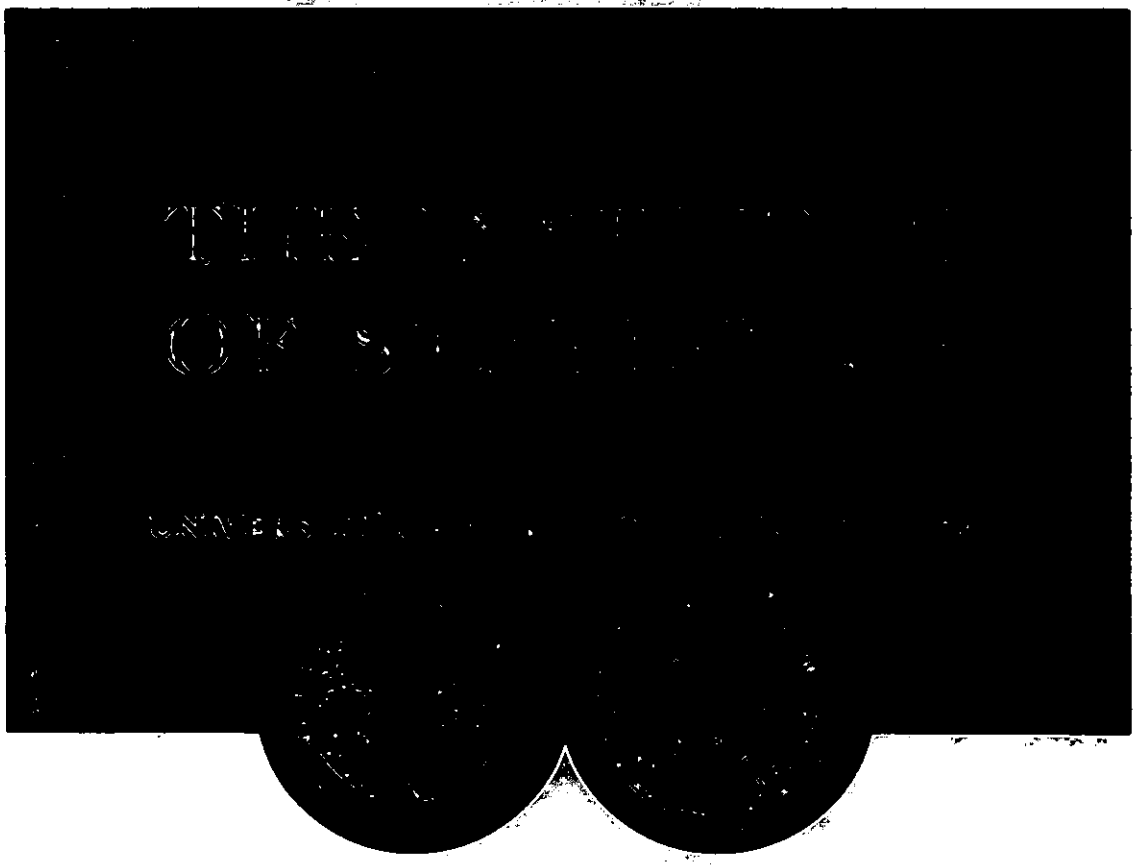

INSTRUMENTAL VARIABLE ESTIMATION IN GENERALIZED LINEAR MEASUREMENT ERROR MODELS

by

Jeffrey S. Buzas

Institute of Statistics Mimeograph Series No. 2257T

May 1993

NORTH CAROLINA STATE UNIVERSITY
Raleigh, North Carolina

ABSTRACT

BUZAS, JEFFREY SANDOR. Instrumental Variable Estimation in Generalized Linear Measurement Error Models. (Under the direction of Leonard A. Stefanski.)

The estimation of regression parameters in generalized linear models when some covariates are concealed by normally distributed measurement error is considered. It is assumed additional information in the form of instrumental variables is available.

In an effort to clarify estimation in a probit model, a new approach to instrumental variable estimation in linear models is explored. A new estimator with the same limiting distribution as the maximum likelihood estimator is derived via a constrained optimization problem.

A probit regression model is studied where the covariates, measurement error and instrumental variables are normally distributed. A class of computationally simple estimators is defined and an optimal estimator in this class is identified. The maximum likelihood estimator is derived and shown to have the same asymptotic distribution as the optimal simple estimator. A small simulation study produces results consistent with the theory, and indicates the optimal simple estimator may be robust to the assumption of normally distributed covariates.

Estimation for generalized linear models in canonical form is studied via estimating functions. Optimal estimating functions are obtained for functional and non-parametric structural measurement error models when normally distributed instrumental variables are available. It is shown that the maximum

likelihood estimator for the linear normal theory model is a solution to the optimal estimating function, indicating that the optimal estimating function approach extends instrumental variable estimation to canonical generalized linear models in a natural and coherent fashion. Finally, for the logistic regression model, small sample properties of the estimators are assessed through a simulation study.

ACKNOWLEDGEMENTS

Dr. Len Stefanski, my advisor, introduced me to the problems studied in this dissertation. While I was working out the solutions, I interrupted him with questions about 10^4 times. I am very grateful for both his patience and insights. It was a pleasure working with Len. Also, very importantly, he allowed me to use his wicked fast 486.

I thank Drs. Berger, Bishir, Boos and Gerig for serving on my committee and for constructive comments on my dissertation.

Dr. Marie Davidian deserves special thanks for granting me access to her computer at the crucial, final stage of my research.

Finally, and most importantly, I thank my wife Marcia. Writing a dissertation can be frustrating. She was very understanding and supportive throughout.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Statement of the problem

The problem of finding efficient, consistent estimators of parameters for a general regression function in the presence of covariate measurement error has not been solved. The problem will not be solved in this dissertation. However, efficient, consistent estimators are obtained for a probit model and unbiased, efficient estimating functions are obtained for a class of generalized linear models. The availability of *instrumental variables*, see Definition 1.2, is assumed.

The statistical models studied in this dissertation are of the following form. Given a covariate p-vector $U = u$ and a covariate q-vector $Z = z$, Y has the density

$$f_{Y|UZ} = \exp\left\{\frac{yg(\beta_0 + u^T\beta_1 + z^T\beta_2) - b(g(\beta_0 + u^T\beta_1 + z^T\beta_2))}{a(\phi)} + c(y,\phi)\right\} \quad (1.1)$$

where $g(\cdot)$, $b(\cdot)$, $a(\cdot)$ and $c(\cdot,\cdot)$ are known functions. The covariates Z are observed but the covariates U are not. Rather X is observed which has density, given $U = u$,

$$f_{X|U} = (2\pi)^{-\frac{p}{2}}|\Sigma_{\delta\delta}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2}(x-u)^T\Sigma_{\delta\delta}^{-1}(x-u)\right\}. \quad (1.2)$$

Interest lies in estimating the regression parameters $(\beta_0, \beta_1, \beta_2)$. The problem is how to compensate for not observing U directly. The approach here is to make use of additional observations W called instrumental variables. Motivation for

the use of instrumental variables is given in Section 1.3 and the distributional assumptions are given in Definition 1.2.

Depending on how the data are collected, the covariates U may be considered as fixed or random. The situation is not unlike fixed or random effects in analysis of variance models. Models where the covariates are considered fixed are termed functional models. When the covariates are considered random the model is called structural. See Kendall and Stuart (1979, Ch. 29) for the origins of these terms.

In structural models, the distribution of the observable data (Y, X, W), given $Z = z$, depends on $f_{U|Z}$, the density of the unobservable covariates U, given $Z = z$. If the parametric form of $f_{U|Z}$ is specified, a parametric structural model is obtained. When $f_{U|Z}$ is unspecified a non-parametric structural model is obtained.

This chapter serves as a brief introduction to the problems encountered in the analysis of regression models subject to covariate measurement error. A linear model is used to illustrate problems and some solutions. The chapter is organized as follows. Section 1.2 briefly explores the effect of covariate measurement error and describes some assumptions commonly employed in the analysis of measurement error models. Section 1.3 motivates a definition for instrumental variables by determining some sufficient conditions for obtaining a consistent estimator in the linear model. The situation where more instruments are available than covariates measured with error is discussed. Section 1.4 provides a brief review of the extension of instrumental variable estimation to

non-linear models. Finally, an overview of the dissertation is given in Section 1.5.

1.2 Effects of measurement error

A common assumption in the analysis of structural measurement error models is that, given $U = u$, the response Y is independent of the observed covariate X; that is

$$f_{YX|U} = f_{Y|U} f_{X|U}$$

where , for example, $f_{X|U}$ denotes the conditional density of X given $U = u$. This assumption is equivalent to

$$f_{Y|XU} = f_{Y|U}. \tag{1.3}$$

When (1.3) is satisfied, X is called a surrogate for U. The interpretation is that, given $U = u$, X gives no more information concerning the distribution of Y. Using (1.3) it follows

$$\begin{aligned} E(Y \mid X) &= E(E(Y \mid X) \mid X,\ U) \\ &= E(E(Y \mid X,\ U) \mid X) \\ &= E(E(Y \mid U) \mid X). \end{aligned}$$

Since conditional expectation can be thought of as a smoothing operation, apparently measurement error tends to obscure the relationship between the response and covariates.

To further illustrate, consider the simple linear regression structural measurement error model

$$\begin{aligned} Y &= \beta_0 + \beta_1 U + \epsilon \\ X &= U + \delta \end{aligned} \tag{1.4}$$

where ϵ, δ and U are mutually uncorrelated with zero means and variances $\sigma^2_{\epsilon\epsilon}$, $\sigma^2_{\delta\delta}$ and σ^2_{UU} respectively.

If we also assume joint normality of (ϵ, δ, U), then

$$E(Y \mid X) = \beta_0 + \beta_1 E(U \mid X) = \beta_0 + \beta_1 (\frac{\sigma^2_{UU}}{\sigma^2_{UU} + \sigma^2_{\delta\delta}}) X. \tag{1.5}$$

The magnitude of the slope in the regression $E(Y \mid X)$ is clearly less than that in $E(Y \mid U)$ when $\sigma^2_{\delta\delta} > 0$, exhibiting the smoothing induced by the measurement error.

Another way to view model (1.4) is to write

$$Y = \beta_0 + \beta_1 X + \epsilon - \beta_1 \delta. \tag{1.6}$$

A classical assumption in regression analysis is the regressor is uncorrelated with the error. Clearly X is correlated with the error term $\epsilon - \beta_1 \delta$. Consequently, the usual estimator for slope is inconsistent, see below.

Methods that account for measurement error often require additional knowledge of some parameters in the model. The reason is lack of identifiability of the parameter vector, and/or the dimension of the parameter vector being the same order as the sample size.

DEFINITION 1.1: Let X be a random variable with distribution function $F_X(x; \theta)$ where θ is a vector of parameters in a parameter space Θ. Then θ is *identified* if for any $\theta_1, \theta_2 \in \Theta$, $\theta_1 \neq \theta_2$ implies $\exists x \ni F_X(x; \theta_1) \neq F_X(x; \theta_2)$. A particular element $\theta_{(i)}$ of θ is identified if $\theta_{i1} \neq \theta_{i2}$ implies $\exists x \ni F_X(x; \theta_1) \neq F_X(x; \theta_2)$.

One can easily show that β_1 is not identified in the normal theory version of model (1.4), even though we have specified the parametric families for the

covariates U and the measurement error δ. However, it is curious that β_1 is identified when the distribution of U is non-normal. For further discussion of this, see Fuller (1987, p.73). The normal theory functional model, that is the normal theory version of (1.4) with $U_i = u_i$ fixed, is identified provided $\sum_{i=1}^{n}(u_i - \bar{u})^2 > 0$. However, the likelihood is unbounded; take $u_i = X_i$ and let $\sigma_{\delta\delta} \to 0$. Consistent estimation of (β_0, β_1) in the structural and functional versions of (1.4) seems hopeless without additional knowledge of some model parameters.

Many methods that account for measurement error require knowledge of or the ability to consistently estimate the measurement error variance, see for example Carroll et al (1984) , Stefanski (1985), Burr (1988), Whittmore and Keller (1988). To see how knowledge of $\sigma^2_{\delta\delta}$ will allow consistent estimation of β_1, multiply (1.6) through by X and take expectations to arrive at

$$\sigma_{YX} = \beta_1\sigma^2_{XX} - \beta_1\sigma^2_{\delta\delta} = \beta_1(\sigma^2_{XX} - \sigma^2_{\delta\delta}) \tag{1.7}$$

where $\sigma^2_{XX} = \sigma^2_{UU} + \sigma^2_{\delta\delta}$. If the data consist of n independent observations $\{Y_i, X_i\}_{i=1}^{n}$, a consistent estimator of β_1 is

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{YX}}{\hat{\sigma}^2_{XX} - \sigma^2_{\delta\delta}},$$

where $(\hat{\sigma}_{YX}, \hat{\sigma}^2_{XX})$ denotes a consistent estimator of $(\sigma_{YX}, \sigma^2_{XX})$. For example, $\hat{\sigma}_{YX} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$. Note that the usual least squares estimator from the regression of Y on X is

$$\hat{\beta}_1 = \frac{\hat{\sigma}_{YX}}{\hat{\sigma}^2_{XX}} \xrightarrow{P} \left(\frac{\sigma^2_{UU}}{\sigma^2_{UU} + \sigma^2_{\delta\delta}}\right)\beta_1,$$

compare to (1.5). The inconsistency results from the correlation of the regressor X with the error $\epsilon - \beta_1\delta$.

1.3 Instrumental variables

Now suppose $\sigma^2_{\delta\delta}$ is unknown. Then (1.7) fails to identify a consistent estimator for β_1. The problem is the term $\beta_1\sigma^2_{\delta\delta}$, which is the covariance between X and the error $\epsilon - \beta_1\delta$. Suppose that a random variable W, uncorrelated with ϵ and δ, is available. Then (1.6) leads directly to

$$\sigma_{YW} = \sigma_{XW}\beta_1$$

and provided $\sigma_{XW} = \sigma_{UW} + \sigma_{\delta W} = \sigma_{UW} \neq 0$,

$$\hat{\beta}_{1,IV} = \frac{\hat{\sigma}_{YW}}{\hat{\sigma}_{XW}} \tag{1.8}$$

is consistent for β_1. As the subscript on the estimator suggests, (1.8) is an instrumental variable (IV) estimator and W is an instrumental variable. The following defines what is meant by instrumental variable in this dissertation. In order to exploit the form of the models given by (1.1) and (1.2), the definition is more restrictive then the conditions just encountered. In particular, the parametric form for the distribution of W is specified.

DEFINITION 1.2: A k-vector random variable W is an instrumental variable for an unobservable covariate u if

i) The density of W given $U = u$ and $Z = z$ is of the form

$$f_{W|UZ} = (2\pi)^{-\frac{k}{2}}|\Sigma_{WW|UZ}|^{-\frac{1}{2}} \times \exp\left\{ -\frac{1}{2}(w - \xi - \eta^T u - \gamma^T z)^T \times \Sigma^{-1}_{WW|UZ}(w - \xi - \eta^T u - \gamma^T z)\right\},$$

where $\eta \neq 0$.

ii) Given $U = u$ and $Z = z$, the response Y and the measured covariate X are independent of W, that is

$$f_{YXW|UZ} = f_{YX|UZ} f_{W|UZ}.$$

The next example illustrates that it is necessary to have at least as many instruments as covariates measured with error and that if more instruments are available than covariates measured with error a class of estimators results. Suppose that U is k-dimensional, that is k covariates are measured with error. Suppose also that p instruments W are available. Then

$$Y = X^T\beta_1 + \epsilon - \delta^T\beta_1$$

leads to

$$\Sigma_{WY} = \Sigma_{WX}\beta_1 \tag{1.9}$$

where Σ_{WY} is a $p \times 1$ vector and Σ_{WX} is a $p \times k$ matrix. Assume Σ_{WX} is full rank. When $p < k$, there are infinitely many solutions to (1.9), so that (1.9) fails to identify a consistent solution. When $p \geq k$ there exists a unique solution to (1.9) and a consistent solution can be identified. For $p \geq k$, define

$$\hat{\beta}_{1,IV} = \hat{\Sigma}^{-}_{WX}\hat{\Sigma}_{WY} \tag{1.10}$$

where $\hat{\Sigma}^{-}_{WX}$ represents any generalized inverse for $\hat{\Sigma}_{WX}$. The estimator is consistent provided $\hat{\Sigma}^{-}_{WX}$ and $\hat{\Sigma}_{WY}$ are consistent for Σ^{-}_{WX} and Σ_{WY}. Note that $\hat{\beta}_{1,IV}$ may not be uniquely defined, because if

$$\hat{\Sigma}_{WY} = \hat{\Sigma}_{WX}\beta_1 \tag{1.11}$$

is not a consistent equation, see Definition 1.3 below, $\hat{\beta}_{1,IV}$ will depend on the choice of g-inverse $\hat{\Sigma}^{-}_{WX}$.

DEFINITION 1.3: Let A and B be matrices of constants of dimension $l \times m$ and $l \times p$. Let c be a $p \times m$ matrix of parameters. The equation $A = Bc$ is consistent if there exists c^* such that $A = Bc^*$.

Under normality, $p = k$ implies (1.11) is almost surely consistent, and $p > k$ implies (1.11) is almost surely inconsistent. It turns out that a particular choice of g-inverse is efficient asymptotically. This is pursued, through a different formulation of the problem, in Chapter 2. The situation is analogous to the linear model with correlated errors. The ordinary least squares estimator is consistent, but the generalized least squares estimator is more efficient. The difference between the ordinary and generalized least squares estimators is the choice of g-inverse for the model matrix.

1.4 Non-linear models

Now suppose Y and X are observed according to

$$Y = f(U, \beta) + \epsilon$$

$$X = U + \sigma_{\delta\delta}\delta$$

where $f(\cdot, \cdot)$ is a known function and ϵ, δ and U are as in Section 1.2 except now $E(\delta^2) = 1$. Define $f'(U, \beta) = \frac{\partial}{\partial U} f(U, \beta)$. Then

$$f(X, \beta) = f(U, \beta) + f'(U, \beta)\sigma_{\delta\delta}\delta + o(\sigma_{\delta\delta})$$

so that

$$Y = f(X, \beta) - f'(U, \beta)\sigma_{\delta\delta}\delta + \epsilon + o(\sigma_{\delta\delta}).$$

Analogous to the linear model, suppose W is a random variable with $E(W\delta \mid U) = 0$, $E(W\epsilon) = 0$ and $\mathrm{COV}(W, f(X, \beta)) \neq 0$. The notation $\mathrm{COV}(\cdot, \cdot)$ denotes

the covariance of the arguments. Then

$$\Sigma_{WY} = \Sigma_{Wf}(\beta) + o(\sigma_{\delta\delta})$$

where $\Sigma_{WY} = E(WY)$ and $\Sigma_{Wf}(\beta) = E\{Wf(X,\beta)\}$. When f is linear, there exists a solution in β to

$$\Sigma_{WY} - \Sigma_{Wf}(\beta) = 0. \tag{1.12}$$

This fact enabled identification of a consistent estimator for β even though

$$\hat{\Sigma}_{WY} - \hat{\Sigma}_{Wf}(\beta) = 0,$$

the sample version of (1.12), generally did not admit a solution.

