variable

Table 4.13
The Number of Correct Predictions of the Direct Variable by the Derived Variable

		Direct Variable			
		<3.6	=3.6	>3.6	total
Derived Variable	<3.6	113	179	302	594
	=3.6	2	56	14	72
	>3.6	46	173	3610	3829
	total	161	408	3926	

Table 4.14
Estimation of proportion of wage rate under the NMW by Skinner Imputation and IP weight

	All	Age>22
S-IMP1	0.0819	0.0726
S-IMP2	0.0262	0.0213
S-IMP3	0.0568	0.0446
S-IMP4	0.0221	0.0151
Scobit-IPWLS2	0.0150	0.0079
Hetprob-IPWLS2	0.0134	0.0067

Notes:
1. The S-IMP1 and S-IMP2 estimators use only X in the imputation regression; the S-IMP3 and S-IMP4 estimators use both X and Z in the imputation regression; Only the S-IMP2 and S-IMP4 estimators use the predictive mean matching

Table 4.15
Estimation of proportion of wage rate under the NMW by CDF, ACDF and Conditional Draw

Estimators	CDF		ACDF		Conditional Draw	
	All	Age>22	All	Age>22	All	Age>22
ULS1	0.0712	0.0648	0.1104	0.1012	0.0790	0.0692
Scobit-IPWLS2	0.0232	0.0201	0.0680	0.0614	0.0481	0.0403
Hetprob-IPWLS2	0.0293	0.0249	0.1108	0.1017	0.0725	0.0623
SSML1	0.6281	0.6247	0.6239	0.6206	0.4732	0.4713
SSTS1	0.5265	0.5191	0.5253	0.5183	0.3921	0.3869
MI1	0.0710	0.0645	0.1108	0.1014	0.0793	0.0694
MI2	0.0709	0.0644	0.1122	0.1029	0.0801	0.0703
MI3	0.0348	0.0297	0.0844	0.0736	0.0567	0.0456
MI4	0.0298	0.0250	0.0824	0.0713	0.0560	0.0446
CGMM	0.0204	0.0180	0.0663	0.0489	0.0387	0.0312
CGMM1	0.0129	0.0118	0.0254	0.0235	0.0247	0.0193
PL	0.0257	0.0225	0.0663	0.0601	0.0446	0.0367
13INTREG	0.0518	0.0467	0.0896	0.0816	0.0643	0.0551
13DGMM	0.0174	0.0153	0.0495	0.0448	0.0362	0.0290
19INTREG	0.0513	0.0463	0.0891	0.0812	0.0640	0.0549
19DGMM	0.0174	0.0152	0.0503	0.0455	0.0365	0.0291

Chapter 5
Conclusion

Several MDMs are described in this thesis. The most general MDM considered allows the response probability to vary continuously with both the independent and explanatory variables, namely,

$$\mathcal{P}\{R = 1|Y = y, X = x\}. \tag{5.111}$$

This MDM is referred to as Not Missing at Random or NMAR. Chapter 2 examines (5.111) and its two special cases, which are

$$\mathcal{P}\{R = 1|Y = y, X = x\} = \mathcal{P}\{R = 1|Y = y\}, \tag{5.112}$$

and

$$\mathcal{P}\{R = 1|Y = y, X = x\} = \mathcal{P}\{R = 1|Y = y, X_1 = x_1\}, \tag{5.113}$$

where X_1 is a subset of X. These MDMs are of interest because it is more usual to encounter NMAR data in many economic contexts being commonly referred to as the *self-selection problem*.

The RS's estimation procedure is extended in Chapter 2 to deal with NMAR data generated by these three MDMs. Three types of GMM estimators are then proposed, namely, INRYX-GMM, INRY-GMM and INRYX1-GMM estimators, which are associated with the MDMs in (5.111), (5.112) and (5.113) respectively. The cost of the extension from the RS approach is that some nuisance functions must be estimated prior to the GMM estimation. Thus, all of the proposed estimators are two-step GMM estimators.

Even though these two-step GMM estimators use parametric estimation as the first step estimation to simplify the discussion of their asymptotic properties, it can be replaced by any nonparametric estimation. In fact, we also propose in Chapter 2 the one-step INRY-GMM estimator where the nuisance function is essentially estimated by an empirical estimator[16].

