A MOMENT SUBSTITUTION APPROACH TO FITTING LINEAR REGRESSION MODELS WITH CATEGORICAL COVARIATES SUBJECT TO RANDOMIZED RESPONSE

WANG ZIJIAN GERALD

(B.Sc (Hons), National University of Singapore)

A THESIS SUBMITTED
FOR THE DEGREE OF MASTER OF SCIENCE
DEPARTMENT OF STATISTICS AND APPLIED PROBABILITY
NATIONAL UNIVERSITY OF SINGAPORE
2008

Acknowledgements

I am indebted to many individuals who contributed directly and indirectly to the completion of this study. First and foremost, I would like to express my gratitude to my supervisor, Associate Professor Chua Tin Chiu, for his advice and patience in bringing this project to fruition. His candid comments and suggestions were invaluable in elevating my understanding of this subject, and in improving the quality of this report.

Also, I would like to thank my family, especially my mother, for her understanding, support and guidance throughout this gruelling academic endeavour. She is my guiding light in the darkness, who never fails to remind me to always keep my feet firmly planted in the ground. Lastly, I thank my dear friends and colleagues for the intriguing intellectual exchange, and for all the laughter and fun they have brought into my life.

Contents

Summary

In this paper, we present an alternative approach to Van den Hout and Kooiman (2006) for estimating the linear regression model with categorical covariates subject to randomized response (RR). Specifically, we consider Warner's (1965) scheme of randomization. Our approach essentially consists of moment substitution, where we estimate the latent first, second and cross product moments in the usual least squares estimator for the centred model with their associated observed unbiased estimates. For the problem of estimating subgroup means in a dichotomous population, we show that this moment substitution approach is equivalent to Selen's (1986) estimator under appropriate distributional assumptions. Assuming independent randomizations, this approach is further adapted to the case of multiple linear regression, when some or all of the covariates are subject to RR. Ultimately, it is shown that the estimates yielded by this method are asymptotically equivalent to the measurement error model estimates of Fuller (1987) under suitable transformations.

List of Tables

Chapter 1

Introduction

1.1 Motivation

Randomized response is an interview technique first introduced by Warner (1965) to circumvent the problem of evasive answer bias in survey studies when sensitive questions have to be answered. Examples of such questions can include questions about income, alcohol consumption or criminal history. In any case, regardless of what is actually asked, the reluctance to reveal personal details to a stranger tends to evoke less than truthful responses from an interview respondent.

A typical RR design gets around this by only requiring respondents to answer questions on a probability basis. One crucial aspect of this design is that the interviewer only has access to the probability that a respondent answers a particular question, but not the question being answered, thereby safeguarding the identity of the respondent. In this light, variables subjected to RR can be looked upon as misclassified variables whose conditional misclassification probabilities are known. This warrants the development and use of specialized techniques that take into account this misclassification. For a review of statistical methods associated with such data, see Van den Hout and Van der Heijden (2004); for a concise account of classical RR methodology and techniques, see Chaudhuri and Mukerjee (1988).

However, while there is a substantial interest in RR techniques in the literature, little attention has been paid to the treatment of linear regression models with RR covariates. Specifically, consider a scenario where a response characteristic is believed to depend on a sensitive quality that one cannot easily measure truthfully in a survey study. In this case, a RR design can be used to extract the relevant information, but fitting the usual linear regression model based on this RR covariate will result in erroneous conclusions.

One way of adjusting for the effects of RR covariates in conventional linear regression models was introduced by Van den Hout and Kooiman (2006). By building upon the methods used by Spiegelman et al. (2000) in studying logistic regression models with misclassified covariates and measurement errors, Van den Hout and Kooiman's approach consists of deriving the likelihood function of the regression model with RR covariates, and then maximising this likelihood using an EM algorithm. Through a simulation study, for low values of misclassification probabilities (~ 0.1), they verified that this adjustment results in estimates that do not show any structural bias.

However, an inherent feature of this approach is that it can become computationally inefficient when dealing with large sample sizes, which is typical of RR designs to offset the additional variance introduced due to misclassification. In this thesis, we attempt to overcome this inefficiency by suggesting an alternative moment substitution procedure to the maximum likelihood approach of Van den Hout and Kooiman.

The outline of this thesis is as follows. The following section describes some of the preliminaries and notation that will be used throughout this paper. In Chapter

2, we discuss the application of our procedure to the problem of estimating subgroup means in a dichotomous population. In Chapter 3, we extend this procedure to the case of multiple linear regression. Also, we discuss the use of measurement error models (Fuller, 1987) for this purpose, and show the asymptotic equivalence of our procedure with the measurement error model approach. In Chapter 4, we document a simulation study to compare our approach with that of Van den Hout and Kooiman. Chapter 5 concludes.

1.2 Preliminaries

1.2.1 Notation

For convenience, we adhere to Fuller's (1987) notation whereby we reserve lower-case letters (x_t) for variables that are measured without error, and upper-case letters (X_t) for observed variables. If these letters are in boldface, they denote row vectors. In order to write models in the usual regression form

$$y_t = \beta_0 + \mathbf{x}_t\boldsymbol{\beta}_1 + \epsilon_t, \qquad t = 1, 2, \ldots, n,$$

we let $\mathbf{x}_t$ and $\boldsymbol{\beta}_1$ be r dimensional row and column vectors respectively.

The bold $\boldsymbol{\Sigma}_{ZZ}$ will denote the covariance matrix of the column $\mathbf{Z}'$, while the lower-case letter $\mathbf{m}$, appropriately subscripted, is used for the sample covariance. For example,

$$\mathbf{m}_{ZZ} = (n-1)^{-1}\sum_{t=1}^{n}(\mathbf{Z}_t - \bar{\mathbf{Z}})'(\mathbf{Z}_t - \bar{\mathbf{Z}}),$$

where $\mathbf{Z}_t = (Z_{t,1}, Z_{t,2}, \ldots, Z_{t,r})$ and $\bar{\mathbf{Z}} = n^{-1}\sum_{t=1}^{n}\mathbf{Z}_t$. Upper-case $\mathbf{M}$, appropriately subscripted, is used for the matrix of centred mean squares and cross products. Thus,

$$\mathbf{M}_{ZZ} = n^{-1} \sum_{t=1}^{n} (\mathbf{Z}_t - \bar{\mathbf{Z}})'(\mathbf{Z}_t - \bar{\mathbf{Z}}).$$

In the later part of this thesis, we also adopt the following representation of $\mathbf{M}_{ZZ}$, which we define below

$$\mathbf{M}_{ZZ} = (h_{i,i_*})_{r \times r}, \quad h_{i,i_*} = \overline{Z_i Z_{i_*}} - \bar{Z}_i \bar{Z}_{i_*},$$

where $\overline{Z_i Z_{i_*}} = n^{-1} \sum_{t=1}^{n} Z_{t,i} Z_{t,i_*}$, $\bar{Z}_i = n^{-1} \sum_{t=1}^{n} Z_{t,i}$ and $i, i_* \in (1, 2, \ldots, r)$.

1.2.2 Warner's Randomized Response Model

Throughout this thesis, we restrict our discussion to the randomization scheme of Warner (1965), which we briefly describe as follows. Suppose that every element in a population belongs to one of two disjoint groups (1 or 0), and we are interested in estimating π, the proportion of elements belonging to group 1. A simple random sample of size n is drawn with replacement from the population, and each respondent is furnished with a spinner that points to the number 1 with probability p and to the number 0 with probability $(1 - p)$. During the interview, the respondent is asked to spin the spinner, unobserved by the interviewer, and is required to report whether he/she belongs to the group indicated by the spinner. In such a survey, only yes or no responses are recorded by the interviewer. Assuming that respondents cooperate fully with the design, it is relatively straightforward to derive maximum likelihood or moment estimates for π.

Let

$$x_t = \begin{cases} 1 & \text{if the } t\text{th element belongs to group 1;} \\ 0 & \text{otherwise,} \end{cases} \tag{1.1}$$

$$X_t = \begin{cases} 1 & \text{if the } t\text{th element says yes;} \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

In this setting, note that x_t is latent and only X_t is observed. Hence, we have

$$P(x_t = 1) = \pi,$$

$$P(X_t = 1 \mid x_t = 1) = P(X_t = 0 \mid x_t = 0) = p, \tag{1.3}$$

$$P(X_t = 1 \mid x_t = 0) = P(X_t = 0 \mid x_t = 1) = 1 - p, \tag{1.4}$$

from which we can derive the probability of a yes response

$$P(X_t = 1) = \pi p + (1 - \pi)(1 - p).$$

As a result, by noting that $E(X_t) = P(X_t = 1)$ and $E(x_t) = \pi$, for $p \neq \frac{1}{2}$, we have the unbiased moment estimator for the latent moment $E(x_t)$ as follows

$$\frac{\bar{X} - (1 - p)}{2p - 1}, \tag{1.5}$$

where $\bar{X} = n^{-1} \sum_{t=1}^{n} X_t$ is the sample proportion of yes responses. For the case where $p = \frac{1}{2}$, no useful information can be gleaned from an application of Warner's procedure as this situation is akin to the respondent giving random yes/no responses.