For general f, (1.12) may not have a solution, which seems to preclude development of a consistent estimator. Nevertheless, it is possible to define estimators based on (1.12). Bowden and Turkington (1984, p. 13) propose an estimator $\hat{\beta}$ that minimizes

$$\left\{\hat{\Sigma}_{WY} - \hat{\Sigma}_{Wf}(\beta)\right\}^T M \left\{\hat{\Sigma}_{WY} - \hat{\Sigma}_{Wf}(\beta)\right\} \tag{1.13}$$

where M is a positive definite weight matrix and $\{\hat{\Sigma}_{WY}, \hat{\Sigma}_{Wf}(\beta)\}$ is the sample moment version of $\{\Sigma_{WY}, \Sigma_{Wf}(\beta)\}$. Indeed, minimizing (1.13) when f is linear leads to the class of estimators (1.10) considered in the previous section.

Amemiya (1985,1990a,1990b) studies the asymptotic properties of the estimator defined through (1.13) for general f. He obtains consistency results, but only by simultaneously considering $\sigma^2_{\delta\delta} \to 0$ and $n \to \infty$.

In this dissertation consistent, efficient estimators are identified under the usual asymptotic condition $n \to \infty$. The stronger results are obtained by exploiting the structure in the non-linear models defined through (1.1).

1.5 Dissertation overview

In Chapter 2, the normal theory linear regression model with structural measurement error is analyzed when normally distributed instrumental variables are available. A new estimator is derived that is asymptotically efficient. The purpose of studying the linear model is to illuminate problems and solutions that are encountered in the study of a structural probit measurement error model.

In Chapter 3, a structural probit model is studied when the covariates, measurement error and instrumental variables are normally distributed. A class of computationally simple estimators is defined and an optimal estimator in this class is identified. The maximum likelihood estimator is derived and shown to have the same asymptotic distribution as the optimal simple estimator.

Finally, Chapter 4 studies generalized linear measurement error models in canonical form when instrumental variables are available. Unbiased estimating functions are obtained in the functional model and efficient estimating functions are obtained in the non-parametric structural model.

CHAPTER 2

Instrumental Variable Estimation in Normal Theory Linear Measurement Error Models

2.1 Introduction

The objective of this chapter is to review instrumental variable estimation in the linear model under normality assumptions. Our approach differs from those found in the literature. It is motivated by identification issues in the probit model considered in the next chapter. The purpose of this chapter is to gain insight into the more complicated probit model.

Section 2.2 defines the model and examines the parameter restrictions induced by the measurement error model. Maximum likelihood and method-of-moments instrumental variable estimators are derived in Section 2.3. In Section 2.4 an optimal method-of-moments estimation procedure is defined. Section 2.5 shows the asymptotic variance of the optimal method-of-moments estimator for slope is equivalent to the asymptotic variance for slope of the estimator (2.4.14) given in Fuller (1987, Theorem 2.4.1, p.151).

2.2 Linear regression and parameter restrictions

Consider the measurement error model given by

$$Y = \beta_0 + \beta_1^T U + \epsilon$$

$$X = U + \delta,$$

where $\epsilon \sim N(0,\ \sigma_\epsilon^2)$, $\delta \sim N(0, \Sigma_{\delta\delta})$, $U \sim N(\mu_U, \Sigma_{UU})$, and ϵ, δ, and U are mutually independent. We also have available a vector W of normally distributed instruments, see Definition 1.2. Estimation of β_1 is the primary

interest. For simplicity, assume $\Sigma_{\delta\delta}$ and Σ_{UU} are full rank, implying there are no covariates measured without error.

The model can be described succinctly in terms of the observable vector

$$\begin{pmatrix} Y \\ X \\ W \end{pmatrix} \sim N\left\{ \begin{pmatrix} \beta_0 + \beta_1^T \mu_U \\ \mu_U \\ \mu_W \end{pmatrix}, \begin{pmatrix} \beta_1^T \Sigma_{UU} \beta_1 + \sigma^2_{\epsilon\epsilon} & \beta_1^T \Sigma_{UU} & \beta_1^T \Sigma_{UW} \\ \Sigma_{UU}\beta_1 & \Sigma_{UU} + \Sigma_{\delta\delta} & \Sigma_{UW} \\ \Sigma_{WU}\beta_1 & \Sigma_{WU} & \Sigma_{WW} \end{pmatrix} \right\} \quad (2.1)$$

where $\dim(Y) = 1$, and $\dim(W) \geq \dim(X)$. Throughout this dissertation, means and centered second-moment matrices are indicated by μ and Σ respectively where the subscripts indicate the corresponding distribution, e.g. μ_U denotes the mean of U and $\mu_{U|XW}$ denotes the mean of $U \mid X, W$. For scalar random variables, σ replaces Σ. Note that the distribution of W defined in (2.1) is compatible with Definition 1.2. In order to identify β_1, assume $\text{rank}(\Sigma_{WU}) = \dim(X)$. The data consist of n independent copies of (2.1). Denote the parameters in this model by the vector

$$\theta_0 = (\beta_0, \beta_1^T, \mu_U, \mu_W, \sigma^2_{\epsilon\epsilon}, \text{vech}^T \Sigma_{UU}, \text{vech}^T \Sigma_{\delta\delta}, \text{vech}^T \Sigma_{WU}, \text{vech}^T \Sigma_{WW})^T.$$

This model can be reparameterized to

$$\begin{pmatrix} Y \\ X \\ W \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_Y \\ \mu_X \\ \mu_W \end{pmatrix}, \begin{pmatrix} \sigma^2_{YY} & \Sigma_{YX} & \Sigma_{YW} \\ \Sigma_{XY} & \Sigma_{XX} & \Sigma_{XW} \\ \Sigma_{WY} & \Sigma_{WX} & \Sigma_{WW} \end{pmatrix} \right\}.$$

Inspection of the original parameterization reveals the constraint $\Sigma_{WY} \in C(\Sigma_{WX})$, where $C(\cdot)$ denote column space. Let

$$\theta_1 = (\mu_Y, \mu_X^T, \mu_W, \sigma^2_{YY}, \text{vec}^T \Sigma_{XY}, \text{vech}^T \Sigma_{XX}, \text{vec}^T \Sigma_{WY}, \text{vec}^T \Sigma_{WX}, \text{vech}^T \Sigma_{WW})^T$$

represent the above parameterization. The vech operator forms a column vector from the unique $\frac{k(k+1)}{2}$ elements of a $k \times k$ symmetric matrix. Ignoring the constraint $\Sigma_{WY} \in C(\Sigma_{WX})$ and maximizing the likelihood with respect to θ_1 leads to the class of estimators (1.12) considered in Chapter 1. Maximizing the likelihood with respect to θ_1 subject to $\Sigma_{WY} \in C(\Sigma_{WX})$ leads to the maximum likelihood estimator. Instead of analyzing the problem from this perspective, we consider other parameterizations that are amenable to the probit model considered in Chapter 3. A new class of computationally simple estimators results. See Fuller (1987, Ch.2, p. 148) and Judge et. al. (1985, Ch. 14-15) for other instrumental variable estimators in the linear model.

Letting f_{YXW} denote the density of (Y, X, W), the factorization $f_{YXW} = f_{Y|XW} f_{X|W} f_W$ suggests the regression parameterization,

$$\begin{pmatrix} Y \\ X \\ W \end{pmatrix} \sim N \left\{ \begin{pmatrix} \beta_{Y|\underline{1}XW} + \beta_{Y|1\underline{X}W}^T \mu_X + \beta_{Y|1X\underline{W}}^T \mu_W \\ \beta_{X|\underline{1}W} + \beta_{X|1\underline{W}}^T \mu_W \\ \mu_W \end{pmatrix}, \begin{pmatrix} \sigma^2_{YY|XW} & A^T & B^T \\ A & \Sigma_{XX|W} & C^T \\ B & C & \Sigma_{WW} \end{pmatrix} \right\}$$

where $\mu_X = \beta_{X|\underline{1}W} + \beta_{X|1\underline{W}}^T \mu_W$, and the covariances A, B, and C are defined in terms of

$$\theta_2 = (\beta_{Y|1\underline{XW}}^T, \mathrm{vec}^T \beta_{X|1\underline{W}}, \mu_W^T, \sigma^2_{YY|XW}, \mathrm{vech}^T \Sigma_{XX|W}, \mathrm{vech}^T \Sigma_{WW})^T.$$

The regression coefficients $\beta_{X|\underline{1}W}$ and $\beta_{X|1\underline{W}}$, for example, are the coefficients of the intercept and W respectively of the regression of X on 1 and W. The notation allows the identification of groups of coefficients, e.g. $\beta_{X|1\underline{W}} = (\beta_{X|\underline{1}W}, \beta_{X|1\underline{W}})$.

Note it is possible to write $\theta_2 = g(\theta_1)$ for some 1 to 1 function $g(\cdot)$. Suppose, in general, that a model with parameters θ_1 is subject to the condition $h(\theta_1) = 0$ for some function $h(\cdot)$. If $\theta_2 = g(\theta_1)$, and $g(\cdot)$ is 1 to 1, the constraint in terms of the parameterization θ_2 is $h \circ g^{-1}(\theta_2) = 0$.

LEMMA 2.1. *The regression parameterization is constrained by*

$$\beta_{Y|1X\underline{W}} \in C(\beta_{X|1\underline{W}}). \tag{2.2}$$

PROOF: The relevant relations between parameterizations θ_1 and θ_2 are

$$\Sigma_{WX} = \Sigma_{WW}\beta_{X|1\underline{W}}$$

and

$$\Sigma_{WY} = \Sigma_{WX}\beta_{Y|1\underline{X}W} + \Sigma_{WW}\beta_{Y|1X\underline{W}}.$$

Since $\Sigma_{WY} = \Sigma_{WX}\beta_1$, it follows $\beta_{X|1\underline{W}}\beta_1 = \beta_{X|1\underline{W}}\beta_{Y|1\underline{X}W} + \beta_{Y|1X\underline{W}}$, or rearranging

$$\beta_{X|1\underline{W}}(\beta_1 - \beta_{Y|1\underline{X}W}) = \beta_{Y|1X\underline{W}}. \tag{2.3}$$

Define

$$\lambda = \beta_1 - \beta_{Y|1\underline{X}W}.$$

Then $\beta_{X|1\underline{W}}\lambda = \beta_{Y|1X\underline{W}}$, i.e. $\beta_{Y|1X\underline{W}} \in C(\beta_{X|1\underline{W}})$. •

The constraint (2.2) on the regression parameterization suggests replacing $\beta_{Y|1X\underline{W}}$ by $\beta_{X|1\underline{W}}\lambda$ leading to yet another parameterization

$$\theta_3 = (\beta_{Y|\underline{1}XW}, \beta^T_{Y|1\underline{X}W}, \lambda^T, \mathrm{vec}^T\beta_{X|1\underline{W}}, \sigma^2_{YY|XW}, \mathrm{vech}^T\Sigma_{XX|W}, \mathrm{vech}^T\Sigma_{WW})^T.$$

The difference between parameterizations θ_2 and θ_3 is θ_2 ignores the constraint (2.2) while θ_3 does not. Maximizing the likelihood of independent observations $\{Y_i, X_i, W_i\}_{i=1}^n$ with respect to θ_3 will yield maximum likelihood estimators for β_0 and β_1. The maximum likelihood estimator and method-of-moments type estimators are discussed in the next section.

2.3 Maximum likelihood and method-of-moments estimators

We first derive the maximum likelihood estimators for the regression parameters (β_0, β_1) of model (2.1). This amounts to examining relationships between θ_0 and θ_3.

As defined in the proof of Lemma 2.1, $\lambda = \beta_1 - \beta_{Y|1\underline{X}W}$. By the invariance property of maximum likelihood estimators,

$$\tilde{\beta}_{1,\text{mle}} = \tilde{\beta}_{Y|1\underline{X}W} + \tilde{\lambda},$$

where $\tilde{\beta}_{Y|1\underline{X}W}$ and $\tilde{\lambda}$ denote the estimators for $\beta_{Y|1\underline{X}W}$ and λ obtained by maximizing the likelihood of $\{Y_i, X_i, W_i\}_{i=1}^n$ with respect to θ_3. Next, combining the identities

$$\begin{aligned}
\beta_0 + \mu_X^T \beta_1 &= \beta_{Y|\underline{1}XW} + \beta_{Y|1\underline{X}W}(\beta_{X|\underline{1}W} + \beta_{X|1\underline{W}}\mu_W) + \beta_{X|1\underline{W}}\lambda, \\
\mu_X &= \beta_{X|\underline{1}W} + \beta_{X|1\underline{W}}\mu_W, \\
\lambda &= \beta_1 - \beta_{Y|1\underline{X}W},
\end{aligned}$$

and solving for β_0 gives

$$\tilde{\beta}_{0,\text{mle}} = \tilde{\beta}_{Y|\underline{1}XW} - \tilde{\beta}_{X|\underline{1}W}\tilde{\lambda}.$$

The method-of-moments estimators are quite similar. They are derived from relations between parameterizations θ_2 and θ_0.

From the proof of Lemma 2.1, $\beta_{X|1\underline{W}}(\beta_1 - \beta_{Y|1\underline{X}W}) = \beta_{Y|1X\underline{W}}$ so that

$$\hat{\beta}_{1,\mathrm{mm}} = \hat{\beta}_{Y|1\underline{X}W} + \hat{\beta}^{-}_{X|1\underline{W}}\hat{\beta}_{Y|1X\underline{W}},$$

where $\hat{\beta}^{-}_{X|1\underline{W}}$ denotes a generalized inverse for $\hat{\beta}_{X|1\underline{W}}$. In an analogous fashion to the derivation of $\hat{\beta}_{0,\mathrm{mle}}$, we arrive at

$$\hat{\beta}_{0,\mathrm{mm}} = \hat{\beta}_{Y|1XW} - \hat{\beta}^{T}_{X|1W}\hat{\beta}^{-}_{X|1\underline{W}}\hat{\beta}_{Y|1X\underline{W}}.$$

The estimators $\hat{\beta}_{Y|1\underline{X}\underline{W}}$ and $\hat{\beta}_{X|1\underline{W}}$ are obtained by maximizing the likelihood of $\{Y_i, X_i, W_i\}_{i=1}^{n}$ with respect to θ_2, or equivalently via least squares regressions of Y on X and W and X on W respectively.

When the regression parameters in (2.3) are replaced by the corresponding estimates, the resulting equation is not necessarily consistent in the sense of Definition 1.3. Consequently, the properties of the method-of-moments estimators depend on the choice of a generalized inverse for $\beta_{X|1\underline{W}}$. This is explored in Section 2.4. Note if $\dim(W) = \dim(X)$, that is we have the same number of instruments as covariates measured with error, then (2.3) is almost surely consistent since $\beta_{X|1\underline{W}}$ is non-singular with probability one. The method-of-moments estimators and the maximum likelihood estimators are equivalent in this case.

2.4 The optimal method-of-moments estimator

In this section we show that the asymptotic distribution of $\hat{\beta}_{1,\mathrm{mm}}$ depends on the choice of a generalized inverse (g-inverse) for $\hat{\beta}_{X|1\underline{W}}$. For clarity, assume

that $\dim(X) = 1$, and replace $\Sigma_{XX|W}$ with $\sigma^2_{XX|W}$.

Assume the data are generated according to (2.1). Also assume

$$E(W) = 0. \tag{2.4}$$

It can be shown that no generality is lost with this additional assumption. See Section 3.5 for the argument in the probit model.

Define $\hat{\beta}_{Y|1\underline{X}W}$, $\hat{\beta}_{X|1\underline{W}}$, and $(\hat{\sigma}^2_{YY|XW}, \hat{\sigma}^2_{XX|W})$ as the solutions to

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\theta_2; Y_i, X_i, W_i,) =$$

$$\frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \psi_1(\theta_2; Y_i, X_i, W_i) \\ \psi_2(\theta_2; X_i, W_i) \\ \psi_3(\theta_2; Y_i, X_i, W_i) \end{pmatrix} = 0$$

where, suppressing function arguments,

$$\psi_1 = \sigma^{-2}_{YY|XW}(Y_i - \beta_{Y|1\underline{X}W} - \beta_{Y|1\underline{X}W}X_i - \beta_{Y|1X\underline{W}}W_i) \begin{pmatrix} 1 \\ X_i \\ W_i \end{pmatrix}$$

$$\psi_2 = \sigma^{-2}_{XX|W}(X_i - \beta_{X|1\underline{W}} - \beta_{X|1\underline{W}}W_i) \begin{pmatrix} 1 \\ W_i \end{pmatrix}$$

and

$$\psi_3 = \begin{pmatrix} \sigma^2_{YY|XW} - (Y_i - \beta_{Y|1\underline{X}W} - \beta_{Y|1\underline{X}W}X_i - \beta_{Y|1X\underline{W}}W_i)^2 \\ \sigma^2_{XX|W} - (X_i - \beta_{X|1\underline{W}} - \beta_{X|1\underline{W}}W_i)^2 \end{pmatrix}.$$

Note ψ_1 and ψ_2 lead to the familiar normal equations from linear models. Then appealing to standard maximum likelihood theory (or Theorem 3.5 in Chapter 3; the regularity conditions are easily verified),

$$\begin{pmatrix} \hat{\beta}_{Y|1\underline{X}W} \\ \hat{\beta}_{X|1\underline{W}} \end{pmatrix} \quad \text{is} \quad AN\left\{ \begin{pmatrix} \beta_{Y|1\underline{X}W} \\ \beta_{X|1\underline{W}} \end{pmatrix}, \begin{pmatrix} \frac{1}{n}\Sigma_1 & 0 \\ 0 & \frac{1}{n}\Sigma_2 \end{pmatrix} \right\}$$

where,

$$\Sigma_1 = \left\{E(\psi_1\psi_1^T)\right\}^{-1} \quad \text{and} \quad \Sigma_2 = \left\{E(\psi_2\psi_2^T)\right\}^{-1}.$$

Partition the asymptotic covariance matrix of $\beta_{Y|1\underline{XW}}$ as

$$\Sigma_1 = \sigma^2_{YY|XW}\begin{pmatrix} A_{11} & A_{1X} & A_{1W} \\ A_{X1} & A_{XX} & A_{XW} \\ A_{W1} & A_{WX} & A_{WW} \end{pmatrix}$$

where $\text{ACOV}(\beta_{Y|1\underline{X}W}, \beta_{Y|1X\underline{W}}) = \sigma^2_{YY|XW}A_{XW}$ for example. The notation $\text{ACOV}(\cdot,\cdot)$ denotes the covariance of the limiting distributions of the arguments. Similarly, partition

$$\Sigma_2 = \sigma^2_{XX|W}\begin{pmatrix} B_{11} & 0 \\ 0 & B_{WW} \end{pmatrix}.$$

The following Lemma is needed in the derivation of an optimal g-inverse.