The RS GMM and PL estimators, which are based on (5.112), are also introduced in Chapter 2. The RS GMM estimator is developed for cases where Y is a discrete variable. To utilise this GMM estimator for continuous Y, the values of Y must be discretised into a finite number of groups; a method which entails certainly a loss in information. Thus, the INRY-GMM estimator should be more efficient than the RS GMM estimator in this circumstance. Furthermore, since the INRY-GMM estimator extracts more information on the parameters of interest from the likelihood of the nonrespondent units than the PL estimator, it should also be more efficient than the PL estimator.

Chapter 3 presents Monte Carlo evidence on the finite sample performance of the two-step INRY-GMM, RS GMM and PL estimators in comparison to other estimators for missing data. These alternative estimators are IPWLS estimators, unweighted estimators and SS model estimators. The SS model estimators maintain the MDM in (5.111) whereas the unweighted estimators assume that

$$\mathcal{P}\{R=1|Y=y,X=x\}=\mathcal{P}\{R=1|X=x\}, \tag{5.114}$$

---

[16] The moment indicators of this one-step GMM estimator can be derived because the MDM in (5.112) does not depend on X.

that is, the MDM is Missing at Random or MAR. Since the MDM is specified and estimated in the IPWLS estimation, this type of estimator can assume that the MDM depends *either* on the covariates only as in (5.114) *or* on the covariates and some additional fully observed variables. A particular case of interest is whenever the MDM is

$$\mathcal{P}\{R = 1|Y = y, X = x, Z = z\} = \mathcal{P}\{R = 1|X = x, Z = z\}, \tag{5.115}$$

where Z is fully observed and highly correlated with Y.

No estimator dominates in all Monte Carlo experiments of Chapter 3 since they impose different assumption on the MDM, i.e., either (5.111), (5.112), (5.114) or (5.115). Under (5.112), it is found in the chapter that the two-step INRY-GMM estimator dominates the RS GMM and PL estimators in terms of RMSE in all Monte Carlo experiments that their underlying assumptions hold. This Monte Carlo evidence supports the above theoretical conjecture that the INRY-GMM estimator should be more efficient than the RS GMM and PL estimators. Furthermore, if only the conditional density function of Y given X is misspecified, it seems that the estimates for the slope coefficients from these three estimators remain consistent. In such a case, the performance of the two-step INRY-GMM estimator is however inferior to that of the RS GMM and PL estimators.

Although (5.115) can be classified as MAR assumption, it allows the MDM to be correlated with Y through the observable Z. In the Monte Carlo experiments, we found that even when the MDM is a function of Y only, the mean bias of the IPWLS can be small if a variable such as Z is available.

Even though the SS model estimators permit the MDM to be (5.111), their drawback is that they also require an exclusion restriction to guarantee identification. There are some

experiments in Chapter 3 where the MDM is a linear function of Y only and the conditional mean function of Y given X is also linear. In such cases, the MDMs in (5.111) and (5.112) are shown to coincide and, as a consequence, the GMM, PL and SS model estimators are all applicable. We found that if there is no exclusion restriction and the explanatory power of the covariates in the MDM is low, the SS model estimators are dominated and perform poorly especially in terms of RMSE. However, the Monte Carlo investigation also suggests that if there are valid exclusion restrictions and the MDM is explained well by the variation of the covariates, the SS model estimators will outperform the two-step INRY-GMM, RS GMM and PL estimators. This is due partly to the fact that these GMM and PL estimators do not allow the MDM to depend on the covariates.

With the availability of variables such as Z, the IPWLS estimator is able to cope with the endogenous selection problem. An advantage of the IPWLS estimator over the SS model estimators is that it may be easier to find a fully observed variable which is correlated with Y than finding an exclusion restriction. In contrast, the GMM and PL estimators require neither exclusion restriction nor Z in dealing with the self-selection problem.