Chapter 2

Estimating Subgroup Means in a Dichotomous Population

In this chapter, we examine the problem of obtaining estimates for the subgroup means of a response variable in a dichotomous population. Assuming that the allocation of elements to the subgroups is subjected to the RR design described in Section 1.2.2, we consider two estimators that take into account this information.

2.1 Selén's Estimator

For a population consisting of k disjoint subgroups, where each element belongs to one and only one subgroup, Selén (1986) proposed a method of adjusting the subgroup means when there are errors in the classification of elements to their subgroups. Selén's estimator is a moment estimate which essentially consists of obtaining linear combinations of the averages of the *recorded* subgroups to offset the bias introduced due to misclassification. In this section, we consider the case of a dichotomous population, where it is of interest to estimate the subgroup means

(μ_0 and μ_1) of a response variable y. It is assumed that the sample is obtained through simple random sampling, and that the response y_t for the tth element is measured without error.

We denote x_t to be the true class of the tth element as defined in (1.1) and x_t^r its recorded class. That is,

$$x_t^r = \begin{cases} 1 & \text{if the } t\text{th element is classified to subgroup 1;} \\ 0 & \text{otherwise.} \end{cases}$$

Also, let $\bar{\mathbf{y}}^r = (\bar{y}_0^r, \bar{y}_1^r)$ be the vector of averages of the recorded subgroups, where

$$\bar{y}_0^r = \frac{\sum_{t=1}^n y_t(1 - x_t^r)}{\sum_{t=1}^n (1 - x_t^r)},$$

and

$$\bar{y}_1^r = \frac{\sum_{t=1}^n y_t x_t^r}{\sum_{t=1}^n x_t^r}.$$

The probability that an element is classified to subgroup j given that it belongs to subgroup i is denoted by $P(x_t^r = j \mid x_t = i) = p_{ij}$. The matrix of classification probabilities can then be represented by

$$\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix},$$

where $p_{01} = 1 - p_{00}$ and $p_{10} = 1 - p_{11}$. If we denote $\hat{\boldsymbol{\mu}}^S = (\hat{\mu}_0^S, \hat{\mu}_1^S)$, Selén's estimator is then given as

$$\hat{\boldsymbol{\mu}}^S = \bar{\mathbf{y}}^r \left[diag(\boldsymbol{\pi})\mathbf{P}' diag(\mathbf{P}\boldsymbol{\pi}')^{-1} \right]^{-1}, \tag{2.1}$$

where $\boldsymbol{\pi} = (1 - \pi, \pi)$ and $diag$ is the diagonalization operator. See Selén (1986).

Here, the matrix $[diag(\boldsymbol{\pi})\mathbf{P}'diag(\mathbf{P}\boldsymbol{\pi}')^{-1}]^{-1}$ can be regarded as an adjustment term to account for the bias introduced by misclassification. In expanded form, the adjusted estimates for the subgroup means is a weighted average of the averages of the recorded subgroups

$$\begin{aligned} \hat{\mu}_0^S &= \lambda\bar{y}_0^r + (1-\lambda)\bar{y}_1^r, && (2.2) \\ \hat{\mu}_1^S &= \nu\bar{y}_1^r + (1-\nu)\bar{y}_0^r, && (2.3) \end{aligned}$$

where $\lambda = \frac{p_{11}[p_{00}(1-\pi)+\pi(1-p_{00})]}{(1-\pi)[p_{00}p_{11}-(1-p_{00})(1-p_{11})]}$, $\nu = \frac{p_{00}[\pi p_{11}-(1-\pi)(1-p_{11})]}{\pi[p_{00}p_{11}-(1-p_{00})(1-p_{11})]}$ and $p_{00}+p_{11} \neq 1$.

The problem with these estimators is that $\mathbf{P}$ and π are normally unknown. If the classification device mentioned above is assumed to be the RR design in Section 1.2.2, we can use (1.5) to estimate π. Furthermore, we can think of the RR design as a classification device which classifies elements into yes or no groups. Thus $\mathbf{P}$ is also known to us, since we have $p_{00} = p_{11} = p$. As a result, by denoting

$$\bar{y}_1^* = \frac{\sum_{t=1}^n y_t X_t}{\sum_{t=1}^n X_t},$$

and

$$\bar{y}_0^* = \frac{\sum_{t=1}^n y_t(1-X_t)}{\sum_{t=1}^n (1-X_t)},$$

to be the averages of the yes and no groups respectively, (2.2) and (2.3) becomes

$$\begin{aligned} \hat{\mu}_0^S &= \hat{\lambda}\bar{y}_0^* + (1-\hat{\lambda})\bar{y}_1^*, \\ \hat{\mu}_1^S &= \hat{\nu}\bar{y}_1^* + (1-\hat{\nu})\bar{y}_0^*, \end{aligned}$$

where $\hat{\lambda} = \frac{p[p(1-\hat{\pi})+\hat{\pi}(1-p)]}{(1-\hat{\pi})(2p-1)}$, $\hat{\nu} = \frac{p[\hat{\pi}p-(1-\hat{\pi})(1-p)]}{\hat{\pi}(2p-1)}$ and $p \neq \frac{1}{2}$. This reduces to

$$\hat{\mu}_0^S = \frac{p\bar{y} - \overline{yX}}{p - \bar{X}}, \tag{2.4}$$

$$\hat{\mu}_1^S = \frac{(1-p)\bar{y} - \overline{yX}}{1 - p - \bar{X}}. \tag{2.5}$$

In the following section, we consider an alternative approach to the above problem.

2.2 Moment Substitution

In this section, we consider the previous problem from a sampling perspective. For a population of size N, when there are no errors in the classification process and with equal weights assigned to each x_t, the subpopulation means of group 1 and 0 are defined as

$$\mu_1 = \frac{\sum_{t=1}^N y_t x_t}{\sum_{t=1}^N x_t} = \frac{E(y_t x_t)}{E(x_t)} \tag{2.6}$$

and

$$\mu_0 = \frac{\sum_{t=1}^N y_t(1-x_t)}{\sum_{t=1}^N (1-x_t)} = \frac{E(y_t) - E(y_t x_t)}{1 - E(x_t)} \tag{2.7}$$

respectively, where $E(w) = N^{-1}\sum_{t=1}^N w_t$ is the population mean of w. For a simple random sample of size n drawn with replacement, the sampling estimators are

$$\hat{\mu}_1 = \frac{\sum_{t=1}^n y_t x_t}{\sum_{t=1}^n x_t} = \frac{\overline{yx}}{\bar{x}}, \tag{2.8}$$

and

$$\hat{\mu}_0 = \frac{\sum_{t=1}^{n} y_t(1-x_t)}{\sum_{t=1}^{n}(1-x_t)} = \frac{\bar{y} - \overline{yx}}{1-\bar{x}}. \tag{2.9}$$

Alternatively, this problem can be formulated in the form of a simple linear regression model

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t, \qquad t = 1, 2, \ldots, n. \tag{2.10}$$

In addition to the usual assumptions for the simple linear regression model, we further assume that the mean of errors in each subpopulation is 0, i.e. $E(\epsilon_t \mid x_t = 1) = E(\epsilon_t \mid x_t = 0) = 0$. Hence, the subpopulation means are

$$\begin{aligned} \mu_1 &= E(y_t \mid x_t = 1) = \beta_0 + \beta_1, && (2.11) \\ \mu_0 &= E(y_t \mid x_t = 0) = \beta_0, && (2.12) \end{aligned}$$

while the least squares estimators for β_1 and β_0 are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sum_{t=1}^{n}(x_t - \bar{x})^2}, && (2.13) \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}. && (2.14) \end{aligned}$$

Since x_t takes on values 1 and 0, (2.13) reduces to

$$\hat{\beta}_1 = \frac{\overline{yx} - \bar{y}\bar{x}}{\bar{x}(1-\bar{x})},$$

and the regression estimates for μ_1 and μ_0 in (2.11) and (2.12) eventually reduce to the sampling estimates of (2.8) and (2.9) respectively.