LEMMA 2.2. *Under assumptions (2.1) and (2.4)*,

$$ACOV(\beta_{Y|1X\underline{W}}, \beta_{Y|1\underline{X}W}) = \sigma^2_{YY|XW}A_{WX} = -\sigma^2_{YY|XW}\beta_{X|1\underline{W}}A_{XX},$$
$$ACOV(\beta_{Y|1X\underline{W}}, \beta_{Y|\underline{1}XW}) = -\sigma^2_{YY|XW}A_{WX}\mu_X.$$

PROOF: Recall $E(W_i) = 0$ so that

$$E(\psi_1\psi_1^T) = E\left\{E(\psi_1\psi_1^T \mid XW)\right\} = \sigma^{-2}_{YY|XW}E\left\{\begin{pmatrix} 1 \\ X \\ W \end{pmatrix}(1 \quad X \quad W)\right\}$$
$$= \sigma^{-2}_{YY|XW}\begin{pmatrix} 1 & \mu_X^T & 0 \\ \mu_X & \mu_{XX} & \Sigma_{XW} \\ 0 & \Sigma_{WX} & \Sigma_{WW} \end{pmatrix}.$$

Apply result A.2 of the appendix for inverting partitioned matrices to obtain

$$\Sigma_1 = \sigma^2_{YY|XW}\begin{pmatrix} A_{11} & A_{1X} & A_{1W} \\ A_{X1} & A_{XX} & A_{XW} \\ A_{W1} & A_{WX} & A_{WW} \end{pmatrix}$$

where

$$A_{11} = 1 + \mu_X^T A_{WW} \mu_X$$

$$A_{XX} = (\sigma_{XX} - \Sigma_{XW}\Sigma_{WW}^{-1}\Sigma_{WX})^{-1}$$

$$A_{WW} = (\Sigma_{WW} - \Sigma_{WX}\Sigma_{XX}^{-1}\Sigma_{XW})^{-1}$$

$$A_{X1} = -A_{XX}\mu_X$$

$$A_{W1} = -A_{WX}\mu_X$$

$$A_{WX} = -\Sigma_{WW}\Sigma_{WX}A_{XX}.$$

Since $\Sigma_{WW}^{-1}\Sigma_{WX} = \beta_{X|1\underline{W}}$, the result follows. •

The next task is to determine how the asymptotic covariance of $\hat{\beta}_{1,\mathrm{mm}}$ depends on $\beta_{X|1\underline{W}}^{-}$ and then try to identify an optimal estimation strategy. It is useful to note that the g-inverses for a full rank matrix M of dimension $u \times v$, $u \geq v$ are characterized by $M^{-}M = I$. For convenience, denote g-inverses of $\beta_{X|1\underline{W}}$ by G^T, i.e. G^T satisfies $G^T\beta_{X|1\underline{W}} = 1$.

THEOREM 2.1. *Under assumptions (2.1) and (2.4) , the asymptotic variance of* $\hat{\beta}_{1,\mathrm{mm}}$ *is given by*

$$-\sigma_{YY|XW}^2 A_{XX} + G^T\left(\sigma_{YY|XW}^2 A_{WW} + \lambda^2\sigma_{XX|W}^2 B_{WW}\right)G.$$

This expression is minimized over G^T *satisfying* $G^T\beta_{X|1\underline{W}} = 1$ *by*

$$\underline{G}^T = \left\{\beta_{X|1\underline{W}}^T(\sigma_{YY|XW}^2 A_{WW} + \lambda^2\sigma_{XX|W}^2 B_{WW})^{-1}\beta_{X|1\underline{W}}\right\}^{-1} \times \beta_{X|1\underline{W}}^T(\sigma_{YY|XW}^2 A_{WW} + \lambda^2\sigma_{XX|W}^2 B_{WW})^{-1}.$$

Note the underbar notation is used to denote the optimal g-inverse.

PROOF: Using the identities $\beta_{Y|1X\underline{W}} = \beta_{X|1\underline{W}}\lambda$, $G^T\beta_{X|1\underline{W}} = 1$, $\hat{G}^T\hat{\beta}_{X|1\underline{W}} = 1$ and assuming $\hat{G}^T \xrightarrow{P} G^T$ it follows

$$\begin{aligned}
&\hat{G}^T\hat{\beta}_{Y|1X\underline{W}} - G^T\beta_{Y|1X\underline{W}} \\
&\quad = (\hat{G}^T - G^T)\hat{\beta}_{Y|1X\underline{W}} + G^T(\hat{\beta}_{Y|1X\underline{W}} - \beta_{Y|1X\underline{W}}) \\
&\quad = \hat{G}^T(I - \hat{\beta}_{X|1\underline{W}}G^T)\hat{\beta}_{Y|1X\underline{W}} + G^T(\hat{\beta}_{Y|1X\underline{W}} - \beta_{Y|1X\underline{W}}) \\
&\quad = \hat{G}^T(I - \hat{\beta}_{X|1\underline{W}}G^T)\beta_{Y|1X\underline{W}} + G^T(\hat{\beta}_{Y|1X\underline{W}} - \beta_{Y|1X\underline{W}}) + o_p(n^{-\frac{1}{2}}) \\
&\quad = -G^T(\hat{\beta}_{X|1\underline{W}} - \beta_{X|1\underline{W}})\lambda + G^T(\hat{\beta}_{Y|1X\underline{W}} - \beta_{Y|1X\underline{W}}) + o_p(n^{-\frac{1}{2}}).
\end{aligned}$$

The third equality is justified by an appeal to Slutsky's Theorem and the last equality follows from the identities $\beta_{Y|1X\underline{W}} = \beta_{X|1\underline{W}}\lambda$ and $G^T\beta_{X|1\underline{W}} = 1$ and an appeal to Slutsky's Theorem. Then

$$\begin{aligned}
(\hat{\beta}_{1,\mathrm{mm}} - \beta_1) = (\hat{\beta}_{Y|1\underline{X}W} - \beta_{Y|1\underline{X}W}) \\
- G^T(\hat{\beta}_{X|1\underline{W}} - \beta_{X|1\underline{W}})\lambda + G^T(\hat{\beta}_{Y|1X\underline{W}} - \beta_{Y|1X\underline{W}}) + o_p(n^{-\frac{1}{2}})
\end{aligned}$$

so that

$$\begin{aligned}
&\mathrm{ACOV}(\hat{\beta}_{1,\mathrm{mm}}, \hat{\beta}_{1,\mathrm{mm}}) = \\
&\quad \sigma^2_{YY|XW}(A_{XX} + A_{XW}G + G^TA_{WX} + G^TA_{WW}G) \\
&\qquad + G^T\lambda^2\sigma^2_{XX|W}B_{WW}G \\
&\quad = -\sigma^2_{YY|XW}A_{XX} + G^T(\sigma^2_{YY|XW}A_{WW} + \sigma^2_{XX|W}\lambda^2B_{WW})G.
\end{aligned}$$

The last equality follows from Lemma 2.2.

The solution to $\min_{G^T}\left\{G^T(\sigma^2_{YY|XW}A_{WW}+\sigma^2_{XX|W}\lambda^2 B_{WW})G\right\}$ subject to $G^T\beta_{X|1\underline{W}}=1$ is found in Rao (1973, p.60). The solution is

$$\underline{G}^T=$$
$$\left\{\beta^T_{X|1\underline{W}}(\sigma^2_{YY|XW}A_{WW}+\lambda^2\sigma^2_{XX|W}B_{WW})^{-1}\beta_{X|1\underline{W}}\right\}^{-1}$$
$$\times\,\beta^T_{X|1\underline{W}}(\sigma^2_{YY|XW}A_{WW}+\lambda^2\sigma^2_{XX|W}B_{WW})^{-1}. \quad \bullet$$

Note that $\mathrm{ACOV}(\hat{\beta}_{1,\mathrm{mm}},\hat{\beta}_{1,\mathrm{mm}})$ depends on the choice of G^T but not on the method of estimating G^T, i.e., any consistent estimator of G^T will result in the same asymptotic distribution for $\beta_{1,\mathrm{mm}}$. The optimal method-of-moments estimation strategy is now clear. Define $\hat{\beta}_{1,\mathrm{mm}}=\hat{\beta}_{Y|1\underline{X}W}+\hat{\underline{G}}^T\hat{\beta}_{Y|1X\underline{W}}$, where $\hat{\underline{G}}^T$ satisfies $\hat{\underline{G}}^T\hat{\beta}_{X|1\underline{W}}=1$ and $\hat{\underline{G}}^T\xrightarrow{P}\underline{G}^T$. From Theorem 2.1, the sequence of estimators will have minimum asymptotic variance in the class of method of moments estimators. An obvious choice for the estimated optimal g-inverse is

$$\hat{\underline{G}}^T=$$
$$\left\{\hat{\beta}^T_{X|1\underline{W}}(\hat{\sigma}^2_{YY|XW}\hat{A}_{WW}+\hat{\sigma}^2_{XX|W}\hat{\lambda}^2\hat{B}_{WW})^{-1}\hat{\beta}_{X|1\underline{W}}\right\}^{-1}$$
$$\times\,\hat{\beta}_{X|1\underline{W}}(\hat{\sigma}^2_{YY|XW}\hat{A}_{WW}+\hat{\sigma}^2_{XX|W}\hat{\lambda}^2\hat{B}_{WW})^{-1},$$

where $\hat{\lambda}=\hat{G}^T\hat{\beta}_{Y|1X\underline{W}}$ and $\hat{G}^T$ satisfies $\hat{G}^T\hat{\beta}_{X|1\underline{W}}=1$ and $\hat{G}^T\xrightarrow{P}G^T$. For example, $\hat{G}^T=(\hat{\beta}^T_{X|1\underline{W}}\hat{\beta}_{X|1\underline{W}})^{-1}\hat{\beta}^T_{X|1\underline{W}}$. Then $\hat{\underline{G}}^T\xrightarrow{P}\underline{G}^T$ follows from the consistency of the linear regression estimators and the sample moments of the normal distribution.

2.5 Asymptotic variances of the optimal method-of-moments and maximum likelihood estimators

Using Lemma 2.2 it can be shown $\underline{G}^T$ minimizes the asymptotic variance of the joint asymptotic distribution of $(\hat{\beta}_{0,\mathrm{mm}},\hat{\beta}_{1,\mathrm{mm}})^T$. It can then be shown this

variance is equivalent to the asymptotic variance of the maximum likelihood estimator. The proof of this fact is omitted since it is cumbersome and very similar to the analogous result for the probit model, see Section 3.5. However, the skeptic may wish to read the remainder of this section. It is shown the asymptotic variance of the optimal method-of-moments estimator for slope is equivalent to the asymptotic variance of the estimator (2.4.14) in Fuller (1987, Theorem 2.4.1, p.151).

The asymptotic variance of the estimator (2.4.14) in Fuller is given in his Theorem 2.4.1. The variance for the slope, under the assumption $E(W) = 0$ and converted to our notation is

$$\frac{\sigma^2_{\epsilon\epsilon} + \beta_1^2 \sigma^2_{\delta\delta}}{\Sigma_{XW}\Sigma^{-1}_{WW}\Sigma_{WX}}. \tag{2.5}$$

The variance of the optimal method-of-moments estimator, as seen from Theorem 2.1, is

$$-\sigma^2_{YY|XW}A_{XX} + \{\beta^T_{X|1\underline{W}}(\sigma^2_{YY|XW}A_{WW} + \lambda^2\sigma^2_{XX|W}B_{WW})^{-1}\beta_{X|1\underline{W}}\}^{-1}. \tag{2.6}$$

The task is to show (2.5) and (2.6) are equivalent. We first show

$$\begin{aligned}\{\beta^T_{X|1\underline{W}}(\sigma^2_{YY|XW}A_{WW} + \lambda^2\sigma^2_{XX|W}B_{WW})^{-1}\beta_{X|1\underline{W}}\}^{-1} \\ = \frac{\sigma^2_{YY|XW} + \lambda^2\sigma^2_{XX|W}}{\Sigma_{XW}\Sigma^{-1}_{WW}\Sigma_{WX}} + A_{XX}\sigma^2_{YY|XW}.\end{aligned}$$

Using the representations

$$\begin{aligned}A_{WW} &= \Sigma^{-1}_{WW} + \beta_{X|1\underline{W}}A_{XX}\beta^T_{X|1\underline{W}} \\ B_{WW} &= \Sigma^{-1}_{WW}\end{aligned}$$

and result A.1 in the appendix we have

$$(\sigma^2_{YY|XW}A_{WW} + \lambda^2\sigma^2_{XX|W}B_{WW})^{-1}$$
$$= k\Sigma_{WW} - \frac{k^2\sigma^2_{YY|XW}\Sigma_{WW}\beta_{X|1\underline{W}}A_{XX}\beta^T_{X|1\underline{W}}\Sigma_{WW}}{1 + k\sigma^2_{YY|XW}A_{XX}\beta^T_{X|1\underline{W}}\Sigma_{WW}\beta_{X|1\underline{W}}}$$
$$= \frac{k\Sigma_{WW} + k^2\sigma^2_{YY|XW}A_{XX}(\Sigma_{WW}\beta^T_{X|1\underline{W}}\Sigma_{WW}\beta_{X|1\underline{W}} - \Sigma_{WW}\beta_{X|1\underline{W}}\beta^T_{X|1\underline{W}}\Sigma_{WW})}{1 + k\sigma^2_{YY|XW}A_{XX}\beta^T_{X|1\underline{W}}\Sigma_{WW}\beta_{X|1\underline{W}}}$$

where $k = (\sigma^2_{YY|XW} + \lambda^2\sigma^2_{XX|W})^{-1}$. Then,

$$\{\beta^T_{X|1\underline{W}}(\sigma^2_{YY|XW}A_{WW} + \lambda^2\sigma^2_{XX|W}B_{WW})^{-1}\beta_{X|1\underline{W}}\}^{-1}$$
$$= \left(\frac{k\beta^T_{X|1\underline{W}}\Sigma_{WW}\beta_{X|1\underline{W}}}{1 + k\sigma^2_{YY|XW}A_{XX}\beta^T_{X|1\underline{W}}\Sigma_{WW}\beta_{X|1\underline{W}}}\right)^{-1}$$
$$= \frac{1}{k\beta^T_{X|1\underline{W}}\Sigma_{WW}\beta_{X|1\underline{W}}} + \sigma^2_{YY|XW}A_{XX}$$
$$= \frac{\sigma^2_{YY|XW} + \lambda^2\sigma^2_{XX|W}}{\Sigma_{XW}\Sigma^{-1}_{WW}\Sigma_{WX}} + \sigma^2_{YY|XW}A_{XX}$$

where we have used $\beta_{X|1\underline{W}} = \Sigma^{-1}_{WW}\Sigma_{WX}$. It remains to show $\sigma^2_{YY|XW} + \lambda^2\sigma^2_{XX|W} = \sigma^2_{\epsilon\epsilon} + \beta_1^2\sigma^2_{\delta\delta}$. There are several ways to do this. We use the identities

$$\sigma^2_Y = \beta_1^2\sigma^2_{XX} - 2\beta_1^2\sigma^2_{\delta\delta} + \sigma^2_{\epsilon\epsilon} + \beta_1^2\sigma^2_{\delta\delta}$$
$$\sigma^2_Y = \beta^2_{Y|1\underline{X}W}\sigma^2_{XX} + \beta^T_{Y|1X\underline{W}}\Sigma_{WW}\beta_{Y|1X\underline{W}} + 2\beta_{Y|1\underline{X}W}\Sigma_{XW}\beta_{Y|1X\underline{W}} + \sigma^2_{YY|XW}$$
$$\sigma^2_{XX} = \beta^T_{X|1\underline{W}}\Sigma_{WW}\beta_{X|1\underline{W}} + \sigma^2_{XX|W}$$
$$\beta_{Y|1\underline{X}W} = \beta_1 - \lambda$$
$$\beta_{Y|1X\underline{W}} = \beta_{X|1\underline{W}}\lambda = \Sigma^{-1}_{WW}\Sigma_{WX}\lambda.$$

The first three identities result from the regressions of Y on X, Y on X and

W, and X on W, respectively. The first two identities give

$$\sigma^2_{\epsilon\epsilon} + \beta_1^2\sigma^2_{\delta\delta} =$$
$$\sigma^2_{YY|XW} + \lambda^2\sigma^2_{XX|W} - \lambda^2\sigma^2_{XX|W} - \beta_1^2\sigma^2_{XX} + 2\beta_1^2\sigma^2_{\delta\delta}$$
$$+ \beta^2_{Y|1\underline{X}W}\sigma^2_{XX} + \beta^T_{Y|1X\underline{W}}\Sigma_{WW}\beta_{Y|1X\underline{W}} + 2\beta_{Y|1\underline{X}W}\Sigma_{XW}\beta_{Y|1X\underline{W}}.$$

Using the last three identities, the right hand side of the equation above simplifies to

$$\sigma^2_{YY|XW} + \lambda^2\sigma^2_{XX|W} - 2\beta_1\lambda\sigma^2_{XX|W} + 2\beta_1^2\sigma^2_{\delta\delta}.$$

If we can show $\lambda\sigma^2_{XX|W} = \beta_1\sigma^2_{\delta\delta}$, then we are done. Using the definition of $\beta_{Y|1\underline{X}W}$ and some identities established in the Proof of Lemma 2.2 we have

$$\begin{aligned}
\beta_{Y|1\underline{X}W} &= A_{XX}\Sigma_{XY} + A_{XW}\Sigma_{WY} \\
&= \beta_1(A_{XX}\sigma^2_{UU} - A_{XX}\beta^T_{X|1\underline{W}}\Sigma_{WX}) \\
&= \beta_1 A_{XX}(\sigma^2_{XX} - \Sigma_{XW}\Sigma^{-1}_{WW}\Sigma_{WX} - \sigma^2_{\delta\delta}) \\
&= \beta_1 - \beta_1 A_{XX}\sigma^2_{\delta\delta}
\end{aligned}$$

so that $\lambda = \beta_1 - \beta_{Y|1\underline{X}W} = \beta_1 A_{XX}\sigma^2_{\delta\delta}$. Note that $\sigma^2_{XX|W} = A^{-1}_{XX}$, so that

$$\lambda\sigma^2_{XX|W} = \beta_1\sigma^2_{\delta\delta}$$

as required.

CHAPTER 3

Instrumental Variable Estimation in a Probit Measurement Error Model

3.1 Introduction

In this chapter a probit regression model is studied wherein normally distributed covariates are subject to normally distributed measurement errors. Under the assumption that instrumental variables are available, the parameters in the probit model are shown to be identified. A class of computationally simple estimators is defined and an optimal estimator in this class is identified. The maximum likelihood estimator is derived and shown to have the same asymptotic distribution as the optimal simple estimator.

Much of the literature concerning estimation of regression parameters when some covariates are subject to measurement error assumes knowledge of or the ability to consistently estimate the measurement error variance. Examples pertaining to binary regression models include Carroll *et. al.* (1984), Stefanski and Carroll (1985), and Burr (1988). It is often assumed that replicate measurements on covariates subject to error are available since this allows estimation of the error variance. However, the assumption that additional measurements are replicates is not always tenable and, when violated, will result in additional bias in estimators. In linear models, this has been recognized and studied through instrumental variables models, see Fuller (1987).

The extensions of instrumental variable theory to non-linear models has been limited. Amemiya (1985, 1990a, 1990b) and Turkington and Bowden (1984) describe methods that apply quite generally, and Stefanski and Buzas (1992) describe methods applicable to generalized linear models. However,

the methods described by these authors usually result in only approximately consistent estimators where the strength of the approximation depends on the size of the measurement error variance. In this chapter, computationally simple and fully efficient estimators are derived for a probit measurement error model with normally distributed instruments.