Chapter 4 compares all estimators employed in the Monte Carlo experiments of Chapter 3 using real data. In addition to the aforementioned estimators, MICE or Multiple Imputation by Chained Equations is also used in the empirical investigation. All estimators are applied to analyse the UK wage distribution. The chapter considers the estimation of (i) coefficient parameters of the conditional mean function of wage rate and (ii) the proportion of the UK wage rate below the NMW. A main objective is to examine the

sensitivity of the outcomes of interest with respect to the assumption used in the estimation procedures; especially assumptions which are imposed on the MDM.

It is clear from the results in Chapter 4 that the outcomes of interest are sensitive to both assumptions on the MDM and estimation procedures employed. For example, although the ULS, IPWLS and MICE estimators assumes that the MDM is (5.114), the results from the IPWLS estimators are considerable different from the other two because IPWLS estimation is also sensitive to the parametric specification of the MDM.

Some evidence suggests that the MDM of this empirical application depends significantly on the dependent variable, i.e., there is the endogenous selection problem. Thus, the estimators which are based on (5.111), (5.112) and (5.115) should give reliable results. In Chapter 4, these estimators are the SS model, GMM, PL, IPWLS and MICE estimators. For (5.115), Z is the log of the derived variable in this empirical context. It can be considered as the dependent variable plus measurement error.

The SS model estimators do not however produce plausible results in this chapter because there is no exclusion. It illustrates again that having valid exclusion restrictions is crucial for a successful application of the SS model estimators. Given (5.115), the performances of the MICE estimators are better than those of the IPWLS estimators because the estimates from IPWLS estimation are sensitive to the parametric specification of the MDM. Further, the INRY-GMM, RS GMM and PL estimators are shown to yield reasonable estimates.

As before, the availability of Z is very important for any estimation procedure which is based on (5.115). Without such a variable, the IPWLS and MICE estimators have to be

based on (5.114) and their results should be unreliable in the presence of the self-selection problem. An advantage of the proposed estimators is that they allow the MDM to vary with the dependent variable without requiring Z. They are also shown to be more efficient than the RS GMM and PL estimators in the simulation studies.

However, their main disadvantage is that they require the correct parametric specification of the conditional density function of Y given X. If a variable like Z can be acquired then an imputation method such as MICE may be more attractive since it does not parametrically specify the conditional density function.

In the future research, one should replace some of the estimators used with their semiparametric counterparts. Various semiparametric SS model estimators are already available in the literature. For IPWLS estimation, one can apply a number of semiparametric binary response models as the first-step estimation. An interesting choice of such methods would be an estimator for distribution free heteroskedastic binary response model developed by Khan (2006). The virtue of this estimator is that, while heteroskedasticity is allowed and no distributional specification is assumed, it is also possible to jointly estimate the regression coefficients and the choice probabilities. In addition, it will also be interesting to compare these semiparametric estimators to the INRYX-GMM estimator which allows the MDM to depend on both Y and X.

An immediate extension of the theoretical work in Chapter 2 concerns elucidation of the asymptotic properties of the proposed GMM estimators if first-step estimation is non-parametric, i.e., when these GMM estimators are special cases of the two-step semiparametric estimator.

Moreover, for all of our GMM estimators, we need to assume the correct specification of $\mathcal{P}\{Y = y|X = x\}$. It may be possible to relax this assumption by using a semi-nonparametric approach. For example, the moment indicator of the INRYX-GMM estimator is

$$\theta : r\frac{\partial \log f(y|x;\theta)}{\partial\theta} - (1-r)\frac{1}{1-\int_{\mathcal{Y}}(h(y,x,r=1;\beta)/f_X(x;\alpha))dy}\int_{\mathcal{Y}}\frac{h(y,x,r=1;\beta)}{f_X(x;\alpha)}\frac{\partial \log f(y|x;\theta)}{\partial\theta}d$$

Whereas $h(y,x,r=1;\beta)$ and $f_X(x;\alpha)$ may be estimated non-parametrically, $f(y|x;\theta)$ cannot be estimated from the observed sample due to missing values in Y. Nevertheless, it may be possible to replace $f(y|x;\theta)$ with a semi-nonparametric estimator such as that due to Gallant and Nychka (1987).