However, since x_t is latent in an RR design, the estimators in (2.8) and (2.9) do not apply as we cannot directly use the observed sample means of X_t and Y_tX_t to estimate the expectations in (2.6) and (2.7). Therefore, alternative estimators need to be considered. To this end, as remarked earlier, since $E(x_t) = \pi$, a moment estimator for $E(x_t)$ is given by (1.5). For $E(y_tx_t)$, an estimator in terms of X_t is presented as follows.

Proposition 1 *Let X_t be the randomized response of x_t according to the process in Section 1.2.2 for $t = 1, 2, \ldots, n$. If y_t and X_t are conditionally independent on x_t, an unbiased moment estimator for $E(y_tx_t)$ is*

$$\frac{\overline{yX} - (1-p)\bar{y}}{2p-1}. \tag{2.15}$$

Proof. See Appendix A.1.

Using the results of (1.5) and Proposition 1, the moment substitution estimators for μ_1 and μ_0 in (2.6) and (2.7) are

$$\hat{\mu}_1^M = \frac{\overline{yX} - (1-p)\bar{y}}{\bar{X} - (1-p)} = \frac{(1-p)\bar{y} - \overline{yX}}{1 - p - \bar{X}}$$

and

$$\hat{\mu}_0^M = \frac{\bar{y} - \frac{\overline{yX} - (1-p)\bar{y}}{2p-1}}{1 - \frac{\bar{X} - (1-p)}{2p-1}} = \frac{p\bar{y} - \overline{yX}}{p - \bar{X}},$$

which are identical to Selén's estimators in (2.4) and (2.5). This suggests that the above moment substitution approach can be adapted to a more general setting. Hence, in the next chapter, we apply this approach to the fitting of multiple linear regression models with RR covariates.

Chapter 3

RR Covariates in Multiple Linear Regression

In this chapter, we propose an alternative approach to Van den Hout and Kooiman (2006) for estimating the linear regression model with RR categorical covariates. In their paper, Van den Hout and Kooiman considered a normal linear regression model where some or all of the covariates are subjected to RR. Specifically, for $t = 1, 2, \ldots, n$, the model considered is

$$\begin{aligned} y_t &= \beta_0 + \beta_1 x_{t,1} + \cdots + \beta_r x_{t,r} + \beta_{r+1} x_{t,r+1} + \cdots + \beta_k x_{t,k} + \epsilon_t \\ &= \beta_0 + \mathbf{x}_t \boldsymbol{\beta}_1 + \epsilon_t, \end{aligned} \tag{3.1}$$

where $\mathbf{x}_t = (\mathbf{x}_{t\circ 1}, \mathbf{x}_{t\circ 2})$ and ϵ_t is $N(0, \sigma^2)$. Here, $\mathbf{x}_{t\circ 1} = (x_{t,1}, x_{t,2}, \ldots, x_{t,r})$ is the row vector of continuous covariates which are measured without error; on the other hand, $\mathbf{x}_{t\circ 2} = (x_{t,r+1}, x_{t,r+2}, \ldots, x_{t,k})$ is the row vector of categorical covariates subject to RR. To implement this model, Van den Hout and Kooiman derived the likelihood function of (3.1) and obtained the associated parameter estimates by maximising this likelihood function with an EM algorithm. However, as an EM algorithm can potentially be a computational burden as the sample size gets large, we extend the moment substitution approach discussed in the previous chapter to (3.1).

3.1 Moment Substitution continued

Following the previous chapter, we limit our discussion to the case where $x_{t,j}$, for $j = r+1, r+2, \ldots, k$, are binary categorical variables subject to Warner's RR design in Section 1.2.2. In particular, we consider a general scenario where

$$\begin{aligned}
P(x_{t,j} = 1) &= \pi_j, \\
P(X_{t,j} = 1 \mid x_{t,j} = 1) &= P(X_{t,j} = 0 \mid x_{t,j} = 0) = p_j, \\
P(X_{t,j} = 1 \mid x_{t,j} = 0) &= P(X_{t,j} = 0 \mid x_{t,j} = 1) = 1 - p_j, \\
P(X_{t,j} = 1) &= \pi_j p_j + (1-\pi_j)(1-p_j).
\end{aligned}$$

On top of the usual regression assumptions applied to (3.1), we also assume that:

1. RR is independently applied to each $x_{t,j}$ for $j = r+1, r+2, \ldots, k$, i.e., for $\mathbf{X}_{t\circ 2} = (X_{t,r+1}, X_{t,r+2}, \ldots, X_{t,k})$ and $\mathbf{x}_{t\circ 2} = (x_{t,r+1}, x_{t,r+2}, \ldots, x_{t,k})$,

$$P\Big(\mathbf{X}_{t\circ 2} = (X^*_{t,r+1}, X^*_{t,r+2}, \ldots, X^*_{t,k}) \mid \mathbf{x}_{t\circ 2} = (x^*_{t,r+1}, x^*_{t,r+2}, \ldots, x^*_{t,k})\Big) = \prod_{j=r+1}^{k} P(X_{t,j} = X^*_{t,j} \mid x_{t,j} = x^*_{t,j}).$$

2. $E(\epsilon_t \mid \mathbf{x}_t) = 0$.

3. $X_{t,j}$ is conditionally independent of y_t given $x_{t,j}$ for $j \in (r+1, r+2, \ldots, k)$.

4. $X_{t,j}$ is conditionally independent of $x_{t,i}$ given $x_{t,j}$ for $i \in (1, 2, \ldots, r)$ and $j \in (r+1, r+2, \ldots, k)$.

Also, we use the centred regression model instead of the model in (3.1), where given a sample size of n, we have

$$\bar{y} = \beta_0 + \beta_1\bar{x}_1 + \cdots + \beta_r\bar{x}_r + \beta_{r+1}\bar{x}_{r+1} + \cdots + \beta_k\bar{x}_k + \bar{\epsilon}.$$

Subtracting this from (3.1), we arrive at

$$y_t - \bar{y} = \beta_1(x_{t,1} - \bar{x}_1) + \beta_2(x_{t,2} - \bar{x}_2) + \cdots + \beta_k(x_{t,k} - \bar{x}_k) + \epsilon_t^*, \tag{3.2}$$

where $\epsilon_t^* = \epsilon_t - \bar{\epsilon}$ for $t = 1, 2, \ldots, n$. We write these n equations in matrix form as

$$\begin{pmatrix} y_1 - \bar{y} \\ y_2 - \bar{y} \\ \vdots \\ y_n - \bar{y} \end{pmatrix} = \begin{pmatrix} x_{1,1} - \bar{x}_1 & x_{1,2} - \bar{x}_2 & \cdots & x_{1,k} - \bar{x}_k \\ x_{2,1} - \bar{x}_1 & x_{2,2} - \bar{x}_2 & \cdots & x_{2,k} - \bar{x}_k \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} - \bar{x}_1 & x_{n,2} - \bar{x}_2 & \cdots & x_{n,k} - \bar{x}_k \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1^* \\ \epsilon_2^* \\ \vdots \\ \epsilon_n^* \end{pmatrix}$$

or

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}_c\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}^*. \tag{3.3}$$

The least squares estimators for $\boldsymbol{\beta}_1$ and β_0 are then

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1 &= (\tilde{\mathbf{X}}_c'\tilde{\mathbf{X}}_c)^{-1}\tilde{\mathbf{X}}_c'\tilde{\mathbf{y}} = \mathbf{M}_{xx}^{-1}\mathbf{M}_{xy}, \\ \hat{\beta}_0 &= \bar{y} - \bar{\mathbf{x}}\hat{\boldsymbol{\beta}}_1. \end{aligned} \tag{3.4}$$