The statistical model has the following form. Given $U = u$ and $Z = z$, assume the binary response Y satisfies

$$\Pr(Y = 1 \mid U = u, Z = z) = \Phi(\beta_0 + \beta_1^T u + \beta_2^T z), \tag{3.1}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Model (3.1) is a probit regression model. Interest lies in estimating $(\beta_0, \beta_1, \beta_2)$. Suppose U cannot be observed but rather $X = U + \delta$ is observed where δ is normally distributed measurement error. The availability of *instrumental variables* W is also assumed. In particular, it is assumed that the observable data consist of n independent, identically distributed observations $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ satisfying the distributional assumptions

$$E(Y \mid X = x, Z = z, W = w) = E(E(Y \mid U, Z) \mid X = x, Z = z, W = w) \tag{3.2}$$

and

$$\begin{pmatrix} X \\ U \\ Z \\ W \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu_U \\ \mu_U \\ \mu_Z \\ \mu_W \end{pmatrix}, \begin{pmatrix} \Sigma_{UU} + \Sigma_{\delta\delta} & \Sigma_{UU} & \Sigma_{UZ} & \Sigma_{UW} \\ \Sigma_{UU} & \Sigma_{UU} & \Sigma_{UZ} & \Sigma_{UW} \\ \Sigma_{ZU} & \Sigma_{ZU} & \Sigma_{ZZ} & \Sigma_{ZW} \\ \Sigma_{WU} & \Sigma_{WU} & \Sigma_{WZ} & \Sigma_{WW} \end{pmatrix} \right\}. \tag{3.3}$$

Assumption (3.2) is equivalent to the assertion that the conditional distribution of $Y \mid U, X, Z, W$ is the same as the conditional distribution of $Y \mid U, Z$. Implicit in (3.3) is that $\mathrm{COV}(W, \delta) = 0$ and it is assumed $\mathrm{COV}(W, U) = \Sigma_{WU} \neq 0$. The distributional assumptions imposed on W are compatible with Definition 1.2.

This chapter is organized as follows. In Section 3.2, two parameterizations for (3.1)-(3.3) lead to two estimation procedures for $(\beta_0, \beta_1, \beta_2)$; the maximum likelihood estimator and method-of-moments estimators. In Section 3.3 an optimal simple estimator is identified and the main theorem is stated; the optimal simple estimator and the maximum likelihood estimator have the same asymptotic distribution. Section 3.4 provides rigorous consistency and asymptotic normality arguments for the parameterizations of Section 3.2. The proof of the main theorem is developed in Sections 3.5 and 3.6. Expressions for the asymptotic variance of the optimal simple estimator are derived in Section 3.7. Finally, a small simulation study is presented in Section 3.8 and a summary is given in Section 3.9.

3.2 Estimators for $(\beta_0, \beta_1, \beta_2)$

3.2.1 *Simple estimators*

In structural linear models, instrumental variable estimation usually proceeds from the assumption $E(Y \mid U, Z, W) = E(Y \mid U, Z)$. This assumption implies

$$\begin{aligned} E(Y \mid Z, W) &= E(E(Y \mid U, Z) \mid Z, W) \\ &= \beta_0 + \beta_1^T E(U \mid Z, W) + \beta_2^T Z \\ &= \beta_0 + \beta_1^T E(X \mid Z, W) + \beta_2^T Z. \end{aligned}$$

Then $(\beta_0, \beta_1, \beta_2)$ can be estimated using the regressions of Y and X on Z and W. For the probit model, it is not possible to directly recover $(\beta_0, \beta_1, \beta_2)$ using the strategy above, see Stefanski and Buzas (1992). In fact, writing the joint density as $f_{YXZW} = f_{Y|XZW} f_{X|ZW} f_{ZW}$ suggests consideration of the conditional distribution of Y given (X, Z, W). Using assumptions (3.2) and

(3.3) it follows that

$$E(Y \mid X = x, Z = z, W = w) = \Phi\left(\frac{\beta_0 + \beta_1^T \mu_{U|XZW} + \beta_2^T z}{\sqrt{1 + \beta_1^T \Sigma_{UU|XZW} \beta_1}}\right). \qquad (3.4)$$

The conditional densities of $(U, X) \mid (Z, W)$, $X \mid (Z, W)$ and $U \mid (X, Z, W)$ satisfy

$$\frac{f_{UX|ZW}}{f_{X|ZW}} = f_{U|XZW},$$

i.e., the regression of $UX \mid (Z, W)$ on $X \mid (Z, W)$ is equivalent to the regression of U on (X, Z, W). Then using properties of conditional normal distributions it follows that

$$\mu_{U|XZW} = (I - \Delta^T)(\beta_{U|\underline{1}WZ} + \beta_{U|1W\underline{Z}}^T z + \beta_{U|1\underline{W}Z}^T w) + \Delta^T x \qquad (3.5)$$

where

$$\Delta = (\Sigma_{UU|ZW} + \Sigma_{\delta\delta})^{-1} \Sigma_{UU|ZW} = \Sigma_{XX|ZW}^{-1} \Sigma_{UU|ZW}.$$

The distribution of (Y, X, Z, W) can be defined using the parameterization

$$\theta_0 = (\beta_0, \beta_1^T, \beta_2^T, \mathrm{vec}^T \Delta, \mathrm{vec}^T \beta_{X|\underline{1}Z\underline{W}}, \mathrm{vech}^T \Sigma_{XX|ZW})^T.$$

However, it is not immediately clear whether $(\beta_0, \beta_1, \beta_2)$ are identified. Analogous to the linear model of Chapter 2, we obtain a regression parameterization showing that $(\beta_0, \beta_1, \beta_2)$ are identified and suggesting a simple estimation scheme.

Note that the regression of Y on (X, Z, W) has the form

$$\Pr(Y = 1 \mid X = x, Z = z, W = w) = \Phi(\gamma_1 + \gamma_X^T x + \gamma_Z^T z + \gamma_W^T w)$$

where

$$\begin{aligned}
\gamma_1 &= r_\gamma(\beta_0 + \beta_1^T(I-\Delta^T)\beta_{X|\underline{1}ZW}) \\
\gamma_X &= r_\gamma \Delta \beta_1 \\
\gamma_Z &= r_\gamma(\beta_{X|1\underline{Z}W}(I-\Delta)\beta_1 + \beta_2) \\
\gamma_W &= r_\gamma \beta_{X|1Z\underline{W}}(I-\Delta)\beta_1
\end{aligned} \tag{3.6}$$

and

$$r_\gamma^{-1} = \sqrt{\beta_1^T \Delta^T \Sigma_{XX|ZW}(I-\Delta)\beta_1 + 1}.$$

If $\dim(X) = 1$ then $0 < \Delta \leq 1$ and since $0 < r_\gamma \leq 1$ it follows that $|\gamma_X| \leq |\beta_1|$, i.e., measurement error has attenuated the magnitude of the regression coefficient of X. Define $\theta_1 = (\gamma_1, \gamma_X^T, \gamma_Z^T, \gamma_W^T, \mathrm{vec}^T \beta_{X|\underline{1}Z\underline{W}}, \mathrm{vech}^T \Sigma_{XX|ZW})^T$. The present task is to show that $(\beta_0, \beta_1^T, \beta_2^T)^T$ is identified using the correspondence between θ_0 and θ_1. Using the relationship

$$\gamma_W = r_\gamma \beta_{X|1Z\underline{W}}(I-\Delta)\beta_1 \tag{3.7}$$

and assuming that $\mathrm{rank}(\beta_{X|1Z\underline{W}}) = \dim(X)$ it follows that

$$r_\gamma(I-\Delta)\beta_1 = \beta^-_{X|1Z\underline{W}}\gamma_W \tag{3.8}$$

and the choice of g-inverse is irrelevant. It easily follows from (3.6) and (3.8) that

$$\begin{aligned}
\beta_0 &= r_\gamma^{-1}(\gamma_1 - \beta^T_{X|\underline{1}ZW}\beta^-_{X|1Z\underline{W}}\gamma_W), \\
\beta_1 &= r_\gamma^{-1}(\gamma_X + \beta^-_{X|1Z\underline{W}}\gamma_W), \\
\beta_2 &= r_\gamma^{-1}(\gamma_Z - \beta_{X|1\underline{Z}W}\beta^-_{X|1Z\underline{W}}\gamma_W).
\end{aligned} \tag{3.9}$$

Finally, using $\gamma_X = r_\gamma \Delta\beta$ and (3.8) , the identities

$$r_\gamma^{-1} = \sqrt{\beta_1^T \Delta^T \Sigma_{XX|ZW}(I-\Delta)\beta_1 + 1}$$

$$= \sqrt{r_\gamma^{-1}\gamma_X^T \Sigma_{XX|ZW} r_\gamma^{-1} \beta^-_{X|1\underline{ZW}}\gamma_W + 1}$$

$$= r_\gamma^{-1}\sqrt{\gamma_X^T \Sigma_{XX|ZW}\beta^-_{X|1\underline{ZW}}\gamma_W + r_\gamma^2} \tag{3.10}$$

imply

$$r_\gamma = \sqrt{1 - \gamma_X^T \Sigma_{XX|ZW}\beta^-_{X|1\underline{ZW}}\gamma_W}. \tag{3.11}$$

Thus the model is identified in the parameterization θ_1. The parameters $(\beta_0, \beta_1^T, \beta_2^T)^T$ are identified through (3.9) and (3.11). A simple estimation scheme is suggested by (3.9) and (3.11):

- Obtain $\hat{\beta}_{X|1\underline{ZW}}$ and $\hat{\Sigma}_{XX|ZW}$ via linear regression of X on Z and W;
- Obtain $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_X^T, \hat{\gamma}_Z^T, \hat{\gamma}_W^T)^T$ via probit regression of Y on X, Z and W;
- Use (3.9) and (3.11) to obtain estimates of $\beta_i, i = 0, 1, 2$, by substitution.

It is important to realize that the relation

$$\hat{\gamma}_W = r_\gamma \hat{\beta}_{X|1\underline{ZW}}(I-\Delta)\beta_1 \tag{3.12}$$

is not necessarily consistent in the sense of Definition 1.3. If $\dim(W) > \dim(X)$, then (3.12) is almost surely inconsistent and the model is said to be over identified. The finite sample and asymptotic properties of the simple estimators defined above will depend on the choice of a g-inverse for $\hat{\beta}_{X|1\underline{ZW}}$. If $\dim(W) = \dim(X)$ then (3.12) is almost surely consistent since $\hat{\beta}_{X|1\underline{ZW}}$ is non-singular and the simple estimators are maximum likelihood estimators; see below. These

observations are identical to those noted in the linear model studied in Chapter 2. As in the linear measurement error model, we are able to identify an optimal g-inverse. This is explored in Section 3.3.

3.2.2 *Maximum likelihood estimators*

The key to constructing a parameterization amenable to the study of the maximum likelihood estimator is the fact that (3.7) imparts the restriction

$$\gamma_W \in C(\beta_{X|1\underline{ZW}}) \tag{3.13}$$

on the parameterization θ_1; compare with (2.2). Define λ via

$$\gamma_W = \beta_{X|1\underline{ZW}}\lambda$$

and let

$$\theta_2 = (\gamma_1, \gamma_X^T, \gamma_Z^T, \lambda^T, \text{vec}^T\beta_{X|1\underline{ZW}}, \text{vech}^T\Sigma_{XX|ZW})^T.$$

The parameterizations θ_1 and θ_2 differ in that θ_2 automatically incorporates the restriction (3.13). Maximizing the likelihood of $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ with respect to θ_2 yields maximum likelihood estimators provided the resulting estimates satisfy

$$0 < \gamma_X^T\Sigma_{XX|ZW}\lambda < 1, \tag{3.14}$$

see (3.10) and (3.11). Inspection of (3.6) yields no other restrictions. Under the assumed model, the probability that condition (3.14) is satisfied approaches one as n tends to infinity. Therefore, (3.14) can be ignored in the derivation of the asymptotic theory for the maximum likelihood estimator, see result A.3 in the appendix. By the invariance property of maximum likelihood estimators, the form of the maximum likelihood estimators for $(\beta_0, \beta_1^T, \beta_2^T)^T$ is given by (3.9)

and (3.11) with the change that λ replaces $\beta^{-}_{X|1Z\underline{W}}\gamma_W$. If $\dim(X) = \dim(W)$, (3.13) is satisfied and the parameterizations θ_1 and θ_2 are equivalent. In this case the simple estimators are maximum likelihood estimators. These observations are identical to those noted in the linear model.

3.3 The optimal simple estimator and the main theorem

Using (3.9)-(3.11) and the Δ-Theorem it is possible to obtain an expression for the variance of the asymptotic distribution of the simple estimators. When there are more instruments available than covariates measured with error, i.e., when $\dim(W) > \dim(X)$, the asymptotic variances, denoted $\Sigma(G)$, will depend on the choice of generalized inverse $G^T \equiv \beta^{-}_{X|1Z\underline{W}}$. Under the additional assumptions of Theorem 3.1 it can be shown that $\Sigma(G) \geq_{p.d.} \Sigma(\underline{G})$ where

$$\underline{G}^T = \left\{\beta^T_{X|1Z\underline{W}} M^{-1} \beta_{X|1Z\underline{W}}\right\}^{-1} \beta^T_{X|1Z\underline{W}} M^{-1}, \tag{3.15}$$

and $M = \mathrm{ACOV}(\hat{\gamma}_W - \hat{\beta}_{X|1Z\underline{W}}\lambda, \hat{\gamma}_W - \hat{\beta}_{X|1Z\underline{W}}\lambda)$. Simple estimators defined with $\hat{\underline{G}}$ where $\hat{\underline{G}}$ satisfies $\hat{\underline{G}}^T \hat{\beta}_{X|1Z\underline{W}} = I$ and $\hat{\underline{G}} \xrightarrow{P} \underline{G}$ are termed optimal simple estimators.

The optimality of $\hat{\underline{G}}$ can be proved directly as in Theorem 2.1. We obtain the result indirectly by showing that the maximum likelihood estimator is asymptotically normal with covariance $\Sigma(\underline{G})$. However, the result can also be obtained by formal manipulation of asymptotic expressions. Since $\gamma_W = \beta_{X|1Z\underline{W}}\lambda$ it follows that

$$\hat{\gamma}_W = \hat{\beta}_{X|1Z\underline{W}}\lambda + \epsilon \tag{3.16}$$

where $\epsilon = (\hat{\gamma}_W - \hat{\beta}_{X|1Z\underline{W}}\lambda) - (\gamma_W - \beta_{X|1Z\underline{W}}\lambda)$. Asymptotically, $\sqrt{n}\epsilon$ converges in distribution to

$$N\left\{0, \mathrm{ACOV}(\hat{\gamma}_W - \hat{\beta}_{X|1Z\underline{W}}\lambda, \hat{\gamma}_W - \hat{\beta}_{X|1Z\underline{W}}\lambda)\right\}.$$

It follows that $\hat{\lambda} = \hat{\underline{G}}^T \hat{\gamma}_W$ can be thought of as the generalized least squares estimator of λ in the (approximately) linear model (3.16).

In what follows estimators of parameters in θ_1 and θ_2 are denoted by placing hats and tildes over the parameters, respectively. Recall the relationships

$$\begin{aligned} \beta_0 &= r_\gamma^{-1}(\gamma_1 - \beta_{X|1\underline{Z}W}^T \lambda) \\ \beta_1 &= r_\gamma^{-1}(\gamma_X + \lambda) \\ \beta_2 &= r_\gamma^{-1}(\gamma_Z - \beta_{X|1\underline{Z}W} \lambda) \ , \end{aligned} \tag{3.17}$$

where $r_\gamma = \sqrt{1 - \gamma_X^T \sigma^2_{XX|ZW} \lambda}$. The optimal simple estimators are obtained by maximizing the likelihood of $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ with respect to θ_1 and substituting into (3.17) using

$$\hat{\lambda} = \hat{\underline{G}}^T \hat{\gamma}_W. \tag{3.18}$$

The maximum likelihood estimators are obtained by maximizing the likelihood of $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ with respect to θ_2 and substituting into (3.17) using

$$\tilde{\lambda} = \tilde{\lambda}. \tag{3.19}$$

The simple estimators are easy to compute. The required computations consist of probit and linear regressions, which can be done with most statistical software packages. Theorem 3.1 tells us our simple estimation strategy also has optimal asymptotic properties. To simplify notation and for ease of presentation, the theorem is stated and proved for the case $\dim(X) = 1$.

THEOREM 3.1. *In addition to (3.1) -(3.3) assume that*

$$E(W) = 0, \qquad COV(W, Z) = \Sigma_{WZ} = 0, \tag{3.20}$$

and $dim(X) = 1$. *Then the optimal simple estimators, defined through (3.17) and (3.18) , and the maximum likelihood estimators, defined through (3.17) and (3.19) , are both consistent and asymptotically normally distributed. Furthermore, the covariance matrices of the limiting distributions are identical.*

Before proceeding with the proof it is necessary to establish consistency and asymptotic normality of $\hat{\theta}_1$ and $\tilde{\theta}_2$. This is accomplished in Section 3.4. The proof of Theorem 3.1 is developed in Section 3.5. In Section 3.6 it is shown no generality is lost assuming (3.20) , provided one appropriately transforms the instrumental variables.

3.4 Asymptotic distributions

In this section consistency and asymptotic normality of $\hat{\theta}_1$ and $\tilde{\theta}_2$ are established. Before delving into the theorems, it is convenient to establish a notation for to the likelihood functions and their derivatives. Define

$$\Phi_i = \Phi(\gamma_1 + \gamma_X X_i + \gamma_Z^T Z_i + \gamma_W^T W_i)$$

$$\mu_i = \beta_{X|1ZW} + \beta_{X|1\underline{Z}W}^T Z_i + \beta_{X|1Z\underline{W}}^T W_i.$$

and Φ_i' similarly where $\Phi'(u) = \frac{d}{du}\Phi(u)$. Let

$$P_i = \frac{\Phi_i'}{\Phi_i(1-\Phi_i)},$$

and P_i' similarly where $P' = \frac{d}{du}P(u)$. The density of an observation $\{Y_i, X_i, Z_i, W_i\}$ in the parameterization θ_1 is $f_{Y|XZW}f_{X|ZW}f_{ZW}$ where

$$f_{Y|XZW}f_{X|ZW} = \Phi_i^{Y_i}(1-\Phi_i)^{1-Y_i}\sigma_{XX|ZW}^{-1}\Phi'\left(\frac{X_i-\mu_i}{\sigma_{XX|ZW}}\right) \equiv f_i(\theta_1). \quad (3.21)$$

Since f_{ZW} has no role in the estimation of θ_1, the relevant part of the log-likelihood of the observable data $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$, multiplied through by $\frac{1}{n}$,

is

$$l_n(\theta_1) = \frac{1}{n}\sum_{i=1}^{n} \log f_i(\theta_1).$$

The relevant piece of the log-likelihood in terms of θ_2 is the same except $\beta_{X|1Z\underline{W}}\lambda$ replaces γ_W. This log-likelihood is written $\tilde{l}_n(\theta_2)$. Then $\hat{\theta}_1$ and $\tilde{\theta}_2$ maximize $l_n(\theta_1)$ and $\tilde{l}_n(\theta_2)$ respectively. The derivative of $l_n(\theta_1)$ with respect to θ_1 is

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\theta_1; Y_i, X_i, Z_i, W_i) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \psi_1(\theta_1; Y_i, X_i, Z_i, W_i) \\ \psi_2(\theta_1; Y_i, X_i, Z_i, W_i) \\ \psi_3(\theta_1; Y_i, X_i, Z_i, W_i) \end{pmatrix}$$

where, suppressing function arguments,

$$\psi_1 = (Y_i - \Phi_i)P_i \begin{pmatrix} 1 \\ X_i \\ Z_i \\ W_i \end{pmatrix}, \tag{3.22}$$

$$\psi_2 = \sigma^{-2}_{XX|ZW}(X_i - \mu_i) \begin{pmatrix} 1 \\ Z_i \\ W_i \end{pmatrix}$$

and

$$\psi_3 = \frac{1}{2\sigma^4_{XX|ZW}}(X_i - \mu_i)^2 - \frac{1}{2\sigma^2_{XX|ZW}}.$$

When working with the parameterization θ_2, Φ_i and P_i are defined as before except that γ_W is replaced by $\beta_{X|1Z\underline{W}}\lambda$. The derivative of $\tilde{l}_n(\theta_2)$ with respect to θ_2 is

$$\frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}(\theta_2; Y_i, X_i, Z_i, W_i) = \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} \tilde{\psi}_1(\theta_2; Y_i, X_i, Z_i, W_i) \\ \tilde{\psi}_2(\theta_2; Y_i, X_i, Z_i, W_i) \end{pmatrix}$$

where

$$\tilde{\psi}_1 = (Y_i - \Phi_i)P_i \begin{pmatrix} 1 \\ X_i \\ Z_i \\ \beta^T_{X|1Z\underline{W}} W_i \\ 0 \\ 0 \\ \lambda W_i \end{pmatrix} + \frac{1}{2\sigma^2_{XX|ZW}}(X_i - \mu_i) \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ Z_i \\ W_i \end{pmatrix},$$

and

$$\tilde{\psi}_2 = \frac{1}{2\sigma^4_{XX|ZW}}(X_i - \mu_i)^2 - \frac{1}{2\sigma^2_{XX|ZW}}.$$

Note that $\tilde{\theta}_2$ can be obtained by maximizing $l_n(\theta_1)$ with the constraint $h(\theta_1) = 0$ where

$$h(\theta_1) \equiv (I - \beta_{X|1\underline{ZW}}\beta^-_{X|1\underline{ZW}})\gamma_W. \tag{3.23}$$

Why then bother with the parameterization θ_2? The reason is that the parameterization θ_2 facilitates a comparison between the asymptotic covariance matrices of the optimal simple estimator (derived from θ_1) and the maximum likelihood estimators (derived from θ_2). Nevertheless, consistency of $\tilde{\theta}_2$ is established through consideration of $l_n(\theta_1)$ subject to $h(\theta_1) = 0$.