As noted in Stewart (2005), the semi-nonparametric approach of Gallant and Nychka (1987) approximates an unknown density function using a Hermite form. For INRYX, if $y = x'\beta + \epsilon$ where $E[\epsilon|x] = 0$, then we can replace $f(y|x;\theta)$ with $f(\epsilon;\theta)$ and this density may be approximated by

$$f_K(\epsilon) = \frac{1}{\lambda}\left(\sum_{k=0}^{K}\alpha_k\epsilon^k\right)^2\phi(\epsilon),$$

where $\phi(\cdot)$ is the standard normal density and

$$\lambda = \int_{-\infty}^{\infty}\left(\sum_{k=0}^{K}\alpha_k\epsilon^k\right)^2\phi(\epsilon)d\epsilon.$$

With a similar approximation for $\partial f(y|x;\theta)/\partial\theta$, the above moment indicators may re-formulated without the necessity of assuming a correct specification of $\mathcal{P}\{Y = y|X = x\}$.

To sum up, a main contribution of this thesis is the extension of the semiparametric estimation procedures developed in RS to include the case where the response variable is continuous and where the MDM depends on both the response variable and covariates.

We found that, in many circumstances, our proposed estimators perform better than the other estimators described; especially when there is no exclusion restriction or other additional information available. We also demonstrated that the estimators based on the MAR assumption in (5.115), such as the IPWLS or MICE estimators, allow the MDM to be correlated with Y through the observable Z. In practice, since the true MDM is unlikely to be known in any empirical study, we suggest that several estimators which impose different assumptions on the MDM should be used together to examine the sensitivity of the outcomes of interest with respect to the assumptions made and the estimation procedures adopted.

Bibliography

Beissel-Durrant, G. and C. Skinner (2003), "Estimation of the Distribution of Hourly Pay from Household Survey Data", CEMMAP working paper No. CWP12/03.

Chatterjee, N., Chen, Y. and N. E. Breslow (2003) "A Pseudoscpre Estimator for Regression Problems With Two-Phase Sampling", *Journal of the American Statistical Association*, 98, 158-186.

Chen, C. (2001), "Parametric Models for Response-Biased Sampling", *Journal of the Royal Statistical Society*. B 63, 775-89.

Cosslett, S. R. (1981), "Maximum Likelihood Estimation for Choice-Based Samples", *Econometrica*, 49, 1289-1316.

Crockett, A. (2007), "*Weighting the Social Surveys ESDS Goverment*". Available: http://www.esds.ac.uk/government/docs/weighting.pdf. Last accessed 28 September 2007.

Dolton, P. (2002) "Reducing Bias using Targeted Refreshment Sampling and Matched Imputation", Manuscript. Department of Economics, University of Newcastle.

Gallant A.R. and D.N. Nychka (1987), "Semi-Nonparametric Maximum Likelihood Estimation", *Econometrica*, 55, 363-390.

Heckman, J.J. (1976), "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, 475-492.

Heckman, J.J (1979), "Sample Selection Bias as a Specification Error", *Econometrica*, 47, 153-161.

Hirano, K., Imbens, G.W., Ridder, G. and D.B. Rubin (2001) "Combining Panel Data Sets with Attrition and Refreshment Samples", *Econometrica*, 69, 1645-1659.

Hsieh, D.A., Manski, C.F. and D. McFadden (1985), "Estimation of Response Probabilities from Augmented Retrospective Observations", *Journal of the American Statistical Association,* 80, 651-662

Imbens, G.W. (1992), "An Efficient Method of Moments Estimator for Discrete Choice Models with Choice-Based Sampling", *Econometrica*, 60, 1187-1214.

Khan, S. (2006), "Distribution Free Estimation of Heteroskedastic Binary Response Models Using Probit/Logit Criterion Functions", Manuscript. Department of Economics, Duke University.

Liang, K.Y. and J. Qin (2000), "Regression Analysis under Non-Standard Situations: A Pairwise Pseudolikelihood Approach", *Journal of the Royal Statistical Society*. B 62, 773-86.

Little, R.J.A. and D.B. Rubin (2002), *Statistical Analysis with Missing Data*, 2nd edition. New York: Wiley.