However, as before, we do not observe all the entries in $\mathbf{M}_{xx}$, $\mathbf{M}_{xy}$ and $\bar{\mathbf{x}}$. Consider a partition of $\mathbf{M}_{xx}$ below

$$\mathbf{M}_{xx} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}, \tag{3.5}$$

where $\mathbf{A}_{11} = (a_{i,i_*})_{r \times r}$, $a_{i,i_*} = \overline{x_i x_{i_*}} - \bar{x}_i \bar{x}_{i_*}$ for $i, i_* \in (1, 2, \ldots, r)$; $\mathbf{A}_{12} = (a_{i,j})_{r \times (k-r)}$, $a_{i,j} = \overline{x_i x_j} - \bar{x}_i \bar{x}_j$ for $i \in (1, 2, \ldots, r)$ and $j \in (r+1, r+2, \ldots, k)$; $\mathbf{A}_{22} = (a_{j,j_*})_{(k-r) \times (k-r)}$, $a_{j,j_*} = \overline{x_j x_{j_*}} - \bar{x}_j \bar{x}_{j_*}$ for $j, j_* \in (r+1, r+2, \ldots, k)$ and $\mathbf{A}_{21} = \mathbf{A}'_{12}$. Similarly, we also consider partitions of $\mathbf{M}_{xy}$ and $\bar{\mathbf{x}}$ as follows

$$\mathbf{M}_{xy} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \end{pmatrix}, \quad \bar{\mathbf{x}} = (\mathbf{C}_1, \mathbf{C}_2), \tag{3.6}$$

where $\mathbf{B}_1 = (b_i)_{r \times 1}$, $b_i = \overline{y x_i} - \bar{y} \bar{x}_i$ for $i \in (1, 2, \ldots, r)$; $\mathbf{B}_2 = (b_j)_{(k-r) \times 1}$, $b_j = \overline{y x_j} - \bar{y} \bar{x}_j$ for $j \in (r+1, r+2, \ldots, k)$; $\mathbf{C}_1 = (c_i)_{1 \times r}$, $c_i = \bar{x}_i$ for $i \in (1, 2, \ldots, r)$ and $\mathbf{C}_2 = (c_j)_{1 \times (k-r)}$, $c_j = \bar{x}_j$ for $j \in (r+1, r+2, \ldots, k)$.

In this problem, only $\mathbf{A}_{11}$, $\mathbf{B}_1$ and $\mathbf{C}_1$ are observed, while the rest of the entries in $\mathbf{A}_{12}$, $\mathbf{A}_{22}$, $\mathbf{B}_2$ and $\mathbf{C}_2$ are latent and need to be estimated. From (1.5) and Proposition 1, we can readily get estimates of $\mathbf{B}_2$ and $\mathbf{C}_2$, which we denote by $\widehat{\mathbf{B}}_2$ and $\widehat{\mathbf{C}}_2$ respectively, as follows

$$\begin{aligned} \hat{b}_j &= \frac{\overline{y X_j} - \bar{y} \bar{X}_j}{2p_j - 1}, \\ \hat{c}_j &= \frac{\bar{X}_j - (1 - p_j)}{2p_j - 1}. \end{aligned}$$

To estimate $\mathbf{A}_{12}$ and $\mathbf{A}_{22}$, we present the following results.

Proposition 2 *Let $X_{t,j}$ be the randomized response of $x_{t,j}$ according to the process in Section 1.2.2 for $t = 1, 2, \ldots, n$. If $X_{t,j}$ and $x_{t,i}$ are conditionally independent on $x_{t,j}$ for $i \in (1, 2, \ldots, r)$ and $j \in (r+1, r+2, \ldots, k)$, an unbiased moment estimator for $E(x_{t,i} x_{t,j})$ is*

$$\frac{\overline{x_i X_j} - (1 - p_j) \bar{x}_i}{2p_j - 1}. \tag{3.7}$$

Proof. *See Appendix A.2.*

Proposition 3 *Let $X_{t,j}$ be the randomized response of $x_{t,j}$ according to the process in Section 1.2.2 for $t = 1, 2, \ldots, n$. If RR is independently applied to $x_{t,j}$ and x_{t,j_*} for $j, j_* \in (r+1, r+2, \ldots, k)$ and $j \neq j_*$, an unbiased moment estimator for $E(x_{t,j}x_{t,j_*})$ is*

$$\frac{\overline{X_j X_{j_*}} - (1-p_j)\bar{X}_{j_*} - (1-p_{j_*})\bar{X}_j + (1-p_j)(1-p_{j_*})}{(2p_j - 1)(2p_{j_*} - 1)}. \tag{3.8}$$

For $j = j_*$, an unbiased moment estimator is given in (1.5).

Proof. See Appendix A.3.

Hence, from (1.5) and Proposition 2, we have $\widehat{\mathbf{A}}_{12} = (\hat{a}_{i,j})_{r\times(k-r)}$, where

$$\hat{a}_{i,j} = \frac{\overline{x_i X_j} - \bar{x}_i \bar{X}_j}{2p_j - 1}.$$

To estimate $\mathbf{A}_{22}$, from (1.5) and Proposition 3, we have $\widehat{\mathbf{A}}_{22} = (\hat{a}_{j,j_*})_{(k-r)\times(k-r)}$, where

$$\hat{a}_{j,j_*} = \begin{cases} \frac{\overline{X_j X_{j_*}} - \bar{X}_j \bar{X}_{j_*}}{(2p_j-1)(2p_{j_*}-1)} & \text{if } j \neq j_*; \\ \frac{\bar{X}_j(1-\bar{X}_j) - p_j(1-p_j)}{(2p_j-1)^2} & \text{if } j = j_*. \end{cases}$$

As a result, we have the moment substitution estimators for $\boldsymbol{\beta}_1$ and β_0 as follows

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1^M &= \widehat{\mathbf{M}}_{xx}^{-1}\widehat{\mathbf{M}}_{xy}, \\ \hat{\beta}_0^M &= \bar{y} - \hat{\bar{\mathbf{x}}}\hat{\boldsymbol{\beta}}_1^M, \end{aligned} \tag{3.9}$$

where

$$\widehat{\mathbf{M}}_{xx} = \begin{pmatrix} \mathbf{A}_{11} & \widehat{\mathbf{A}}_{12} \\ \widehat{\mathbf{A}}'_{12} & \widehat{\mathbf{A}}_{22} \end{pmatrix}, \quad \widehat{\mathbf{M}}_{xy} = \begin{pmatrix} \mathbf{B}_1 \\ \widehat{\mathbf{B}}_2 \end{pmatrix}, \quad \hat{\bar{\mathbf{x}}} = (\mathbf{C}_1, \widehat{\mathbf{C}}_2). \tag{3.10}$$

In the next section, we consider the use of measurement error models to fit the model in (3.1).

3.2 Measurement Error Models

A measurement error model is a regression model with substantial measurement errors in the variables. An example of such models is as follows, where

$$y_t = \beta_0 + \mathbf{x}_t\boldsymbol{\beta}_1 + \epsilon_t, \quad \mathbf{X}_t = \mathbf{x}_t + \mathbf{e}_t, \tag{3.11}$$

for $t = 1, 2, \ldots, n$. The first equation of (3.11) is a classical regression specification, but the true explanatory variables $\mathbf{x}_t$ are not observed directly; instead, the measurement $\mathbf{X}_t$ is observed. In this setting, Fuller (1987) considered maximum likelihood estimation for the normal multiple linear regression model, where it is assumed that

$$\begin{bmatrix} \mathbf{x}_t' \\ \epsilon_t \\ \mathbf{e}_t' \end{bmatrix} \sim N\left(\begin{bmatrix} \boldsymbol{\mu}_x' \\ 0 \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{xx} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_{\epsilon\epsilon} & \boldsymbol{\Sigma}_{\epsilon e} \\ \mathbf{0} & \boldsymbol{\Sigma}_{e\epsilon} & \boldsymbol{\Sigma}_{ee} \end{bmatrix} \right)$$

and both $\boldsymbol{\Sigma}_{ee}$ and $\boldsymbol{\Sigma}_{e\epsilon}$ are known. The maximum likelihood estimators are then given by

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1^F &= (\mathbf{m}_{XX} - \boldsymbol{\Sigma}_{ee})^{-1}(\mathbf{m}_{Xy} - \boldsymbol{\Sigma}_{e\epsilon}), \\ \hat{\beta}_0^F &= \bar{y} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}_1^F. \end{aligned} \tag{3.12}$$

For the likelihood function which led to the above estimators, see Fuller (1987). These estimators were shown by Fuller to be asymptotically unbiased.

3.2.1 Adaptation for RR covariates

As mentioned earlier, RR covariates can be looked upon as misclassified categorical variables whose misclassification probabilities are known; in the light of measurement error models, these can also be viewed as variables with measurement errors, which suggests that Fuller's afore-mentioned estimators can be adapted to models with RR covariates. To this end, we write (3.1) in the form of (3.11), where $\boldsymbol{\beta}_1$ is a k dimensional column vector and $\mathbf{e}_t = (\mathbf{0}_{1\times r}, e_{t,r+1}, \ldots, e_{t,k})$. To evaluate $\boldsymbol{\Sigma}_{e\epsilon}$, we present the following result.