The log-likelihood $l_n(\theta_1)$ is the sum of the component from the probit regression of Y on (X, Z, W) and the linear regression of X on (Z, W). That is

$$l_n(\theta_1) = \frac{1}{n}\sum_{i=1}^{n} \log f_{Y_i|X_iZ_iW_i}(\gamma) + \frac{1}{n}\sum_{i=1}^{n} \log f_{X_i|Z_iW_i}(\beta_{X|1\underline{ZW}}, \sigma^2_{XX|ZW})$$

where $\gamma = (\gamma_1, \gamma_X, \gamma_Z^T, \gamma_W^T)^T$. Since the constraint $h(\theta_1) = 0$ is independent of $\sigma^2_{XX|ZW}$, both the constrained and unconstrained maximizations of $l_n(\theta_1)$ can proceed by first maximizing $l_n(\theta_1)$ with respect to $\sigma^2_{XX|ZW}$. Let $\tau = (\gamma^T, \beta^T_{X|1\underline{ZW}})^T$ so that $\theta_1 = (\tau^T, \sigma^2_{XX|ZW})^T$. Then considering τ fixed,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)^2 = \arg \max_{\sigma^2_{XX|ZW}} l_n(\theta_1)$$

so that

$$\max_{\sigma^2_{XX|ZW}} l_n(\theta_1) \equiv m_n(\tau) = \frac{1}{n}\sum_{i=1}^{n} \log f_{Y_i|X_iZ_iW_i}(\gamma) - \frac{1}{2}\log\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)^2\right] - \frac{1}{2}\{\log(2\pi) + 1\}.$$

The consistency proof exploits the concavity of $\sum_{i=1}^{n} \log f_{Y_i|X_i Z_i W_i}(\gamma)$ and the convexity of $\sum_{i=1}^{n}(X_i - \mu_i)^2$.

DEFINITION 3.1: Let E be a convex set. A function f on E is concave if $f(\lambda x + (1-\lambda)y) \geq \lambda f(x) + (1-\lambda)f(y)\ \forall x, y \in E\ , 0 \leq \lambda \leq 1$. A function f is convex if $-f$ is concave.

The following outlines the argument used to establish consistency. A zero subscript on an expectation operator denotes evaluation at the parameter value $\theta_1^0 = \{\tau^{0^T}, (\sigma^0_{XX|ZW})^2\}^T$. The symbol λ is often used in the literature concerning concave and convex functions. It is used here in conjunction with such functions and should not be confused with the *parameter* λ defined earlier.

i) There exists τ_n such that $m_n(\tau_n) > m_n(\tau)$ when $\tau \neq \tau_n$ and $m_n(\tau_n + \lambda_1\tau) > m_n(\tau_n + \lambda_2\tau)$ when $0 \leq \lambda_1 < \lambda_2$.

ii) $m_n(\tau) \xrightarrow{a.s.} m(\tau)$ uniformly for $\tau \in \mathcal{C}$ where $\mathcal{C}$ is a compact set and $m(\tau) = E_0\{m_n(\tau)\}$. Furthermore, $m(\tau^0) > m(\tau)$ when $\tau \neq \tau^0$.

iii) Combine i) and ii) to show $\hat{\theta}_1$ and $\tilde{\theta}_1$ are consistent for θ_1^0.

LEMMA 3.1. *Let $f(\cdot)$ be a function and suppose $\exists\ x_0 \ni\ f(x_0) > f(x)$ whenever $x \neq x_0$ and $x \in E$, E a convex set. Also suppose x_0 satisfies $h(x_0) = 0$ for some continuous function $h(\cdot)$. Let $f_n(\cdot)$ be a sequence of functions such that for each n, $\exists\ x_n \ni f_n(x_n) > f_n(x)$ whenever $x \neq x_n$ and $x \in E$. Also suppose that $f_n(x_n + \lambda_1 x) > f_n(x_n + \lambda_2 x)$ whenever $0 \leq \lambda_1 < \lambda_2$ and that $f_n(\cdot)$ converges uniformly to $f(\cdot)$ on any compact set $\mathcal{C} \subset E$. Fix any $\epsilon > 0$ and define $A_1 = \{x : \|x - x_0\| \leq \epsilon\}$. Then $\exists\ N \ni n > N$ implies*

$$\sup\{f_n(x) : x \in E, h(x) = 0\} = \sup\{f_n(x) : x \in A_1, h(x) = 0\}.$$

PROOF: First note that $h(x_0) = 0$ and $h(\cdot)$ continuous imply that $\{x : x \in$

$A_1, h(x) = 0\}$ is non-empty. Define $A_2 = \{x : \epsilon < \|x - x_0\| \leq 2\epsilon\}$. Then $\exists\ \delta > 0 \ \ni \ f(x) < f(x_0) - \delta$ for $x \in A_2$. To see why, define $\bar{A}_2 = \{x : \epsilon \leq \|x - x_0\| \leq 2\epsilon\}$. Then $\exists\ \bar{x} \in \bar{A}_2 \ \ni \ f(\bar{x}) = \sup_{x \in \bar{A}_2} f(x)$. But $f(x_0) > f(\bar{x})$ by assumption. Take $\delta = \frac{1}{2}(f(x_0) - f(\bar{x}))$.

Next, by uniform convergence, $\exists\ N \ \ni \ n > N$ implies

$$|f_n(x) - f(x)| < \frac{\delta}{4} \quad \forall\ x \in A_1 \cup A_2.$$

Then $f_n(x_0) > f(x_0) - \frac{\delta}{4}$ and, for $x \in A_2$, $f_n(x) < f(x_0) - \frac{3}{4}\delta$. Thus $f_n(x_0) > f_n(x)$ for $x \in A_2$.

Recall $f_n(x_n) > f_n(x)$. Suppose $x_n \in (A_1 \cup A_2)^c \cap E$. It will be shown this leads to a contradiction. With $y = x_0 - x_n$, we have $x_0 = x_n + y$ and clearly $\exists\ \lambda \in (0,1) \ni x^* = x_n + \lambda y \in A_2$. Then

$$f_n(x^*) = f_n(x_n + \lambda y) > f_n(x_n + y) = f_n(x_0)$$

which contradicts $f_n(x_0) > f_n(x)$ for $x \in A_2$. •

LEMMA 3.2. *If $f(\cdot)$ is concave and $\exists\ x_0 \in E \ni f(x_0) > f(x)\ \forall\ x \neq x_0$, then $f(x_0 + \lambda_1 x) > f(x_0 + \lambda_2 x)$ whenever $0 \leq \lambda_1 < \lambda_2$ and $\{x_0 + \lambda_1 x\},\ \{x_0 + \lambda_2 x\} \in E$.*

PROOF: Since $f(\cdot)$ is concave, for all $0 \leq \alpha \leq 1$,

$$f(\alpha x_0 + (1-\alpha)(x_0 + \lambda_2 x)) \geq \alpha f(x_0) + (1-\alpha) f(x_0 + \lambda_2 x)$$

so that

$$f(x_0 + (1-\alpha)\lambda_2 x) \geq \alpha f(x_0) + (1-\alpha) f(x_0 + \lambda_2 x) > f(x_0 + \lambda_2 x).$$

Take $1 - \alpha = \frac{\lambda_1}{\lambda_2}$ and the result follows. •

Remark: An analogous result holds for convex functions.

LEMMA 3.3. *For each* n, $\exists\ \tau_n \ni m_n(\tau_n) > m_n(\tau)$ *when* $\tau \neq \tau_n$. *Furthermore,* $m_n(\tau_n + \lambda_1\tau) > m_n(\tau_n + \lambda_2\tau)$ *whenever* $0 \leq \lambda_1 < \lambda_2$.

PROOF: Wedderburn (1976) has established that $\frac{1}{n}\sum_{i=1}^{n} \log f_{Y_i|X_iZ_iW_i}(\gamma)$ is concave in γ and has a unique finite maximum. It is well known $\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)^2$ is strictly convex in $\beta_{X|1ZW}$ and has a unique minimum. Since $-\log(\cdot)$ is a strictly decreasing continuous function, $-\log\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)^2$ has a unique maximum. An application of Lemma 3.2, again noting $-\log(\cdot)$ is strictly decreasing, completes the proof. •

THEOREM 3.2. *Let* E *be an open convex subset of* $\Re^p$ *and let* $l_1, l_2, \ldots$ *be a sequence of random concave functions on* E *such that* $\forall\ x \in E$, $l_n(x) \xrightarrow{a.s.} l(x)$ *as* $n \to \infty$, *where* $l(\cdot)$ *is some real function on* E. *Then* $l(\cdot)$ *is concave and for all compact* $C \subset E$,

$$\sup_{x \in C} |l_n(x) - l(x)| \xrightarrow{a.s.} 0 \quad as \quad n \to \infty.$$

PROOF: The proof is an immediate almost sure generalization of Theorem 10.8 in Rockafellar (1970), see Andersen and Gill (1982). •

THEOREM 3.3. *Let* $\{Y_i, X_i, Z_i, W_i\}$, $1 \leq i \leq n$ *be i.i.d. each with density* $f(\theta_1^0)f_{ZW}$ *given by (3.21). Assume* θ_1^0 *is a fixed element of a convex parameter space* Θ *and satisfies* $h(\theta_1^0) = 0$. *The function* $h(\cdot)$ *is defined by (3.23). Let* $\hat{\theta}_1$ *maximize* $l_n(\theta_1)$ *and* $\tilde{\theta}_1$ *maximize* $l_n(\theta_1)$ *subject to* $h(\theta_1) = 0$. *Then, as* $n \to \infty$

$$\hat{\theta}_1 \xrightarrow{a.s.} \theta_1^0$$

$$\tilde{\theta}_1 \xrightarrow{a.s.} \theta_1^0.$$

PROOF: It is assumed the model is identified. By the strong law of large numbers and Theorem 3.2, as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} \log f_{Y_i|X_i Z_i W_i}(\gamma) \xrightarrow{a.s.} E_0 \log f_{Y|XZW}(\gamma) \quad \text{uniformly for} \quad \gamma \in \mathcal{C}_1$$

where $\mathcal{C}_1$ is a compact set. Using result ii) of Rao (1973, p.59),

$$E_0 \log f_{Y|XZW}(\gamma^0) > E_0 \log f_{Y|XZW}(\gamma) \quad \text{when} \quad \gamma \neq \gamma^0.$$

Again using the strong law of large numbers and Theorem 3.2,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)^2 \xrightarrow{a.s.} (\sigma^0_{XX|ZW})^2 + E_0\{a^T R a\} \equiv g(\beta_{X|\underline{1ZW}})$$

uniformly for $\beta_{X|\underline{1ZW}} \in \mathcal{C}_2$ where $\mathcal{C}_2$ is a compact set and

$$a = (\beta^0_{X|\underline{1}ZW} - \beta_{X|\underline{1}ZW}, \beta^{0T}_{X|1\underline{Z}W} - \beta^T_{X|1\underline{Z}W}, \beta^{0T}_{X|1Z\underline{W}} - \beta^T_{X|1Z\underline{W}})^T$$

and

$$R = \begin{pmatrix} 1 \\ Z \\ W \end{pmatrix} (1, \quad Z, \quad W).$$

Clearly,

$$0 < (\sigma^0_{XX|ZW})^2 = g(\beta^0_{X|\underline{1ZW}}) < g(\beta_{X|\underline{1ZW}}) < M < \infty$$

when $\beta_{X|\underline{1ZW}} \neq \beta^0_{X|\underline{1ZW}}$ and $\beta_{X|\underline{1ZW}} \in \mathcal{C}_2$ where M is a constant. Now, $-\log(\cdot)$ is strictly decreasing and uniformly continuous on $[(\sigma^0_{XX|ZW})^2, M]$ so that

$$-\log\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_i)^2\right\} \xrightarrow{a.s.} -\log\left\{g(\beta_{X|\underline{1ZW}})\right\} \quad \text{uniformly for} \quad \beta_{X|\underline{1ZW}} \in \mathcal{C}_2,$$

and $-\log\{g(\beta^0_{X|\underline{1ZW}})\} > -\log\{g(\beta_{X|\underline{1ZW}})\}$ when $\beta_{X|\underline{1ZW}} \neq \beta^0_{X|\underline{1ZW}}$. We have argued that

$$m_n(\tau) \xrightarrow{a.s.} E_0 \log f_{Y|XZW}(\gamma) + g(\beta_{X|\underline{1ZW}}) = m(\tau)$$

uniformly for $\tau \in \mathcal{C}$ a compact set and that $m(\tau^0) > m(\tau)$ whenever $\tau \neq \tau^0$. Then by Lemma 3.1, $\hat{\tau}$ and $\tilde{\tau}$ are ultimately trapped in the set $A_1 = \{\tau : \|\tau - \tau^0\| \leq \epsilon\}$. Since $\epsilon > 0$ is arbitrary, strong consistency follows. Strong consistency of $\hat{\sigma}^2_{XX|ZW}$ and $\tilde{\sigma}^2_{XX|ZW}$ follow from the strong law of large numbers and the consistency of $\hat{\tau}$ and $\tilde{\tau}$. •

Remarks: Since, excepting degenerate cases in the data, the probit and normal theory log-likelihood are strictly concave and possess unique, finite maxima, $\hat{\theta}_1$ can be obtained as the unique root of $\frac{\partial}{\partial \theta_1} l_n(\theta_1) = 0$. The maximum of $l_n(\theta_1)$ subject to $h(\theta_1) = 0$ can be obtained by maximizing $\tilde{l}_n(\theta_2)$. Since $\tilde{l}_n(\theta_2)$ is a differentiable function, it's maximum, which with probability one is finite for n large, will be among the roots of $\frac{\partial}{\partial \theta_2} \tilde{l}_n(\theta_2) = 0$.

THEOREM 3.4. *Let l be a function on $\mathcal{X} \times \Theta$ where $\mathcal{X}$ is a Euclidean space and Θ is a compact subset of a Euclidean space. Let $l(x, \theta)$ be a continuous function of θ for each x and a measurable function of x for each θ. Assume also that $\|l(x, \theta)\| \leq g(x)$ for all x and θ, where g is integrable with respect to a distribution function F on $\mathcal{X}$. If $X_1, \ldots, X_n$ is a random sample from F, then as $n \to \infty$,*

$$\frac{1}{n} \sum_{i=1}^{n} l(X_i, \theta) \xrightarrow{a.s.} \int l(x, \theta) dF(x) \quad \text{uniformly for all} \quad \theta \in \Theta.$$

PROOF: Jennrich (1969). •

THEOREM 3.5. *Asymptotic normality of M-estimators. Let $X_1, \ldots, X_n$ be i.i.d., possibly vector valued, each with density $f(x, \theta^0)$ where θ^0 is in the interior of $\Theta \subseteq \Re^p$ and suppose the following regularity conditions hold.*

i) $\hat{\theta}$ solves $\sum_{i=1}^{n} \psi(X_i, \theta) = 0$ where $E_\theta \{\psi(X_i, \theta)\} = 0$, $tr\left[E(\psi\psi^T)\right] < \infty$.

ii) $\hat{\theta} \xrightarrow{a.s.} \theta^0$.

iii) $\dot{\psi}(x,\theta) = \frac{\partial}{\partial\theta}\psi(x,\theta)$ *exists and is continuous in* θ.

iv) $|\{\dot{\psi}(x,\theta)\}_{i,j}| \leq g_{i,j}(x)$ *for all* $x, 1 \leq i,j \leq p$ *and* $\theta \in \mathcal{C}$ *where* $Eg_{i,j}(X) < \infty$, $\mathcal{C}$ *a compact subset of* Θ.

v) $\det\left(E_0\left\{\dot{\psi}(X_1,\theta^0)\right\}\right) \neq 0$.

Then

$$\sqrt{n}(\hat{\theta} - \theta^0) \xrightarrow{D} N\left(0, A^{-1}BA^{-T}\right)$$

where $A = E_0\left\{\dot{\psi}(X_1,\theta^0)\right\}$ *and* $B = E_0\left\{\psi(X_1,\theta^0)\psi(X_1,\theta^0)^T\right\}$.