Manski, C.F. (1989), "Anatomy of the Selection Pronlem", *Journal of Human Resources*. Vol. 24, No. 3, 343–360.

Manski, C.F. (1994), "The Selection Problem.", In C. Sims, ed., *Advances in Econometrics*. Cambridge: Cambridge University Press.

Manski, C.F. (2003), *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Newey, W.K. and D. McFadden (1994), "Large Sample Estimation and Hypothesis Testing", in *Handbook of Econometrics,* Volume 4, ed. R.F. Engle and D. McFadden. Amsterdam: North Holland, 2111-2245.

Ramalho, E.A. and R.J. Smith (2003), "Discrete Choice Nonresponse", CEMMAP working paper No. CWP07/03..

Robins, J.M. and A. Rotnitzky (1995), "Semiparametric Efficiency in Multivariate Regression Models with Missing Data", *Journal of the American Statistical Association* 90, 122-129.

Robins, J.M., Rotnitzky, A. and L.P. Zhao (1995), "Analysis of Semiparametirc Regression models for Repeated Outcomes in the Presence of Missing Data", *Journal of the American Statistical Association* 90, 106-121.

Royston, P. (2004), "Multiple Imputation of Missing Values", *Stata Journal*, 4(3), 227-241.

Royston, P. (2005a), "Multiple Imputation of Missing Values: Update", *Stata Journal*, 5(2), 188-201.

Royston, P. (2005b), "Multiple Imputation of Missing Values: Update of Ice", *Stata Journal*, 5(4), 527-536.

Rubin, D.B. (1976), "Inference and missing data" (with discussion), *Biometrika*, 63, 581-592.

Schafer, J.L. (1997), *Analysis of incomplete Multivariate Data*, London: Chapman and Hall.

Scharfstein, D.O., Rotnitzky, A. and J.M. Robins (1999), "Adjusting for non-ignorable drop-out using semiparametric non-response models", *Journal of the American Statistical Association*, 94, 1096-1120.

Skinner, C., Stuttard, N., Beissel-Durrant, G. and J. Jenkins (2002), "The Measurement of Low Pay in the UK Labour Force Survey", *Oxford Bulletin of Economics and Statistics*, 64, 653-676.

Stewart, M.B. (1983), "On Least Squares Estimation when the Dependent Variable is Grouped", *The Review of Economic Studies*, 50, 737-753.

Stewart, M.B. (2005), "A Comparison of Semiparametric Estimators for the Ordered Response Model", *Computational Statistics & Data Analysis* 49, 555-573.

Stuttard, N. and J. Jenkins (2001), "Measuring Low Pay using the New Earnings Survey and the Labour Force Survey", *Labour Market Trends*, January 2001, 55-66.

Tang, G., Little, R.J.A. and T.E. Raghunathan (2003), "Analysis of Multivariate Missing Data with Nonignorable Nonresponse", *Biometrica*, 90, 747-764.

Tripathi, G. (2003) "GMM and Empirical Likelihood with Incomplete Data", Manuscript. Department of Economics, University of Wisconsin-Madison.

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data*. New York: Springer-Verlag.

van Buuren S., Boshuizen, H. C. and D. L. Knook. 1999. "Multiple Imputation of Missing Blood Pressure Covariates in Survival Analysis", *Statistics in Medicine*, 18, 681-694.

van der Laan, M.J. and J.M. Robins (2003), *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer-Verlag.

Vella, F. (1998), "Estimating Models with Sample Selection Bias: A Surevey", *Journal of Human Resources*. Vol. 33, No. 1, 127–169.

Werner, B. (2006), "Reflections on Fifteen Years of Change in using the Labour Force Survey", *Labour Market Trends*, Vol. 114, No. 08, 257-277.

Wooldridge, J. M. (2002a), "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification", *Portuguese Economic Journal* 1, 117-139.

Wooldridge, J. M. (2002b), *Econometric Analysis of Cross Section and Panel Data.* Cambridge, MA: MIT Press.

Wooldridge, J. M. (2003), "Inverse Probability Weighted Estimation for General Missing Data Problems", Manuscript. Department of Economics, Michigan State University.