Proposition 4 *Let $X_{t,j}$ be the randomized response of $x_{t,j}$ according to the process in Section 1.2.2 for $t = 1, 2, \ldots, n$. If $X_{t,j}$ is conditionally independent of y_t given $x_{t,j}$ for $j \in (r+1, r+2, \ldots, k)$, then*

$$cov(\epsilon_t, e_{t,j}) = 0.$$

Proof. See Appendix A.4.

Using this result, we can easily verify that $\boldsymbol{\Sigma}_{e\epsilon} = \mathbf{0}_{k\times 1}$. The remaining task is to find $\boldsymbol{\Sigma}_{ee}$.

However, before evaluating $\boldsymbol{\Sigma}_{ee}$, we observe that when $X_{t,j}$ is a RR variable for $j = r+1, r+2, \ldots, k$, we have in general,

$$\begin{aligned} E(e_{t,j}) &= E(X_{t,j} - x_{t,j}) \\ &= p_j\pi_j + (1-p_j)(1-\pi_j) - \pi_j \\ &= (1-p_j)(1-2\pi_j) \\ &\neq 0. \end{aligned}$$

To circumvent this problem, we introduce the following transformation on $X_{t,j}$, where we define

$$W_{t,j} = \frac{X_{t,j} - (1 - p_j)}{2p_j - 1} = x_{t,j} + u_{t,j}.$$

Under this transformation, (3.11) becomes

$$y_t = \beta_0 + \mathbf{x}_t\boldsymbol{\beta}_1 + \epsilon_t, \quad \mathbf{W}_t = \mathbf{x}_t + \mathbf{u}_t, \tag{3.13}$$

where $\mathbf{u}_t = (\mathbf{0}_{1\times r}, u_{t,r+1}, \ldots, u_{t,k})$ and we have $E(u_{t,j}) = E(u_{t,j} \mid x_{t,j}) = 0$ as required. Hence, when RR covariates are included in the model, Fuller's estimators for $\boldsymbol{\beta}_1$ and β_0 are

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1^F &= (\mathbf{m}_{WW} - \boldsymbol{\Sigma}_{uu})^{-1}\mathbf{m}_{Wy}, \\ \hat{\beta}_0^F &= \bar{y} - \bar{\mathbf{W}}\hat{\boldsymbol{\beta}}_1^F. \end{aligned} \tag{3.14}$$

Furthermore, we note that since $x_{t,i}$ is measured without error for $i = 1, 2, \ldots, r$, $\boldsymbol{\Sigma}_{uu}$ is of the following form

$$\boldsymbol{\Sigma}_{uu} = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}^* \end{pmatrix},$$

where $\boldsymbol{\Sigma}^*$ is the covariance matrix of the vector $(u_{t,r+1}, u_{t,r+2}, \ldots, u_{t,k})$. Hence, to evaluate $\boldsymbol{\Sigma}^*$, we have the following result.

Proposition 5 *For $j, j_* \in (r + 1, r + 2, \ldots, k)$,*

1. $var(u_{t,j}) = \dfrac{p_j(1 - p_j)}{(2p_j - 1)^2}.$

2. For $j \neq j_*$, if RR is independently applied to $x_{t,j}$ and x_{t,j_*},

$$cov(u_{t,j}, u_{t,j_*}) = 0.$$

Proof. See Appendix A.5.

From Proposition 5, we thus have

$$\mathbf{\Sigma}^* = diag\biggl(\frac{p_{r+1}(1-p_{r+1})}{(2p_{r+1}-1)^2}, \frac{p_{r+2}(1-p_{r+2})}{(2p_{r+2}-1)^2}, \ldots, \frac{p_k(1-p_k)}{(2p_k-1)^2}\biggr)$$

as required. In the next section, we show that Fuller's measurement error model estimates for linear regression models with RR covariates are asymptotically equivalent to that of the moment substitution approach mentioned in the preceding section.

3.2.2 Asymptotic Equivalence with Moment Substitution

To show the asymptotic equivalence of Fuller's estimates with that of our proposed moment substitution approach, consider the following modification of Fuller's estimators in (3.14), where

$$\begin{aligned} \hat{\boldsymbol{\beta}}_1^{F*} &= (\mathbf{M}_{WW} - \mathbf{\Sigma}_{uu})^{-1}\mathbf{M}_{Wy}, \\ \hat{\beta}_0^{F*} &= \bar{y} - \bar{\mathbf{W}}\hat{\boldsymbol{\beta}}_1^{F*}. \end{aligned} \tag{3.15}$$

To show the desired result, let us consider a partition of the matrix $\mathbf{M}_{WW} - \mathbf{\Sigma}_{uu}$ below

$$\mathbf{M}_{WW} - \mathbf{\Sigma}_{uu} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12}^* \\ \mathbf{A}_{21}^* & \mathbf{A}_{22}^* \end{pmatrix}, \tag{3.16}$$

where $\mathbf{A}_{11}$ is as defined in (3.5). Note that for $W_{t,j} = \frac{X_{t,j}-(1-p_j)}{2p_j-1}$, we have $\bar{W}_j = \frac{\bar{X}_j-(1-p_j)}{2p_j-1}$ and $W_{t,j} - \bar{W}_j = \frac{X_{t,j}-\bar{X}_j}{2p_j-1}$. Thus, we have $\mathbf{A}^*_{12} = (a^*_{i,j})_{r\times(k-r)}$, where

$$a^*_{i,j} = \frac{\overline{x_i X_j} - \bar{x}_i \bar{X}_j}{2p_j - 1} = \hat{a}_{i,j},$$

for $i \in (1, 2, \ldots, r)$ and $j \in (r+1, r+2, \ldots, k)$; $\mathbf{A}^*_{22} = (a^*_{j,j_*})_{(k-r)\times(k-r)}$, where

$$a^*_{j,j_*} = \left\{ \begin{array}{ll} \frac{\overline{X_j X_{j_*}} - \bar{X}_j \bar{X}_{j_*}}{(2p_j-1)(2p_{j_*}-1)} & \text{for } j \neq j_*, \\ \frac{\bar{X}_j(1-\bar{X}_j)-p_j(1-p_j)}{(2p_j-1)^2} & \text{for } j = j_*, \end{array} \right\} = \hat{a}_{j,j_*}$$

for $j, j_* \in (r+1, r+2, \ldots, k)$. Also, $\mathbf{A}^*_{21} = \mathbf{A}^{*\prime}_{12}$. Thus, from the above, we can easily see that $\mathbf{A}^*_{12} = \widehat{\mathbf{A}}_{12}$ and $\mathbf{A}^*_{22} = \widehat{\mathbf{A}}_{22}$.

Similarly, we also consider partitions of $\mathbf{M}_{Wy}$ and $\bar{\mathbf{W}}$ as follows

$$\mathbf{M}_{Wy} = \begin{pmatrix} \mathbf{B}_1 \\ \mathbf{B}^*_2 \end{pmatrix}, \quad \bar{\mathbf{W}} = (\mathbf{C}_1, \mathbf{C}^*_2), \tag{3.17}$$

where $\mathbf{B}_1$ is as defined in (3.6); $\mathbf{B}^*_2 = (b^*_j)_{(k-r)\times 1}$, where

$$b^*_j = \frac{\overline{yX_j} - \bar{y}\bar{X}_j}{2p_j - 1} = \hat{b}_j,$$

for $j \in (r+1, r+2, \ldots, k)$; $\mathbf{C}_1$ is as defined in (3.6) and $\mathbf{C}^*_2 = (c^*_j)_{1\times(k-r)}$, where

$$c^*_j = \frac{\bar{X}_j - (1-p_j)}{2p_j - 1} = \hat{c}_j,$$

for $j \in (r+1, r+2, \ldots, k)$. Once again, we can easily see that $\mathbf{B}^*_2 = \widehat{\mathbf{B}}_2$ and $\mathbf{C}^*_2 = \widehat{\mathbf{C}}_2$. As a result, recalling the results from (3.9) and (3.10), we note that

$\mathbf{M}_{WW} - \mathbf{\Sigma}_{uu} = \widehat{\mathbf{M}}_{xx}$, $\mathbf{M}_{Wy} = \widehat{\mathbf{M}}_{xy}$ and $\bar{\mathbf{W}} = \hat{\bar{\mathbf{x}}}$. Hence, we have

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1^{F*} &= \hat{\boldsymbol{\beta}}_1^{M},\\ \hat{\beta}_0^{F*} &= \hat{\beta}_0^{M}.\end{aligned}$$

Furthermore, note that

$$\mathbf{M}_{WW} = \frac{n-1}{n}\mathbf{m}_{WW}.$$

Consequently, we can express

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1^{M} &= \hat{\boldsymbol{\beta}}_1^{F*}\\ &= (\mathbf{M}_{WW} - \mathbf{\Sigma}_{uu})^{-1}\mathbf{M}_{Wy}\\ &= \left(\frac{n-1}{n}\mathbf{m}_{WW} - \mathbf{\Sigma}_{uu}\right)^{-1} \cdot \frac{n-1}{n}\mathbf{m}_{Wy}\\ &\to (\mathbf{m}_{WW} - \mathbf{\Sigma}_{uu})^{-1}\mathbf{m}_{Wy}\\ &= \hat{\boldsymbol{\beta}}_1^{F} \quad \text{as } n \to \infty.\end{aligned}$$

We can therefore conclude that the moment substitution approach is asymptotically equivalent to the measurement error model approach in (3.14). Furthermore, since measurement error model estimates are asymptotically unbiased, it follows that estimates obtained via moment substitution are also asymptotically unbiased. In the next chapter, we examine the effectiveness of this approach in fitting linear regression models with RR covariates through a simulation study.