PROOF: Let $\rho \in \Re^p$ be a fixed but otherwise arbitrary vector. The mean value theorem, applicable by iii), together with i) imply

$$0 = \sum_{i=1}^{n} \rho^T \psi(X_i,\hat{\theta}) = \sum_{i=1}^{n} \rho^T \psi(X_i,\theta^0) + \sum_{i=1}^{n} \rho^T \dot{\psi}(X_i,\bar{\theta}_{n,\rho})(\hat{\theta} - \theta^0)$$

where $\bar{\theta}_{n,\rho}$ is such that $\|\bar{\theta}_{n,\rho} - \theta^0\| \leq \|\hat{\theta} - \theta^0\|$. Rearranging,

$$-\left[\frac{1}{n}\sum_{i=1}^{n} \rho^T \dot{\psi}(X_i,\bar{\theta}_{n,\rho})\right]\sqrt{n}(\hat{\theta} - \theta^0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \rho^T \psi(X_i,\theta^0).$$

Condition i) implies the Central Limit Theorem is applicable, i.e. that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \rho^T \psi(X_i,\theta^0) \xrightarrow{D} N\{0, \rho^T(E_0\psi\psi^T)\rho\}.$$

Conditions iii) and iv) allow application of Theorem 3.4 so that

$$\frac{1}{n}\sum_{i=1}^{n} \rho^T \dot{\psi}(X_i,\theta) \xrightarrow{a.s.} E_0\left\{\rho^T \dot{\psi}(X_1,\theta)\right\}$$

uniformly on every compact set $\mathcal{C} \subset \Theta$. Since $\hat{\theta} \xrightarrow{a.s.} \theta^0$, it follows $\bar{\theta}_{n,\rho} \xrightarrow{a.s.} \theta^0$ and that we may restrict attention to a compact set $\mathcal{C}$ with θ^0 in the interior. These facts together with the continuity of $E_0\left\{\dot{\psi}(X,\theta)\right\}$ imply

$$\frac{1}{n}\sum_{i=1}^{n} \rho^T \dot{\psi}(X_i,\bar{\theta}_{n,\rho}) \xrightarrow{a.s.} E_0\left\{\rho^T \dot{\psi}(X_1,\theta^0)\right\}.$$

The theorem follows from application of Slutsky's theorem and an appeal to the Cramer-Wold device. •

LEMMA 3.4. *There exists constants* k_1 *and* k_2 *such that the function* $P(x) = \Phi'(x)/(\Phi(x)(1-\Phi(x)))$ *satisfies*

$$P(x) < k_1 + k_2|x| \quad \forall x \in \Re.$$

PROOF: The inequality

$$\frac{\Phi'(x)}{1-\Phi(x)} < x + \frac{1}{x}\,, \quad x > 0$$

associated with Mill's ratio is given in Kendall and Stuart (1977, p.155). Fix any $N > 0$. For $0 \leq x \leq N$, $P(x)$ is obviously bounded, say by M_1. For $x > N$, $\exists\ M_2 \in (0,1)\ \ni\ \Phi(x) > M_2$ whenever $x > N$. Hence, for $x \geq 0$,

$$P(x) < M_1 I(0 \leq x \leq N) + \frac{1}{M_2}\left(x + \frac{1}{x}\right) I(x > N) < M_1 + \frac{1}{M_2}\left(x + \frac{1}{N}\right).$$

Since $P(x)$ is an even function, the result follows with $k_1 = M_1 + \frac{1}{M_2 N}$ and $k_2 = \frac{1}{M_2}$. •

COROLLARY 3.1. *Let* $\{Y_i, X_i, Z_i, W_i\}_{i=1}^n$ *be i.i.d., each with density* $f(\theta_1^0) f_{ZW}$ *in (3.21). Let* $\hat{\theta}_1$ *and* $\tilde{\theta}_2$ *be strongly consistent roots of the likelihood equations*

$$\frac{1}{n}\sum_{i=1}^{n} \psi(\theta_1) = 0 \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}(\theta_2) = 0$$

respectively. Then

$$\sqrt{n}(\hat{\theta}_1 - \theta_1^0) \xrightarrow{D} N(0, \Sigma)$$

and

$$\sqrt{n}(\tilde{\theta}_2 - \theta_1^0) \xrightarrow{D} N\left(0, \tilde{\Sigma}\right)$$

where $\Sigma = diag(\Sigma_1, \Sigma_2, \Sigma_3)$, $\tilde{\Sigma} = diag(\tilde{\Sigma}_1, \tilde{\Sigma}_2)$ *and*

$$\begin{aligned}
\Sigma_1 &= \left(E\psi_1\psi_1^T\right)^{-1}, \\
\Sigma_2 &= \left(E\psi_2\psi_2^T\right)^{-1}, \\
\Sigma_3 &= \left(E\psi_3\psi_3^T\right)^{-1}, \\
\tilde{\Sigma}_1 &= \left(E\tilde{\psi}_1\tilde{\psi}_1^T\right)^{-1}, \\
\tilde{\Sigma}_2 &= \left(E\tilde{\psi}_2\tilde{\psi}_2^T\right)^{-1}.
\end{aligned}$$

PROOF: The first task is to check conditions i)-v) of Theorem 3.5. Condition i) is satisfied provided ψ and $\tilde{\psi}$ satisfy

$$\text{tr}\left[E\left(\psi\psi^T\right)\right] < \infty \quad \text{and} \quad \text{tr}\left[E\left(\tilde{\psi}\tilde{\psi}^T\right)\right] < \infty.$$

That the above is true follows from Lemma 3.4 and the fact that moments of all orders are finite for the normal distribution. Conditions ii) and iii) are evidently satisfied. Condition iv) is satisfied for ψ if it is satisfied for $\tilde{\psi}$. Define $\gamma = (\gamma_1, \gamma_X, \gamma_Z^T, \gamma_W^T)^T$ and $V = (Y, X, Z, W)$. Then $\gamma^T V = \gamma_1 + \gamma_X X + \gamma_Z^T Z + \gamma_W^T W$. Since attention is restricted to a compact set, $\exists M_1 \ni \|\gamma\| < M_1 < \infty$. From Lemma 3.4,

$$(Y_i - \Phi_i)P_i < k_1 + k_2|\gamma^T V| \leq k_1 + k_2\|\gamma\|\|V\| < k_1 + k_2 M\|V\|.$$

Next,

$$P'(x) = -\frac{x\Phi'(x)}{\Phi(1-\Phi)} - \{P(x)\}^2(1 - 2\Phi(x)).$$

By Lemma 3.4 there exists constants $\bar{k}_1$,$\bar{k}_2$ and $\bar{k}_3$ such that

$$|P'(x)| < \bar{k}_1 + \bar{k}_2|x| + \bar{k}_3|x|^2,$$

and thus $|P_i'| < \bar{k}_1 + \bar{k}_2 M\|V\| + \bar{k}_3 M^2\|V\|^2$. Finally, it is not hard to show that $|\Phi'(x)P(x)|$ is bounded. The above imply the absolute value of each element of $\dot{\tilde{\psi}}$ and $\dot{\psi}$ are bounded by integrable functions of the data only, so that condition iv) is satisfied. It remains to check condition v).

Define

$$E_0\left\{\frac{\partial}{\partial\gamma}\psi_1(\theta^0)\right\} = E_0\left\{\frac{\partial}{\partial\gamma}\psi_1(\theta_1)_{|_{\theta_1=\theta^0}}\right\}.$$

Similar notation is used for the other components. First note that

$$E_0\left\{\dot{\psi}(\theta^0)\right\} = \mathrm{diag}\Bigg[E_0\left\{\frac{\partial}{\partial\gamma}\psi_1(\theta^0)\right\}, E_0\left\{\frac{\partial}{\partial\beta_{X|\underline{1ZW}}}\psi_2(\theta^0)\right\}, E_0\left\{\frac{\partial}{\partial\sigma^2_{XX|ZW}}\psi_3(\theta^0)\right\}\Bigg].$$

Assuming the covariance matrix Σ of the covariates $\{X, Z, W\}$ is non-singular, it follows that $E_0\left\{\frac{\partial}{\partial\beta_{X|\underline{1ZW}}}\psi_2(\theta^0)\right\}$ and $E_0\left\{\frac{\partial}{\partial\sigma^2_{XX|ZW}}\psi_3(\theta^0)\right\}$ are non-singular. That $E_0\left\{\frac{\partial}{\partial\gamma}\psi_1(\theta^0)\right\}$ is non-singular also follows from the non-singularity of Σ (details are in the proof of Lemma 3.7).

Define $\eta = (\gamma_1, \gamma_X, \gamma_Z^T, \lambda)^T$. Then

$$E_0\left\{\dot{\tilde{\psi}}(\theta^0)\right\} = \mathrm{diag}\left(E_0\left\{\frac{\partial}{\partial\eta, \beta_{X|\underline{1ZW}}}\tilde{\psi}_1(\theta^0)\right\}, E_0\left\{\frac{\partial}{\partial\sigma^2_{XX|ZW}}\tilde{\psi}_2(\theta^0)\right\}\right).$$

It is obvious that $E_0\left\{\frac{\partial}{\partial\sigma^2_{XX|ZW}}\tilde{\psi}_2(\theta^0)\right\}$, a scalar, is non-singular. It can be shown that non-singularity of Σ implies $E_0\left\{\frac{\partial}{\partial\eta,\beta_{X|\underline{1ZW}}}\tilde{\psi}_1(\theta^0)\right\}$ is non-singular. The argument is not given, but follows fairly easily from (3.27)-(3.29). Finally, note that $E\psi\psi^T = -E\dot{\psi}$ and $E\tilde{\psi}\tilde{\psi} = -E\dot{\tilde{\psi}}$. •

3.5 Proof of the main theorem

The purpose of this section is to present a proof of Theorem 3.1. Define

$$\hat{\mu}_S = \begin{pmatrix} \hat{\gamma}_1 \\ \hat{\gamma}_X \\ \hat{\gamma}_Z \\ \hat{\underline{G}}^T \hat{\gamma}_W \\ \hat{\beta}_{X|\underline{1}ZW} \\ \hat{\beta}_{X|1\underline{Z}W} \end{pmatrix} \quad \text{and} \quad \tilde{\mu}_{ML} = \begin{pmatrix} \tilde{\gamma}_1 \\ \tilde{\gamma}_X \\ \tilde{\gamma}_Z \\ \tilde{\lambda} \\ \tilde{\beta}_{X|\underline{1}ZW} \\ \tilde{\beta}_{X|1\underline{Z}W} \end{pmatrix}. \tag{3.24}$$

The estimators $\hat{\sigma}^2_{XX|ZW}$ and $\tilde{\sigma}^2_{XX|ZW}$ are omitted. It is apparent from Corollary 3.1 that $\hat{\sigma}^2_{XX|ZW}$ and $\tilde{\sigma}^2_{XX|ZW}$ have the same limiting distribution and are independent of $\hat{\mu}_S$ and $\tilde{\mu}_{ML}$. The estimators $\hat{\beta}_{X|1Z\underline{W}}$ and $\tilde{\beta}_{X|1Z\underline{W}}$ are omitted since $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ depends on $\hat{\beta}_{X|1Z\underline{W}}$ only through $\hat{\underline{G}}^T \hat{\gamma}_W$ and $(\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2)$ do not involve $\tilde{\beta}_{X|1Z\underline{W}}$. In fact, it can be shown that the asymptotic variances of $\hat{\beta}_{X|1Z\underline{W}}$ and $\tilde{\beta}_{X|1Z\underline{W}}$ are different.

Then by examination of the right hand side of (3.17) , it is apparent that the limiting distribution of the optimal simple estimators and maximum likelihood estimators are the same provided the limiting distributions of $\hat{\mu}_S$ and $\tilde{\mu}_{ML}$ are the same. Using Theorem 3.3 it is easy to show that $\hat{\mu}_S$ and $\tilde{\mu}_{ML}$ are both consistent. By Corollary 3.1, $\tilde{\mu}_{ML}$ is asymptotically normal, and the same corollary coupled with the Δ-Theorem implies that $\hat{\mu}_S$ is asymptotically normal. The present task is to derive the form of the covariance matrices of the limiting distributions.

Recall that Σ_1 and Σ_2 are the asymptotic covariance matrices of $\hat{\gamma} =$

$(\hat{\gamma}_1, \hat{\gamma}_X, \hat{\gamma}_Z^T, \hat{\gamma}_W^T)^T$ and $\hat{\beta}_{X|1\underline{ZW}}$ respectively. Partition

$$\Sigma_1 = \begin{pmatrix} A_{11} & A_{1X} & A_{1Z} & A_{1W} \\ A_{X1} & A_{XX} & A_{XZ} & A_{XW} \\ A_{Z1} & A_{ZX} & A_{ZZ} & A_{ZW} \\ A_{W1} & A_{WX} & A_{WZ} & A_{WW} \end{pmatrix}$$

$$= \begin{pmatrix} \bar{A}_{11} & \bar{A}_{1X} & \bar{A}_{1Z} & \bar{A}_{1W} \\ \bar{A}_{X1} & \bar{A}_{XX} & \bar{A}_{XZ} & \bar{A}_{XW} \\ \bar{A}_{Z1} & \bar{A}_{ZX} & \bar{A}_{ZZ} & \bar{A}_{ZW} \\ \bar{A}_{W1} & \bar{A}_{WX} & \bar{A}_{WZ} & \bar{A}_{WW} \end{pmatrix}^{-1}$$

where $A_{W1} = \mathrm{ACOV}(\hat{\gamma}_1, \hat{\gamma}_W)$ for example. Similarly partition

$$\Sigma_2 = \begin{pmatrix} B_{11} & B_{1Z} & 0 \\ B_{Z1} & B_{ZZ} & 0 \\ 0 & 0 & B_{WW} \end{pmatrix}$$

$$= \begin{pmatrix} \bar{B}_{11} & \bar{B}_{1Z} & 0 \\ \bar{B}_{Z1} & \bar{B}_{ZZ} & 0 \\ 0 & 0 & \bar{B}_{WW} \end{pmatrix}^{-1}.$$

The partitions are used extensively below. Suppose that $\underline{G}^T \beta_{X|1Z\underline{W}} = 1$ and that $\hat{\underline{G}}$ is a consistent estimator of $\underline{G}$ satisfying $\hat{\underline{G}}^T \hat{\beta}_{X|1Z\underline{W}} = 1$. Recall that $\gamma_W = \beta_{X|1Z\underline{W}}\lambda$. Then,

$$\begin{aligned} \hat{\underline{G}}^T\hat{\gamma}_W - \underline{G}^T\gamma_W &= (\hat{\underline{G}}^T - \underline{G}^T)\hat{\gamma}_W + \underline{G}^T(\hat{\gamma}_W - \gamma_W) \\ &= \hat{\underline{G}}^T(I - \hat{\beta}_{X|1Z\underline{W}}\underline{G}^T)\hat{\gamma}_W + \underline{G}^T(\hat{\gamma}_W - \gamma_W) \\ &= \hat{\underline{G}}^T(I - \hat{\beta}_{X|1Z\underline{W}}\underline{G}^T)\gamma_W + \underline{G}^T(\hat{\gamma}_W - \gamma_W) + o_p(n^{-\frac{1}{2}}) \\ &= -\underline{G}^T(\hat{\beta}_{X|1Z\underline{W}} - \beta_{X|1Z\underline{W}})\lambda \\ &\qquad + \underline{G}^T(\hat{\gamma}_W - \gamma_W) + o_p(n^{-\frac{1}{2}}). \end{aligned} \tag{3.25}$$

The second equality follows from $\hat{\underline{G}}^T \hat{\beta}_{X|1Z\underline{W}} = 1$. The third equality is justified by Slutsky's Theorem. The last equality follows from Slutsky's Theorem and the identities $\underline{G}^T \beta_{X|1Z\underline{W}} = 1$ and $\gamma_W = \beta_{X|1Z\underline{W}}\lambda$. With M defined as in

(3.15) and $\hat{M}$ satisfying $\hat{M} \xrightarrow{P} M$ it can be shown that all g-inverses of $\hat{\beta}_{X|1Z\underline{W}}$ consistent for $\underline{G}^T$ are of the form

$$\hat{\underline{G}}^T = (\hat{\beta}^T_{X|1Z\underline{W}} \hat{M}^{-1} \hat{\beta}_{X|1Z\underline{W}})^{-1} \hat{\beta}^T_{X|1Z\underline{W}} \hat{M}^{-1}.$$

The matrix $\hat{M}$ is an estimated weight matrix. Then (3.25) implies that the asymptotic distribution of $\hat{\underline{G}}^T \hat{\gamma}_W$ is the same for any estimator $\hat{\underline{G}}^T$ in this class. This is reminiscent of weighted least squares in linear models; in terms of the asymptotic distribution of the weighted least squares estimator, consistent estimation of weights works as well as knowing the weights.

The asymptotic covariance matrix for $\hat{\mu}_S$ is

$$\Sigma_S \equiv \begin{pmatrix} \Sigma_{S|\gamma} & 0 \\ 0 & \Sigma_{S|\beta} \end{pmatrix} =$$

$$\begin{pmatrix} A_{11} & A_{1X} & A_{1Z} & A_{1W}\underline{G} & 0 & 0 \\ A_{X1} & A_{XX} & A_{XZ} & A_{XW}\underline{G} & 0 & 0 \\ A_{Z1} & A_{ZX} & A_{ZZ} & A_{ZW}\underline{G} & 0 & 0 \\ \underline{G}^T A_{W1} & \underline{G}^T A_{WX} & \underline{G}^T A_{WZ} & \underline{G}^T C \underline{G} & 0 & 0 \\ 0 & 0 & 0 & 0 & B_{11} & B_{1Z} \\ 0 & 0 & 0 & 0 & B_{Z1} & B_{ZZ} \end{pmatrix} \tag{3.26}$$

where $C = A_{WW} + \lambda^2 B_{WW}$.

Recall $\tilde{\Sigma}_1$ is the asymptotic covariance matrix of $(\tilde{\eta}^T, \tilde{\beta}^T_{X|1ZW})^T$ where $\tilde{\eta} = (\tilde{\gamma}_1, \tilde{\gamma}_X, \tilde{\gamma}_Z^T, \tilde{\lambda})^T$. It is not hard to see that

$$\tilde{\Sigma}_1 = \begin{pmatrix} \Lambda_{\eta\eta} & \Lambda_{\eta\beta} \\ \Lambda_{\beta\eta} & \Lambda_{\beta\beta} \end{pmatrix}^{-1},$$

where

$$\Lambda_{\eta\eta} =$$

$$\begin{pmatrix} \bar{A}_{11} & \bar{A}_{1X} & \bar{A}_{1Z} & \bar{A}_{1W}\beta_{X|1Z\underline{W}} \\ \bar{A}_{X1} & \bar{A}_{XX} & \bar{A}_{XZ} & \bar{A}_{XW}\beta_{X|1Z\underline{W}} \\ \bar{A}_{Z1} & \bar{A}_{ZX} & \bar{A}_{ZZ} & \bar{A}_{ZW}\beta_{X|1Z\underline{W}} \\ \beta^T_{X|1Z\underline{W}}\bar{A}_{W1} & \beta^T_{X|1Z\underline{W}}\bar{A}_{WX} & \beta_{X|1Z\underline{W}}\bar{A}_{WZ} & \beta^T_{X|1Z\underline{W}}\bar{A}_{WW}\beta_{X|1Z\underline{W}} \end{pmatrix}, \quad (3.27)$$

$$\Lambda_{\eta\beta} = \begin{pmatrix} 0 & 0 & \bar{A}_{1W}\lambda \\ 0 & 0 & \bar{A}_{XW}\lambda \\ 0 & 0 & \bar{A}_{ZW}\lambda \\ 0 & 0 & \beta^T_{X|1Z\underline{W}}\bar{A}_{WW}\lambda \end{pmatrix}, \quad (3.28)$$

and

$$\Lambda_{\beta\beta} = \begin{pmatrix} \bar{B}_{11} & \bar{B}_{1Z} & 0 \\ \bar{B}_{Z1} & \bar{B}_{ZZ} & 0 \\ 0 & 0 & \lambda^2\bar{A}_{WW} + \bar{B}_{WW} \end{pmatrix}. \quad (3.29)$$

Denote the asymptotic covariance matrix of $\tilde{\mu}_{ML}$, see (3.24) , by Σ_{ML}. Apply result A.2 to see that

$$\Sigma^{-1}_{ML} = \begin{pmatrix} \Sigma^{-1}_{ML|\eta} & 0 \\ 0 & \Sigma^{-1}_{ML|\beta} \end{pmatrix}, \quad (3.30)$$

where

$$\Sigma^{-1}_{ML|\beta} = \begin{pmatrix} \bar{B}_{11} & \bar{B}_{1Z} \\ \bar{B}_{Z1} & \bar{B}_{ZZ} \end{pmatrix}$$

and $\Sigma^{-1}_{ML|\eta} = \Lambda_{\eta\eta} - \Lambda_{\eta\beta}\Lambda^{-1}_{\beta\beta}\Lambda_{\beta\eta}$ has components

$$\left\{\Sigma^{-1}_{ML|\eta}\right\}_{11} = \bar{A}_{11} - \lambda^2\bar{A}_{1W}D\bar{A}_{W1}$$

$$\left\{\Sigma^{-1}_{ML|\eta}\right\}_{1X} = \bar{A}_{1X} - \lambda^2\bar{A}_{1W}D\bar{A}_{WX}$$

$$\left\{\Sigma^{-1}_{ML|\eta}\right\}_{1Z} = \bar{A}_{1Z} - \lambda^2\bar{A}_{1W}D\bar{A}_{WZ}$$

$$\left\{\Sigma^{-1}_{ML|\eta}\right\}_{XW} = \bar{A}_{XW}\beta_{X|1Z\underline{W}} - \lambda^2\bar{A}_{XW}D\bar{A}_{WW}\beta_{X|1Z\underline{W}},$$

etc., where $D = (\lambda^2 \bar{A}_{WW} + \bar{B}_{WW})^{-1}$. The pattern should be clear. The asymptotic distributions of $\hat{\mu}_S$ and $\tilde{\mu}_{ML}$ are the same provided we can show

$$\Sigma_S \Sigma_{ML}^{-1} = I.$$

To do this it is crucial to understand the dependence of the asymptotic covariance matrix of $\hat{\gamma}$ on the vector of linear regression coefficients $\beta_{X|1Z\underline{W}}$. To this end, we need the following series of lemmas.