Chapter 4

Simulation Study

In the previous chapter, we proposed a method of fitting linear regression models with RR covariates using moment substitution; also, we considered an adaptation of the measurement error model of Fuller (1987) to fit such models, and have subsequently shown that these two approaches are asymptotically equivalent. Here, we compare the afore-mentioned methods with the maximum likelihood approach of Van den Hout and Kooiman (2005) using a simulation study. Specifically, data is simulated in the programming environment R and the estimation of a linear regression model is discussed.

4.1 Simulation Setup

For this exercise, we follow the simulation scheme originally set up in Van den Hout and Kooiman (2006), where the plan was to assess the following regression model

$$E(y_t \mid \mathbf{x}_t) = \beta_0 + \beta_1 x_{t,1} + \beta_2 x_{t,2} + \beta_3 x_{t,3} = \mathbf{x}_t \boldsymbol{\beta}, \tag{4.1}$$

for $t = 1, 2, \ldots, n$. Here, $x_{t,1}$ and $x_{t,2}$ are $(1, 0)$ explanatory variables and are expected to be subject to RR. On the other hand, $x_{t,3}$ is a continuous variable and is assumed to be measured without error. The values of these covariates are chosen as follows: $0.2n$ units of $(x_{t,1}, x_{t,2}) = (0, 0)$, $0.3n$ units of $(x_{t,1}, x_{t,2}) = (0, 1)$, $0.3n$ units of $(x_{t,1}, x_{t,2}) = (1, 0)$ and $0.2n$ units of $(x_{t,1}, x_{t,2}) = (1, 1)$. This distribution of $x_{t,1}$ and $x_{t,2}$ values implicitly assume that

$$P(x_{t,j} = 1) = \pi_j = 0.5,$$

for $j = 1, 2$; in addition, $x_{t,3}$ is sampled from a normal distribution with mean 20 and $\sigma^2 = 4$. For this study, we fix $\boldsymbol{\beta} = (8, 4, 15, 8)'$. Given $\mathbf{x}_t$, we then sample y_t from a normal distribution with mean $\mathbf{x}_t\boldsymbol{\beta}$ and $\sigma^2 = 9$ for $t = 1, 2, \ldots, n$.

Next, we define the independent RR processes for $x_{t,1}$ and $x_{t,2}$ as follows

$$\begin{aligned} P(X_{t,j} = 1 \mid x_{t,j} = 1) &= P(X_{t,j} = 0 \mid x_{t,j} = 0) = p, \\ P(X_{t,j} = 1 \mid x_{t,j} = 0) &= P(X_{t,j} = 0 \mid x_{t,j} = 1) = 1 - p, \end{aligned} \tag{4.2}$$

for $j = 1, 2$. Using this setup, we compare the regression coefficients of β_1, β_2 and β_3 using average point estimates and root mean square errors (RMSE) when they are fitted under the moment substitution, measurement error model and maximum likelihood EM approaches. At this point, we point out that when $p = 1$, all three methods reduce to the usual least squares estimate for the linear regression model. Table 4.1 shows the results for various choices of n and p.

4.2 Simulation Results

Table 4.1 is obtained as follows. Given a particular choice of sample size, one simulation consists of generating a sample $y_1, y_2, \ldots, y_n$ from $\mathbf{x}_t$ as described previously and simulating the RR values of $x_{t,1}$ and $x_{t,2}$ according to the process in (4.2). Keeping $\mathbf{x}_t$ fixed, this is then repeated 5000 times to obtain the average point estimates and root mean square errors of β_1, β_2 and β_3.

From Table 4.1, the MLEM estimates perform better in terms of consistency and RMSE when compared to the MS and MEM estimates for all choices of n and p. Specifically, when both n and p are small, the MLEM estimates are much more reliable as the MS and MEM estimates show unusually large values of RMSE. For the estimates obtained by moment substitution, this may be due to the factor $(2p-1)^{-2}$ present in the variance of the individual estimates in (1.5) and Propositions 2, 3 and 5. As p approaches 0.5, the variance of these terms increases exponentially.

For the measurement error model estimates, one possible reason that may account for the inflated values of RMSE is that the RR covariates do not satisfy the normality assumption required for the application of the measurement error model technique. Fortunately, for both cases, this effect becomes less pronounced as the sample size increases. In general, we begin to lose consistency as the value of p decreases. However, for larger values of n and p, the MS and MEM estimates perform reasonably well.

Nevertheless, the results in Table 4.1 are for the case when $\pi_1 = \pi_2 = 0.5$. As RR is typically employed when the values of π_1 and π_2 are small, we repeat the simulations in Table 4.1 but for alternative values of π_1 and π_2. For these simulations, the sample size used is 1000 and the results are shown in Table 4.2.

From Table 4.2, it can be seen that the effect of reducing a particular π_j value for $j \in (1, 2)$ increases the RMSE of its corresponding regression coefficient. Furthermore, if this were to be coupled with a low p value, the reliability of the MS and MEM regression estimates is reduced. In such scenarios, the sample size will need to be larger to ensure better quality estimates.

Table 4.1: Comparison of moment substitution (MS), measurement error model (MEM) and maximum likelihood EM (MLEM) estimates for $\pi_1 = \pi_2 = 0.5$. Average point estimates for 5000 replications are given. RMSE is shown in parentheses.

n	p	Parameter	MS	MEM	MLEM
60	1	β_1	4.00 (0.79)	4.00 (0.79)	4.00 (0.79)
		β_2	15.01 (0.79)	15.01 (0.79)	15.01 (0.79)
		β_3	8.00 (0.20)	8.00 (0.20)	8.00 (0.20)
	0.9	β_1	4.33 (2.53)	4.25 (2.46)	4.07 (1.07)
		β_2	15.63 (2.22)	15.45 (2.13)	15.03 (0.92)
		β_3	8.00 (0.41)	8.00 (0.40)	8.01 (0.20)
	0.8	β_1	5.04 (79.86)	5.00 (75.25)	3.94 (1.61)
		β_2	17.25 (81.48)	17.35 (75.28)	15.04 (1.17)
		β_3	8.02 (7.39)	8.00 (3.89)	7.86 (0.30)
100	1	β_1	3.99 (0.62)	3.99 (0.62)	3.99 (0.62)
		β_2	15.00 (0.62)	15.00 (0.62)	15.00 (0.62)
		β_3	8.00 (0.14)	8.00 (0.14)	8.00 (0.14)
	0.9	β_1	4.23 (1.88)	4.18 (1.85)	4.04 (0.81)
		β_2	15.41 (1.68)	15.32 (1.64)	15.04 (0.70)
		β_3	8.00 (0.31)	8.00 (0.30)	7.99 (0.16)
	0.8	β_1	5.84 (21.34)	4.43 (62.84)	4.08 (1.09)
		β_2	17.70 (20.96)	16.11 (62.40)	15.02 (0.87)
		β_3	8.12 (1.75)	8.02 (3.71)	7.93 (0.21)
1000	1	β_1	4.00 (0.19)	4.00 (0.19)	4.00 (0.19)
		β_2	15.00 (0.19)	15.00 (0.19)	15.00 (0.19)
		β_3	8.00 (0.05)	8.00 (0.05)	8.00 (0.05)
	0.9	β_1	4.02 (0.56)	4.01 (0.56)	4.00 (0.25)
		β_2	15.04 (0.49)	15.03 (0.49)	14.99 (0.22)
		β_3	8.00 (0.10)	8.00 (0.10)	8.01 (0.05)
	0.8	β_1	4.05 (1.24)	4.03 (1.23)	4.00 (0.32)
		β_2	15.14 (0.92)	15.11 (0.91)	14.99 (0.26)
		β_3	8.00 (0.18)	8.00 (0.17)	8.02 (0.06)

Table 4.2: Comparison of moment substitution (MS), measurement error model (MEM) and maximum likelihood EM (MLEM) estimates for alternative values of π_1 and π_2. Sample size is 1000 for each of the 5000 simulations.