LEMMA 3.5. *Suppose* $X \sim N(\mu, \sigma^2)$. *Let* $h : \Re \mapsto \Re$ *be differentiable and such that* $E|h'(X)| < \infty$. *Then*

$$E\{(X-\mu)h(X)\} = \sigma^2 [E\{h'(X)\}]$$

PROOF: See Stein (1981, p. 1136). •

LEMMA 3.6. *Suppose* $V \sim N(\mu, \Sigma)$ *where* $\dim(V) = p$ *and let* $\gamma \in \Re^p$ *be a fixed vector. Let* $h : \Re \mapsto \Re$ *be twice differentiable and such that* $E\{|h(\gamma^T V)|\} < \infty$, $E\{|h'(\gamma^T V)(\gamma^T V - \gamma^T \mu_V) + h(\gamma^T V)|\} < \infty$ *and* $E\{|h''(\gamma^T V)|\} < \infty$. *Then*

$$E\left\{VV^T h(\gamma^T V)\right\} = k_0(\Sigma + \mu\mu^T) + k_1(\mu\gamma^T\Sigma + \Sigma\gamma\mu^T) + k_2\Sigma\gamma\gamma^T\Sigma$$

where $k_0 = E\{h(\gamma^T V)\}$, $k_1 = E\{h'(\gamma^T V)\}$, *and* $k_2 = E\{h''(\gamma^T V)\}$.

PROOF: First note that

$$\begin{aligned} E\left[VV^T h(\gamma^T V)\right] &= E\Bigg[\Big\{\mathrm{VAR}(V \mid \gamma^T V) \\ &\qquad + E(V \mid \gamma^T V)E^T(V \mid \gamma^T V)\Big\}h(\gamma^T V)\Bigg] \\ &= k_0\left\{\Sigma - \Sigma\gamma\gamma^T\Sigma(\gamma^T\Sigma\gamma)^{-1} + \mu\mu^T\right\} \\ &\quad + (\Sigma\gamma\mu^T + \mu\gamma^T\Sigma)(\gamma^T\Sigma\gamma)^{-1}E\left\{(\gamma^T V - \gamma^T\mu)h(\gamma^T V)\right\} \\ &\quad + \Sigma\gamma\gamma^T\Sigma(\gamma^T\Sigma\gamma)^{-2}E\left\{(\gamma^T V - \gamma^T\mu)(\gamma^T V - \gamma^T\mu)h(\gamma^T V)\right\}. \end{aligned}$$

Apply Lemma 3.5 to show the right hand side is

$$k_0(\Sigma + \mu\mu^T) + k_1(\Sigma\gamma\mu^T + \mu\gamma^T\Sigma) + \Sigma\gamma^T\gamma\Sigma(\gamma^T\Sigma\gamma)^{-1}$$
$$\times E\left\{(\gamma^T V - \gamma^T\mu)h'(\gamma^T V)\right\}$$
$$= k_0(\Sigma + \mu\mu^T) + k_1(\Sigma\gamma\mu^T + \mu\gamma^T\Sigma) + k_2(\Sigma\gamma^T\gamma\Sigma). \quad \bullet$$

Next we use Lemma 3.6 to examine the asymptotic covariance matrix of the probit regression estimators.

LEMMA 3.7. *Under the assumptions set forth in Theorem 3.1,*

$$ACOV(\hat{\gamma}_W, \hat{\gamma}_1) = A_{W1} = \beta_{X|1Z\underline{W}} M_1$$
$$ACOV(\hat{\gamma}_W, \hat{\gamma}_X) = A_{WX} = \beta_{X|1Z\underline{W}} M_2$$
$$ACOV(\hat{\gamma}_W, \hat{\gamma}_Z) = A_{XZ} = \beta_{X|1Z\underline{W}} M_3$$

for appropriately defined matrices M_i, $i = 1, 2, 3$.

PROOF: Define $h(x) = (\Phi'(x))^2/(\Phi(x)(1-\Phi(x)))$ and

$$\begin{pmatrix} X \\ Z \\ W \end{pmatrix} \sim N(\mu, \Sigma) \equiv N\left\{ \begin{pmatrix} \mu_X \\ \mu_Z \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{XX} & \Sigma_{XZ} & \Sigma_{XW} \\ \Sigma_{ZX} & \Sigma_{ZZ} & 0 \\ \Sigma_{WX} & 0 & \Sigma_{WW} \end{pmatrix} \right\}$$

where $\Sigma_{XX} = \Sigma_{UU} + \Sigma_{\delta\delta}$. The functions $h(x)$, $h'(x)$ and $h''(x)$ are bounded since each function is continuous and approaches zero as x approaches $\pm\infty$. Then Lemma 3.6 is applicable and used to show that

$$E(\psi_1\psi_1^T) = \begin{pmatrix} k_0 & \bar{A}_{1D} \\ \bar{A}_{D1} & \bar{A}_{DD} \end{pmatrix}$$

where

$$\bar{A}_{D1} = k_1\Sigma\underline{\gamma} + k_0\mu$$
$$\bar{A}_{DD} = k_0\Sigma + k_2\Sigma\underline{\gamma}\underline{\gamma}^T\Sigma + k_1\mu\underline{\gamma}^T\Sigma + k_1\Sigma\underline{\gamma}\mu^T + k_0\mu\mu^T \qquad (3.31)$$

and $\underline{\gamma} = (\gamma_X, \gamma_Z^T, \gamma_W^T)^T$. The constants k_0 , k_1 and k_2 are defined as the expectations of the function $h(\cdot)$ and its first and second derivatives evaluated at $\gamma_1 + \gamma_X X + \gamma_Z^T Z + \gamma_W^T W$. Results A.1 and A.2 of the appendix are applicable provided Σ is non-singular. Then

$$\{E(\psi_1 \psi_1^T)\}^{-1} = \begin{pmatrix} A_{11} & A_{1D} \\ A_{D1} & A_{DD} \end{pmatrix}$$

where

$$A_{DD} = \begin{pmatrix} A_{XX} & A_{XZ} & A_{XW} \\ A_{ZX} & A_{ZZ} & A_{ZW} \\ A_{WX} & A_{WZ} & A_{WW} \end{pmatrix}$$

$$\begin{aligned} &= (k_0 \Sigma + k_3 \Sigma \underline{\gamma} \underline{\gamma}^T \Sigma)^{-1} \\ &= k_0^{-1} \Sigma^{-1} - k_4 \underline{\gamma} \underline{\gamma}^T, \end{aligned} \tag{3.32}$$

and

$$A_{D1} = \begin{pmatrix} A_{X1} \\ A_{Z1} \\ A_{W1} \end{pmatrix} = -k_0^{-1} A_{DD} \bar{A}_{D1} \tag{3.33}$$

with $k_3 = (k_2 - k_1^2 k_0^{-1})$ and $k_4 = k_3 k_0^{-2} (1 + k_3 k_0^{-1} \underline{\gamma}^T \Sigma \underline{\gamma})^{-1}$.

Partition

$$\Sigma^{-1} = \begin{pmatrix} C_{XX} & C_{XZ} & C_{XW} \\ C_{ZX} & C_{ZZ} & C_{ZW} \\ C_{WX} & C_{WZ} & C_{WW} \end{pmatrix}$$

and again apply results for inverting partitioned matrices to see that

$$C_{WX} = -\Sigma_{WW}^{-1} \Sigma_{WX} F = -\beta_{X|1Z\underline{W}} F$$

$$C_{WZ} = \Sigma_{WW}^{-1} \Sigma_{WX} F \Sigma_{XZ} \Sigma_{ZZ}^{-1} = \beta_{X|1Z\underline{W}} F \beta_{X|1\underline{Z}W}^T$$

$$C_{WW} = \Sigma_{WW}^{-1} + \beta_{X|1Z\underline{W}} F \beta_{X|1Z\underline{W}}^T$$

where $F = (\Sigma_{XX} - \Sigma_{XZ} \Sigma_{ZZ}^{-1} \Sigma_{ZX} - \Sigma_{XW} \Sigma_{WW}^{-1} \Sigma_{WX})^{-1}$. Now,

$$\underline{\gamma} \underline{\gamma}^T = \begin{pmatrix} \gamma_X \\ \gamma_Z \\ \gamma_W \end{pmatrix} (\gamma_X, \ \gamma_Z^T, \ \gamma_W^T) = \begin{pmatrix} \gamma_X \\ \gamma_Z \\ \beta_{X|1Z\underline{W}} \lambda \end{pmatrix} (\gamma_X, \ \gamma_Z^T, \ \lambda \beta_{X|1Z\underline{W}}^T)$$

so that , using (3.32) and the results derived above,

$$\begin{aligned}
A_{WX} = k_0^{-1}C_{WX} - k_4\beta_{X|1Z\underline{W}}\lambda\gamma_X &= \beta_{X|1Z\underline{W}}(-k_0^{-1}F - k_4\lambda\gamma_X) \\
&\equiv \beta_{X|1Z\underline{W}}M_2 \\
A_{WZ} = k_0^{-1}C_{WZ} - k_4\beta_{X|1Z\underline{W}}\lambda\gamma_Z^T &= \beta_{X|1Z\underline{W}}(k_0^{-1}F\beta_{X|1\underline{Z}W}^T - k_4\lambda\gamma_Z^T) \\
&\equiv \beta_{X|1Z\underline{W}}M_3 \\
A_{WW} = k_0^{-1}C_{WW} - k_4\beta_{X|1Z\underline{W}}\lambda^2\beta_{X|1Z\underline{W}}^T &= k_0^{-1}\Sigma_{WW}^{-1} \\
&\quad - \beta_{X|1Z\underline{W}}(k_4\lambda^2 - k_0^{-1}F)\beta_{X|1Z\underline{W}} \\
&\equiv k_0^{-1}\Sigma_{WW}^{-1} - \beta_{X|1Z\underline{W}}M_4.
\end{aligned}$$

Finally, using (3.33) and (3.31)

$$\begin{aligned}
A_{W1} &= -k_0^{-1}(A_{WX}, A_{WZ}, A_{WW})(k_1\Sigma\underline{\gamma} + k_0\mu) \\
&= k_0^{-1}(A_{WX}H_1 + A_{WZ}H_2) + k_1k_0^{-1}A_{WW}(\Sigma_{WX}\gamma_X + \Sigma_{WW}\beta_{X|1Z\underline{W}}\lambda) \\
&= k_0^{-1}(A_{WX}H_1 + A_{WZ}H_2) + k_1k_0^{-1}(k_0^{-1}\beta_{X|1Z\underline{W}}(\gamma_X + \lambda)) \\
&\quad - \beta_{X|1Z\underline{W}}M_4(\Sigma_{WX}\gamma_X + \Sigma_{WW}\beta_{X|1Z\underline{W}}\lambda) \\
&\equiv \beta_{X|1Z\underline{W}}M_1
\end{aligned}$$

where

$$\begin{aligned}
H_1 &= k_1(\Sigma_{XX}\gamma_X + \Sigma_{XZ}\gamma_Z + \Sigma_{XW}\beta_{X|1Z\underline{W}}\lambda) + \mu_X k_0 \\
H_2 &= k_1(\Sigma_{ZX}\gamma_X + \Sigma_{ZZ}\gamma_Z) + \mu_Z k_0. \quad \bullet
\end{aligned}$$

We are now in a position to prove Theorem 3.1.

PROOF OF THEOREM 3.1: We need to show that

$$\Sigma_S\Sigma_{ML}^{-1} = I,$$

which follows provided $\Sigma_{S|\gamma}\Sigma^{-1}_{ML|\eta} = I$, see (3.26) and (3.30). Define the sets $\mathcal{G} = \{1, X, Z, W\}$ and $\mathcal{F} = \{1, X, Z\}$. Then, using Lemma 3.5,

$$\beta_{X|1Z\underline{W}}\underline{G}^T A_{Wi} = A_{Wi} \quad \text{for} \quad i \in \mathcal{F}.$$

Note that, by definition,

$$\sum_{i\in\mathcal{G}} A_{ki}\bar{A}_{ij} = \begin{cases} I_{kk}, & j = k \\ 0, & j \neq k. \end{cases} \tag{3.34}$$

The product of the j^{th} column of $\Sigma^{-1}_{ML|\eta}$ with the k^{th} row of $\Sigma_{S|\gamma}$, for $1 \leq j, k \leq 3$ is

$$\begin{aligned}
&\sum_{i\in\mathcal{F}} \left(A_{ki}\bar{A}_{ij}\right) + A_{kW}\underline{G}\beta^T_{X|1Z\underline{W}}\bar{A}_{Wj} \\
&\quad - \lambda^2\left(\sum_{i\in\mathcal{F}} A_{ki}\bar{A}_{iW}D\bar{A}_{Wj}\right) - \lambda^2 A_{kW}\underline{G}\beta^T_{X|1Z\underline{W}}\bar{A}_{WW}D\bar{A}_{Wj} \\
&\quad = \sum_{i\in\mathcal{G}} A_{ki}\bar{A}_{ij} - \lambda^2\left(\sum_{i\in\mathcal{G}} A_{ki}\bar{A}_{iW}\right)D\bar{A}_{Wj} \\
&\quad = \begin{cases} I_{kk}, & j = k \\ 0, & j \neq k. \end{cases}
\end{aligned}$$

The product of the fourth column of $\Sigma^{-1}_{ML|\eta}$ with the k^{th} row of $\Sigma_{S|\gamma}$, for $1 \leq k \leq 3$, is as above except that $\beta_{X|1Z\underline{W}}$ multiplies the equation from the right. So far, we have not used the actual form of the optimal g-inverse $\underline{G}^T$, but rather only that it satisfies $\underline{G}^T\beta_{X|1Z\underline{W}} = 1$. The last steps require we exploit the form of $\underline{G}^T$. Recall that

$$\underline{G}^T = \left\{\beta^T_{X|1Z\underline{W}}(A_{WW} + \lambda^2 B_{WW})^{-1}\beta_{X|1Z\underline{W}}\right\}^{-1}\beta^T_{X|1Z\underline{W}}(A_{WW} + \lambda^2 B_{WW})^{-1}.$$

When $1 \leq j \leq 3$, the product of the j^{th} column of $\Sigma^{-1}_{ML|\eta}$ with the fourth row

of $\Sigma_{S|\gamma}$ is

$$\begin{aligned}
&\underline{G}^T\Bigg\{\sum_{i\in\mathcal{F}}(A_{Wi}\bar{A}_{ij})+(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}\bar{A}_{Wj}\\
&\quad-\lambda^2\left(\sum_{i\in\mathcal{F}}A_{Wi}\bar{A}_{iW}D\bar{A}_{Wj}\right)-\lambda^2(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}\bar{A}_{WW}D\bar{A}_{Wj}\Bigg\}\\
&=\underline{G}^T\Bigg\{-A_{WW}(\lambda^2\bar{A}_{WW}+\bar{B}_{WW})\\
&\qquad+(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}(\lambda^2\bar{A}_{WW}+\bar{B}_{WW})\\
&\qquad-\lambda^2(I-A_{WW}\bar{A}_{WW})-\lambda^2(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}\bar{A}_{WW}\Bigg\}D\bar{A}_{Wj}\\
&=\underline{G}^T\Bigg\{(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}-(A_{WW}+\lambda^2B_{WW})\Bigg\}\bar{B}_{WW}D\bar{A}_{Wj}\\
&=0.
\end{aligned}$$

The last equality follows from the form of $\underline{G}^T$. Finally, the product of the fourth column of $\Sigma^{-1}_{ML|\eta}$ with the fourth row of $\Sigma_{S|\gamma}$ is just as above except j is replaced by W and the equations are multiplied on the right by $\beta_{X|1Z\underline{W}}$. The product is

$$\begin{aligned}
&\underline{G}^T\Bigg\{\sum_{i\in\mathcal{F}}(A_{Wi}\bar{A}_{iW})+(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}\bar{A}_{WW}\\
&\qquad-\lambda^2\left(\sum_{i\in\mathcal{F}}A_{Wi}\bar{A}_{iW}D\bar{A}_{WW}\right)\\
&\qquad-\lambda^2(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}\bar{A}_{WW}D\bar{A}_{WW}\Bigg\}\beta_{X|1Z\underline{W}}\\
&=1-\underline{G}^T\Bigg\{(A_{WW}+\lambda^2B_{WW})\underline{G}\beta^T_{X|1Z\underline{W}}\\
&\qquad-(A_{WW}+\lambda^2B_{WW})\Bigg\}\bar{B}_{WW}D\bar{A}_{WW}\beta_{X|1Z\underline{W}}\\
&=1.\quad\bullet
\end{aligned}$$

3.6 Centering the instrumental variables

We argue that no generality is lost assuming (3.20). Define

$$W^* = W - E(W \mid Z) = (W - \mu_W) - \Sigma_{WZ}\Sigma_{ZZ}^{-1}(Z - \mu_Z).$$

Clearly $E(W^*) = 0$ and $COV(W^*, Z) = 0$. We show the following. The maximum likelihood estimators are unaffected using W^* in place of W. The simple estimators are affected, in that their asymptotic covariance matrix is different. However, the asymptotic distributions of the simple estimators using W^* and using

$$\hat{W}^* = (W - \hat{\mu}_W) - \hat{\Sigma}_{WZ}\hat{\Sigma}_{ZZ}^{-1}(Z - \hat{\mu}_Z)$$

are the same, provided that μ_W, μ_Z, Σ_{WZ}, and Σ_{ZZ} are consistently estimated. It follows the simple estimators defined with $\hat{W}^*$ replacing W are fully efficient. The remainder of the section verifies this claim.