π_1	π_2	p	Parameter	MS	MEM	MLEM
0.5	0.1	1	β_1	4.00 (0.19)	4.00 (0.19)	4.00 (0.19)
			β_2	15.00 (0.30)	15.00 (0.30)	15.00 (0.30)
			β_3	8.00 (0.05)	8.00 (0.05)	8.00 (0.05)
		0.9	β_1	4.00 (0.53)	4.00 (0.53)	3.99 (0.24)
			β_2	15.22 (1.73)	15.20 (1.72)	15.01 (0.35)
			β_3	8.00 (0.10)	8.00 (0.10)	8.01 (0.05)
		0.8	β_1	4.02 (1.26)	4.01 (1.25)	3.99 (0.31)
			β_2	15.87 (3.64)	15.79 (3.59)	15.01 (0.42)
			β_3	8.00 (0.18)	8.00 (0.18)	8.02 (0.05)
0.1	0.1	1	β_1	3.99 (0.30)	3.99 (0.30)	3.99 (0.30)
			β_2	15.00 (0.32)	15.00 (0.32)	15.00 (0.32)
			β_3	8.00 (0.05)	8.00 (0.05)	8.00 (0.05)
		0.9	β_1	4.01 (1.11)	4.01 (1.11)	3.99 (0.46)
			β_2	15.28 (1.92)	15.25 (1.91)	15.00 (0.36)
			β_3	8.00 (0.11)	8.00 (0.11)	8.00 (0.05)
		0.8	β_1	4.02 (3.16)	4.01 (3.13)	3.99 (0.62)
			β_2	16.30 (4.50)	16.21 (4.42)	15.00 (0.41)
			β_3	8.00 (0.20)	8.00 (0.20)	8.00 (0.05)

Chapter 5

Conclusion and Suggestions for Further Study

In this thesis, we attempted to provide a more efficient method for estimating linear regression models with RR covariates. A moment substitution approach is suggested as an alternative to the maximum likelihood EM approach of Van den Hout and Kooiman. In its entirety, the method of Van den Hout and Kooiman is an effective, stable but somewhat slow procedure of estimating such models, and this problem is only exacerbated when sample sizes become large. Specifically, it was observed that computation times for the maximum likelihood method were at least three times as long as the moment substitution approach. Under such circumstances, the efficiency of the moment substitution approach is an advantage, though a trade-off in terms of the accuracy of the regression coefficients is required.

For sufficiently large samples, the measurement error models of Fuller provide a useful and convenient generalization of the moment substitution method. The asymptotic equivalence of these two methods imply that any large sample results associated with measurement error models extend to the moment substitution method as well (Fuller, 1987). Examples include expressions for the asymptotic variance of the moment substitution estimates, of which similar expressions are not readily available from the numerical method of Van den Hout and Kooiman.

In retrospect, the results in this thesis were established under the premise of Warner's RR design. However, the question of whether or not these results extend to other RR designs remains to be answered. As the variability of the moment substitution regression estimates are largely dependent on the type of RR design, it will be interesting to investigate the use of other RR designs to derive alternative moment substitution estimates in future studies. Possible candidates include the unrelated question RR model by Greenberg et al. (1969), a variation of this model by Huang (2005) and the mixture distribution technique by Kuk (1990).

Appendix A

Some Results for Chapter 2 and 3

A.1 Proof of Proposition 1

Before embarking on the proof of Proposition 1, we first work out $E(y_t)$ and $E(y_t x_t)$. From (2.11) and (2.12),

$$E(y_t \mid x_t) = \mu_1 x_t + \mu_0 (1 - x_t).$$

Hence,

$$E(y_t) = \mu_1 \pi + \mu_0 (1 - \pi). \tag{A.1}$$

Also,

$$
\begin{aligned}
E(y_t x_t) &= E(E(y_t x_t \mid x_t)) \\
&= E(x_t E(y_t \mid x_t)) \\
&= E(x_t(\mu_1 x_t + \mu_0(1 - x_t))) \\
&= E(\mu_1 x_t) \\
&= \mu_1 \pi. \qquad \text{(A.2)}
\end{aligned}
$$

Using (1.3) and (1.4), we have

$$
\begin{aligned}
E(X_t \mid x_t = 1) &= p, \\
E(X_t \mid x_t = 0) &= 1 - p.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
E(X_t \mid x_t) &= p x_t + (1 - p)(1 - x_t) \\
&= (2p - 1)x_t + 1 - p.
\end{aligned}
$$

Given that y_t and X_t are conditionally independent on x_t for $t = 1, 2, \ldots, n$,

$$
\begin{aligned}
E(y_t X_t \mid x_t) &= E(y_t \mid x_t) E(X_t \mid x_t) \\
&= [(2p - 1)x_t + 1 - p] E(y_t \mid x_t) \\
&= (2p - 1)x_t E(y_t \mid x_t) + (1 - p) E(y_t \mid x_t) \\
&= (2p - 1) E(y_t x_t \mid x_t) + (1 - p) E(y_t \mid x_t).
\end{aligned}
$$

Taking expectation on both sides,

$$
E(y_t X_t) = (2p - 1) E(y_t x_t) + (1 - p) E(y_t), \qquad \text{(A.3)}
$$

and finally, we arrive at

$$E(y_t x_t) \;=\; \frac{E(y_t X_t) - (1-p)E(y_t)}{2p-1}.$$

Let $\theta = E(y_t x_t)$. We thus have

$$\hat{\theta} \;=\; \frac{\overline{yX} - (1-p)\bar{y}}{2p-1},$$

as required. Furthermore, from (A.1), (A.2) and (A.3), we have

$$\begin{aligned} E(\hat{\theta}) &= E\left(\frac{\overline{yX} - (1-p)\bar{y}}{2p-1}\right) \\ &= \frac{E(y_t X_t) - (1-p)E(y_t)}{2p-1} \\ &= \mu_1 \pi, \end{aligned}$$

which shows that $\hat{\theta}$ is an unbiased estimator of $E(y_t x_t)$. The proof is therefore complete. ■

A.2 Proof of Proposition 2

The proof of Proposition 2 is similar to that of Proposition 1, and will therefore not be presented. ■

A.3 Proof of Proposition 3

Given that RR is independently applied to $x_{t,j}$ and x_{t,j_*}, for $j \neq j_*$, we have

$$\begin{aligned}
& E(X_{t,j}X_{t,j_*} \mid x_{t,j}, x_{t,j_*}) \\
= \; & E(X_{t,j} \mid x_{t,j})E(X_{t,j_*} \mid x_{t,j_*}) \\
= \; & \Big[p_j x_{t,j} + (1-p_j)(1-x_{t,j})\Big]\Big[p_{j_*} x_{t,j_*} + (1-p_{j_*})(1-x_{t,j_*})\Big] \\
= \; & (2p_j - 1)(2p_{j_*} - 1)x_{t,j}x_{t,j_*} + (1-p_j)(2p_{j_*} - 1)x_{t,j_*} \\
& + (1-p_{j_*})(2p_j - 1)x_{t,j} + (1-p_j)(1-p_{j_*}).
\end{aligned}$$

Taking expectation on both sides, we arrive at

$$\begin{aligned}
E(X_{t,j}X_{t,j_*}) = \; & (2p_j - 1)(2p_{j_*} - 1)E(x_{t,j}x_{t,j_*}) + (1-p_j)(2p_{j_*} - 1)E(x_{t,j_*}) \\
& + (1-p_{j_*})(2p_j - 1)E(x_{t,j}) + (1-p_j)(1-p_{j_*}), \qquad \text{(A.4)}
\end{aligned}$$

which implies that

$$\begin{aligned}
E(x_{t,j}x_{t,j_*}) = \; & \Big(E(X_{t,j}X_{t,j_*}) - (1-p_j)(2p_{j_*} - 1)E(x_{t,j_*}) \\
& - (1-p_{j_*})(2p_j - 1)E(x_{t,j}) \\
& - (1-p_j)(1-p_{j_*})\Big)\Big/(2p_j - 1)(2p_{j_*} - 1).
\end{aligned}$$

However, since $E(x_{t,j})$ and $E(x_{t,j_*})$ are also not observed, we make use of their estimators given in (1.5). Let $\gamma = E(x_{t,j}x_{t,j_*})$. Hence, an estimator for $E(x_{t,j}x_{t,j_*})$

is given by

$$\begin{aligned}
\hat{\gamma} &= \Big(\overline{X_j X_{j_*}} - (1-p_j)\Big[\bar{X}_{j_*} - (1-p_{j_*})\Big] \\
&\quad - (1-p_{j_*})\Big[\bar{X}_j - (1-p_j)\Big] \\
&\quad - (1-p_j)(1-p_{j_*})\Big) \Big/ (2p_j - 1)(2p_{j_*} - 1) \\
&= \frac{\overline{X_j X_{j_*}} - (1-p_j)\bar{X}_{j_*} - (1-p_{j_*})\bar{X}_j + (1-p_j)(1-p_{j_*})}{(2p_j-1)(2p_{j_*}-1)}
\end{aligned}$$

as required. For $j = j_*$, since $E(x_{t,j}^2) = E(x_{t,j})$, the result in (1.5) follows. Furthermore, from (A.4) and

$$E(X_{t,j}) = P(X_{t,j} = 1) = \pi_j p_j + (1-\pi_j)(1-p_j),$$

we can easily verify that $E(\hat{\gamma}) = \gamma$. The proof is therefore complete. ■

A.4 Proof of Proposition 4

If $X_{t,j}$ is the randomized response of $x_{t,j}$, for $j \in (r+1, r+2, \ldots, k)$,

$$\begin{aligned}
cov(\epsilon_t, e_{t,j}) &= cov(\epsilon_t, X_{t,j} - x_{t,j}) \\
&= cov(\epsilon_t, X_{t,j}) - cov(\epsilon_t, x_{t,j}) \\
&= E(\epsilon_t X_{t,j}) - E(\epsilon_t)E(X_{t,j}) - cov(\epsilon_t, x_{t,j}).
\end{aligned}$$

From the normality assumption of model (3.11), we have $E(\epsilon_t) = cov(\epsilon_t, x_{t,j}) = 0$. Hence,

$$cov(\epsilon_t, e_{t,j}) = E(\epsilon_t X_{t,j}).$$

To evaluate $E(\epsilon_t X_{t,j})$, we compute its conditional expectation with respect to $\mathbf{x}_t = (x_{t,1}, \ldots, x_{t,r}, x_{t,r+1}, \ldots, x_{t,k})$ as follows

$$\begin{aligned} E(\epsilon_t X_{t,j} \mid \mathbf{x}_t) &= E\Big[(y_t - \beta_0 - \mathbf{x}_t \boldsymbol{\beta}_1) X_{t,j} \mid \mathbf{x}_t\Big] \\ &= E(y_t X_{t,j} \mid \mathbf{x}_t) - (\beta_0 + \mathbf{x}_t \boldsymbol{\beta}_1) E(X_{t,j} \mid \mathbf{x}_t). \qquad \text{(A.5)} \end{aligned}$$

Since $X_{t,j}$ is conditionally independent of y_t given $x_{t,j}$ and $E(y_t \mid \mathbf{x}_t) = \beta_0 + \mathbf{x}_t \boldsymbol{\beta}_1$, (A.5) becomes

$$\begin{aligned} E(\epsilon_t X_{t,j} \mid \mathbf{x}_t) &= E(y_t \mid \mathbf{x}_t) E(X_{t,j} \mid \mathbf{x}_t) - E(y_t \mid \mathbf{x}_t) E(X_{t,j} \mid \mathbf{x}_t) \\ &= 0. \end{aligned}$$

Hence, we have $E(\epsilon_t X_{t,j}) = 0$ as required. The proof is therefore complete. ■

A.5 Proof of Proposition 5

For $j, j_* \in (r+1, r+2, \ldots, k)$, recall that

$$\begin{aligned} E(X_{t,j} \mid x_{t,j} = 1) &= p_j, \\ E(X_{t,j} \mid x_{t,j} = 0) &= 1 - p_j. \end{aligned}$$

As a result,

$$\begin{aligned} var(X_{t,j} \mid x_{t,j} = 1) &= E(X_{t,j}^2 \mid x_{t,j} = 1) - E(X_{t,j} \mid x_{t,j} = 1)^2 \\ &= E(X_{t,j} \mid x_{t,j} = 1) - E(X_{t,j} \mid x_{t,j} = 1)^2 \\ &= p_j(1 - p_j) \\ &= var(X_{t,j} \mid x_{t,j} = 0). \end{aligned}$$

Consequently, in general, $var(X_{t,j} \mid x_{t,j}) = p_j(1-p_j)$. We can thus evaluate

$$\begin{aligned} var(u_{t,j} \mid x_{t,j}) &= var(W_{t,j} - x_{t,j} \mid x_{t,j}) \\ &= var\bigg(\frac{X_{t,j} - (1-p_j)}{2p_j - 1} \mid x_{t,j}\bigg) \\ &= \frac{p_j(1-p_j)}{(2p_j-1)^2} \\ &= var(u_{t,j}) \end{aligned}$$

as required. Furthermore, if RR is independently applied to $x_{t,j}$ and x_{t,j_*}, for $j \neq j_*$,

$$\begin{aligned} & cov(u_{t,j}, u_{t,j_*} \mid x_{t,j}, x_{t,j_*}) \\ = \; & cov\bigg(\frac{X_{t,j} - (1-p_j)}{2p_j - 1}, \frac{X_{t,j_*} - (1-p_{j_*})}{2p_{j_*} - 1} \mid x_{t,j}, x_{t,j_*}\bigg) \\ = \; & \frac{cov(X_{t,j}, X_{t,j_*} \mid x_{t,j}, x_{t,j_*})}{(2p_j-1)(2p_{j_*}-1)} \\ = \; & \frac{E(X_{t,j}X_{t,j_*} \mid x_{t,j}, x_{t,j_*}) - E(X_{t,j} \mid x_{t,j}, x_{t,j_*})E(X_{t,j_*} \mid x_{t,j}, x_{t,j_*})}{(2p_j-1)(2p_{j_*}-1)} \\ = \; & \frac{E(X_{t,j} \mid x_{t,j})E(X_{t,j_*} \mid x_{t,j_*}) - E(X_{t,j} \mid x_{t,j})E(X_{t,j_*} \mid x_{t,j_*})}{(2p_j-1)(2p_{j_*}-1)} \\ = \; & 0. \end{aligned}$$

Also, since $E(u_{t,j} \mid x_{t,j}, x_{t,j_*}) = E(u_{t,j} \mid x_{t,j}) = 0$, we then have

$$\begin{aligned} cov(u_{t,j}, u_{t,j_*}) &= E\Big[cov(u_{t,j}, u_{t,j_*} \mid x_{t,j}, x_{t,j_*})\Big] \\ &\quad + cov\Big[E(u_{t,j} \mid x_{t,j}, x_{t,j_*}), E(u_{t,j} \mid x_{t,j}, x_{t,j_*})\Big] \\ &= 0, \end{aligned}$$

as required. The proof is therefore complete. ■

Bibliography

[1] Chaudhuri, A., Mukerjee, R., 1988. Randomized Response: Theory and Techniques. Marcel Dekker, New York.

[2] Fuller, W. A., 1987. Measurement Error Models. Wiley, New York.

[3] Greenberg B. G., Abul-Ela Abdel-Latif A., Simmons W. R., Horvitz D. G., 1969. The unrelated question RR model: theoretical framework. *Journal of the American Statistical Association.* 64, 520–539.

[4] Huang, K. C., 2005. Estimation of sensitive data from a dichotomous population. *Statistical Papers.* 47, 149–156.

[5] Kuk, A. Y. C., 1990. Asking Sensitive Questions Indirectly. *Biometrika.* 77, 436–438.

[6] Selén, J., 1986. Adjusting for Errors in Classification and Measurement in the Analysis of Partly and Purely Categorical Data. *Journal of the American Statistical Association.* 81, 75–81.

[7] Spiegelman, D., Rosner, B., Logan, R., 2000. Estimation and inference for logistic regression with covariate misclassification and measurement error in main study/validation study designs. *Journal of the American Statistical Association.* 95, 51–61.

[8] Van den Hout, A., Kooiman, P., 2006. Estimating the linear regression model with categorical covariates subject to randomized response. *Computational Statistics and Data Analysis.* 50, 3311–3323.

[9] Van den Hout, A., Van der Heijden, P. G. M., 2002. Randomized response, statistical disclosure control and misclassification: a review. *International Statistical Review.* 70, 269–288.

[10] Warner, S. L., 1965. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association.* 60, 63–69.