Denote estimators obtained by replacing W with W^* as before except superscribed with a *. First consider the maximum likelihood estimators. It is not hard to see that

$$\begin{aligned}
\tilde{\gamma}_1^* &= \tilde{\gamma}_1 + \tilde{\lambda}\tilde{\beta}_{X|1Z\underline{W}}^T(\mu_W - \Sigma_{WZ}\Sigma_{ZZ}^{-1}\mu_Z) \\
\tilde{\gamma}_Z^* &= \tilde{\gamma}_Z + \Sigma_{ZZ}^{-1}\Sigma_{ZW}\tilde{\beta}_{X|1Z\underline{W}}\tilde{\lambda} \\
\tilde{\beta}_{X|\underline{1}ZW}^* &= \tilde{\beta}_{X|\underline{1}ZW} + \tilde{\beta}_{X|1Z\underline{W}}^T(\mu_W - \Sigma_{WZ}\Sigma_{ZZ}^{-1}\mu_Z) \\
\tilde{\beta}_{X|1\underline{Z}W}^* &= \tilde{\beta}_{X|1\underline{Z}W} + \Sigma_{ZZ}^{-1}\Sigma_{ZW}\tilde{\beta}_{X|1Z\underline{W}}.
\end{aligned} \tag{3.35}$$

The remaining estimators are unchanged, that is

$$(\tilde{\lambda}^*, \tilde{\gamma}_X^*, \tilde{\beta}_{X|1Z\underline{W}}^*, \mathrm{vech}^T\tilde{\Sigma}_{XX|ZW}^*)^T = (\tilde{\lambda}, \tilde{\gamma}_X, \tilde{\beta}_{X|1Z\underline{W}}, \mathrm{vech}^T\tilde{\sigma}_{XX|ZW}^2)^T.$$

Substitute (3.35) into (3.17) with $D = \lambda$ to see the maximum likelihood estimators are unaffected when W^* replaces W in the data. Next consider the simple estimators. Similar to (3.35), we have

$$\begin{aligned}
\hat{\gamma}_1^* &= \hat{\gamma}_1 + \hat{\gamma}_W(\mu_W - \Sigma_{WZ}\Sigma_{ZZ}^{-1}\mu_Z) \\
\hat{\gamma}_Z^* &= \hat{\gamma}_Z + \Sigma_{ZZ}^{-1}\Sigma_{ZW}\hat{\gamma}_W \\
\hat{\beta}_{X|1ZW}^* &= \hat{\beta}_{X|1ZW} + \hat{\beta}_{X|1Z\underline{W}}^T(\mu_W - \Sigma_{WZ}\Sigma_{ZZ}^{-1}\mu_Z) \\
\hat{\beta}_{X|1\underline{Z}W}^* &= \hat{\beta}_{X|1\underline{Z}W} + \Sigma_{ZZ}^{-1}\Sigma_{ZW}\hat{\beta}_{X|1Z\underline{W}}. \qquad (3.36)
\end{aligned}$$

The remaining estimators are unchanged. Substitute (3.36) into (3.17) with $D = \underline{G}^T\gamma_W$. The estimator for β_1 is unchanged. The estimators for β_0 and β_2 are changed. Specifically,

$$\begin{aligned}
\hat{\beta}_0^* &= \hat{\beta}_0 + (\hat{\gamma}_W^T - \hat{\gamma}_W^T\hat{G}\hat{\beta}_{X|1Z\underline{W}}^T)(\mu_W - \Sigma_{WZ}\Sigma_{ZZ}^{-1}\mu_Z)\hat{r}_\gamma^{-1} \\
\hat{\beta}_1^* &= \hat{\beta}_1 \\
\hat{\beta}_2^* &= \hat{\beta}_2 + \Sigma_{ZZ}^{-1}\Sigma_{ZW}(\hat{\gamma}_W - \hat{\beta}_{X|1Z\underline{W}}\hat{\underline{G}}^T\hat{\gamma}_W)\hat{r}_\gamma^{-1}.
\end{aligned}$$

Since $(\hat{\gamma}_W - \hat{\beta}_{X|1Z\underline{W}}\hat{\underline{G}}^T\hat{\gamma}_W) \xrightarrow{P} 0$ the additional terms are consistent for zero, so that $(\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^{*T})^T$ is consistent for $(\beta_0, \beta_1, \beta_2^T)^T$. A simple argument using Slutsky's Theorem will show that replacing μ_W, μ_Z, Σ_{WZ}, and Σ_{ZZ} by, for example, the corresponding sample moments results in estimators with the same asymptotic distribution as $(\hat{\beta}_0^*, \hat{\beta}_1^*, \hat{\beta}_2^{*T})^T$.

3.7 Expressions for the asymptotic variance of the simple estimators

The next task is to derive the form of the asymptotic distribution of the optimal simple estimators $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2^T)^T$. Asymptotic normality follows from

the asymptotic normality of $\hat{\mu}_S$ and a straightforward application of the Δ-Theorem. Our primary interest here is obtaining an expression for the asymptotic covariance matrix for $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2^T)^T$.

Let $\hat{\mu} = (\hat{\gamma}^t, \hat{\beta}_{X|1ZW}, \hat{\sigma}^2_{XX|ZW})^T$ denote the probit and linear regression estimators. The task is to obtain matrices V_i, $i = 0, 1, 2$ such that

$$\begin{pmatrix} \hat{\beta}_0 - \beta_0 \\ \hat{\beta}_1 - \beta_1 \\ \hat{\beta}_2 - \beta_2 \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \\ V_3 \end{pmatrix} (\hat{\mu} - \mu) + o_p(n^{-\frac{1}{2}}).$$

This is the Δ-Theorem. Then it follows $\text{ACOV}(\hat{\beta}_i, \hat{\beta}_j) = V_i^T \Sigma V_j$ where

$$\Sigma = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & \Sigma_3 \end{pmatrix},$$

and, as defined previously,

$$\begin{aligned} \Sigma_1 &= \text{ACOV}(\hat{\gamma}, \hat{\gamma}), \\ \Sigma_2 &= \text{ACOV}(\hat{\beta}_{X|1ZW}, \hat{\beta}_{X|1ZW}), \\ \Sigma_3 &= \text{ACOV}(\hat{\sigma}^2_{XX|ZW}, \hat{\sigma}^2_{XX|ZW}). \end{aligned}$$

Since each β_i contains the elements r_γ^{-1} and $\underline{G}^T \gamma_W$ we first derive expressions for them. In fact, we have already seen that

$$\hat{\underline{G}}^T \hat{\gamma}_W - \underline{G}^T \gamma_W = \underline{G}^T (\hat{\gamma}_W - \gamma_W) - \underline{G}^T (\hat{\beta}_{X|1Z\underline{W}} - \beta_{X|1Z\underline{W}}) \lambda + o_p(n^{-\frac{1}{2}}),$$

see (3.25). Next, note the Δ-Theorem is applicable to the function $g(x) = (1-x)^{-\frac{1}{2}}$ provided $x \neq 1$. Then

$$\hat{r}_\gamma^{-1} - r_\gamma^{-1} = \frac{1}{2} r_\gamma^{-3} \left\{ \hat{\gamma}_X^T \hat{\sigma}^2_{XX|ZW} \hat{\underline{G}}^T \hat{\gamma}_W - \gamma_X^T \sigma^2_{XX|ZW} \underline{G}^T \gamma_W \right\} + o_p(n^{-\frac{1}{2}}).$$

Apply the Δ-Theorem again to the function $g(a,b,c) = abc$ to see that

$$\hat{r}_\gamma^{-1} - r_\gamma^{-1} = \frac{1}{2} r_\gamma^{-3} \Big\{ \gamma_W^T G \sigma^2_{XX|ZW} (\hat{\gamma}_X - \gamma_X) + \gamma_X^T (\sigma^2_{XX|ZW} - \sigma^2_{XX|ZW}) \underline{G}^T \gamma_W + \gamma_X \sigma^2_{XX|ZW} (\hat{\underline{G}}^T \hat{\gamma}_W - \underline{G}^T \gamma_W) \Big\} + o_p(n^{-\frac{1}{2}}). \tag{3.37}$$

We can now easily obtain, for example, V_1.

$$\begin{aligned} \hat{\beta}_1 - \beta_1 &= \hat{r}_\gamma^{-1} (\hat{\gamma}_X + \hat{\underline{G}}^T \hat{\gamma}_W) - r_\gamma^{-1} (\gamma_X + \underline{G}^T \gamma_W) \\ &= (\hat{r}_\gamma^{-1} - r_\gamma^{-1})(\gamma_X + \underline{G}^T \gamma_W) + r_\gamma^{-1} (\hat{\gamma}_X - \gamma_X) \\ &\quad + r_\gamma^{-1} (\hat{\underline{G}}^T \hat{\gamma}_W - \underline{G}^T \gamma_W) + o_p(n^{-\frac{1}{2}}). \end{aligned}$$

Using (3.25) and (3.37) ,

$$\begin{aligned} \hat{\beta}_1 - \beta_1 = V_{1gx} (\hat{\gamma}_X - \gamma_X) + V_{1gw}^T (\hat{\gamma}_W - \gamma_W) + V_{1bw}^T (\hat{\beta}_{X|1Z\underline{W}} - \beta_{X|1Z\underline{W}}) \\ + V_{1s} (\hat{\sigma}^2_{XX|ZW} - \sigma^2_{XX|ZW}) + o_p(n^{-\frac{1}{2}}), \end{aligned}$$

where

$$\begin{aligned} V_{1gx} &= (\gamma_X + \underline{G}^T \gamma_W) \frac{1}{2} r_\gamma^{-3} \gamma_W^T \underline{G} \sigma^2_{XX|ZW} + r_\gamma^{-1} \\ V_{1gw} &= (\gamma_X + \underline{G}^T \gamma_W) \frac{1}{2} r_\gamma^{-3} \gamma_X \sigma^2_{XX|ZW} \underline{G}^T + r_\gamma^{-1} \underline{G}^T \\ V_{1bw} &= -\lambda V_{1gw} \\ V_{1s} &= (\gamma_X + \underline{G}^T \gamma_W) \frac{1}{2} r_\gamma^{-3} \gamma_X \underline{G}^T \gamma_W. \end{aligned}$$

Then

$$V_1 = (0, V_{1gx}, 0^T, V_{1gw}^T, 0, V_{1bw}^T, V_{1s})^T.$$

The vectors V_0 and V_2 are constructed similarly. In particular, letting

$$
\begin{aligned}
V_{0g1} &= r_\gamma^{-1} \\
V_{0gx} &= (\gamma_1 - \beta_{X|\underline{1}ZW}\underline{G}^T\gamma_W)\frac{1}{2}r_\gamma^{-3}\gamma_W^T\underline{G}\sigma^2_{XX|ZW} \\
V_{0gw} &= (\gamma_1 - \beta_{X|\underline{1}ZW}\underline{G}^T\gamma_W)\frac{1}{2}r_\gamma^{-3}\gamma_X\sigma^2_{XX|ZW}\underline{G}^T - r_\gamma^{-1}\beta_{X|\underline{1}ZW}\underline{G}^T \\
V_{0b1} &= -r_\gamma^{-1}\underline{G}^T\gamma_W \\
V_{0bw} &= -\lambda V_{0gw} \\
V_{0s} &= (\gamma_1 - \beta_{X|\underline{1}ZW}\underline{G}^T\gamma_W)\frac{1}{2}r_\gamma^{-3}\gamma_X\underline{G}^T\gamma_W
\end{aligned}
$$

and

$$
\begin{aligned}
V_{2gx} &= (\gamma_Z - \beta_{X|1\underline{Z}W}\underline{G}^T\gamma_W)\frac{1}{2}r_\gamma^{-3}\gamma_W^T\underline{G}\sigma^2_{XX|ZW} \\
V_{2gz} &= r_\gamma^{-1} \\
V_{2gw} &= (\gamma_Z - \beta_{X|1\underline{Z}W}\underline{G}^T\gamma_W)\frac{1}{2}r_\gamma^{-3}\gamma_X\sigma^2_{XX|ZW}\underline{G}^T - r_\gamma^{-1}\beta_{X|1\underline{Z}W}\underline{G}^T \\
V_{2bz} &= -r_\gamma^{-1}\underline{G}^T\gamma_W \\
V_{2bw} &= -\lambda V_{2gw} \\
V_{2s} &= (\gamma_Z - \beta_{X|1\underline{Z}W}\underline{G}^T\gamma_W)\frac{1}{2}r_\gamma^{-3}\gamma_X\underline{G}^T\gamma_W
\end{aligned}
$$

we have

$$
\begin{aligned}
V_0 &= (V_{0g1}, V_{0gx}, 0^T, V_{0gw}^T, V_{0b1}, 0^T, V_{0bw}^T, V_{0s})^T \\
V_2 &= (0, V_{2gx}, V_{2gz}^T, V_{2gw}^T, 0, V_{2bz}^T, V_{2bw}^T, V_{2s})^T.
\end{aligned}
$$

3.8 A simulation study

In this section the results of a small simulation study are described. The model specification is a rough match to a subset of data from the Framingham

Heart Study, see Gordon and Kannel (1968). In the Framingham study, the binary response Y indicates the presence or absence of coronary heart disease. Measured covariates include blood pressure and age. The specification for the covariates used in the simulation study is

$$\begin{pmatrix} X \\ U \\ Z \\ W_1 \\ W_2 \end{pmatrix} \sim N \left\{ \begin{pmatrix} 133 \\ 133 \\ 46 \\ 120 \\ 120 \end{pmatrix}, \begin{pmatrix} 302+121 & 302 & 42 & 319 & 303 \\ 302 & 302 & 42 & 319 & 303 \\ 42 & 42 & 73 & 40 & 40 \\ 319 & 319 & 40 & 440 & 368 \\ 303 & 303 & 40 & 368 & 415 \end{pmatrix} \right\}.$$

The specification is such that one covariate is measured with error, one without error and two instruments are available. The covariate X is measured systolic blood pressure at the start of the study and Z is age. The instruments W_1 and W_2 are blood pressure measurements taken two and four years after the start of the study. The analytic expressions for $E(Y)$, CORR(Y,U) and CORR(Y,Z) depend on the parameters $(\beta_0,\beta_1,\beta_2)$. The notation CORR$(\cdot,\cdot)$ denotes the correlation of the arguments. Values for $(\beta_0,\beta_1,\beta_2)$ were obtained by specifying values for $E(Y)$,CORR(Y,U) and CORR(Y,Z) and then solving the analytic expressions for $(\beta_0,\beta_1,\beta_2)$. Here β_0 is the intercept, β_1 the coefficient of U and β_2 the coefficient of Z. In particular, we set $E(Y) = .1$, CORR$(Y,Z) = .15$ and varied CORR(Y,U) at .15, .25 and .35. The value CORR$(Y,U) = .15$ corresponds roughly to a subset of the Framingham data. The sample size for the simulations was set at $n = 1500$ and the number of data sets generated for each of the three specifications was 1,000. Four different estimators were computed: an estimator that ignores measurement error, termed the naive estimator; a simple estimator using the generalized inverse $\hat{G}^T = (\hat{\beta}^T_{X|1Z\underline{W}}\hat{\beta}_{X|1Z\underline{W}})^{-1}\hat{\beta}^T_{X|1Z\underline{W}}$; the optimal simple estimator; and the maximum likelihood estimator.

Results are presented in Tables 3.1-3.5 and are as expected, based on the theory. The optimal simple estimators and the maximum likelihood estimators behaved similarly in terms of BIAS and MSE, see Table 3.1. In Table 3.1, a $\diamond$ between two values indicates the difference was significant at the 5% level but both individual values were not significantly different form zero at the 5% level. Thus, a $\diamond$ indicates statistically significant differences that are probably not practically significant. A * between two values indicates the difference was significant at the 5% level and at least one individual value was significantly different from zero. Thus, a * indicates statistically significant differences that may be practically significant.

The estimators that correct for measurement error performed substantially better than the naive estimator in all cases. For example, as CORR(Y, U) increased from .15 to .35, the relative efficiency between the naive estimator and the optimal simple estimator, defined as $\frac{\text{MSE}(OS)}{\text{MSE}(N)}$, decreased from 50% to 8%.

Presented in Table 3.2 are the results of computing 95% Wald-type confidence intervals. The intervals were computed using the expressions derived in Section 3.7. The proportion of intervals covering was close to the nominal level further indicating the asymptotic theory is relevant at the sample size studied.

To determine whether differences in MSE were practically significant, relative efficiencies between the simple and optimal simple estimators and the maximum likelihood estimator and optimal simple estimators were computed. The results are presented in Tables 3.3 and 3.4 respectively. Based on the entries in Table 3.3, the statistically significant differences between the MSE's of the simple and optimal simple estimators observed in Table 3.1 may be practically significant.

Table 3.4 indicates differences in MSE between the maximum likelihood and optimal simple estimators are not practically significant.

Table 3.5 presents the correlation matrix for the four estimators of β_1 when $\text{CORR}(Y, U) = .25$.

The consistency and efficiency results of the previous sections relied on the assumption of normally distributed covariates, instruments and errors. The estimators will not be consistent when this assumption is violated. A simulation study was done to assess the robustness of the optimal simple estimator. Results are presented in Table 3.6. The distribution of the covariates (U, Z, W) was based on Chi-squared random variables with four degrees of freedom, shifted and scaled to have the same first two moments as above. The measurement error was normally distributed as before. The number of simulations was 1,000.

The naive and optimal simple estimators were compared in terms of bias, mean squared error, and percent coverage of 95% Wald-type confidence intervals. The optimal simple estimator performed much better than the naive estimator in all cases. For example, the relative efficiency of the naive estimator to the optimal simple estimator for β_1 went from 59% to 22% as $\text{CORR}(Y, U)$ increased from .15 to .35.

As expected, comparison of Tables 3.1 and 3.6 indicates the optimal simple estimator computed with Chi-squared distributed covariates has more bias than when the covariates are normally distributed. However, the optimal simple estimator did not lose much in terms of mean squared error and in some cases did better. These results are encouraging, since in practice covariates will not be exactly normally distributed.

3.9 Summary

An analysis of a probit regression model where normally distributed covariates are convolved with normally distributed measurement error was presented. When normally distributed instrumental variables are available, the regression parameters of the probit model were shown to be identified, provided the number of instruments is at least as great as the number of covariates measured with error.

A class of computationally simple estimators of the regression parameters was defined. When the number of instruments is equal to the number of covariates measured with error, the only estimator in the class is the maximum likelihood estimator. When more instruments are available than covariates measured with error, the class of simple estimators is indexed by a class of generalized inverses. A generalized inverse was identified that is optimal in the sense that the associated simple estimator has an asymptotic distribution identical to the maximum likelihood estimator.

A simulation study supported the theory and indicated the optimal simple estimator may be robust to normality assumptions.

Table 3.1. Comparison of BIAS and MSE. Asterisks or diamonds between two values indicates the difference was significant at the 5% level, see the discussion in Section 3.8 for details.

Parameter		Naive		Simple		Opt. Sim.		MLE
				CORR$(Y,U) = .15$				
$\beta_0 = -4.078$	BIAS x10	3.91	*	-0.24		-0.23	◇	-0.21
	MSE	0.29	*	0.21	*	0.19		0.19
$\beta_1 = 0.012$	BIAS x10^3	-3.58	*	.012		0.12	◇	0.13
	MSE x10^5	1.78	*	1.10	*	0.89	*	0.89
$\beta_2 = 0.025$	BIAS x10^5	193.39	*	1.77		1.99	◇	-0.99
	MSE x10^5	3.46	*	3.27	*	3.21		3.21
				CORR$(Y,U) = .25$				
$\beta_0 = -5.595$	BIAS x10	9.05	*	-0.21		-0.30	*	-0.29
	MSE	0.97	*	0.26	*	0.22	*	0.22
$\beta_1 = 0.025$	BIAS x10^3	-7.85	*	0.12		0.20	*	0.22
	MSE x10^5	6.71	*	1.31	*	1.02		1.02
$\beta_2 = 0.019$	BIAS x10^5	370.77	*	-2.48		-7.27	◇	-11.12
	MSE x10^5	4.75	*	3.69		3.63		3.63
				CORR$(Y,U) = .35$				
$\beta_0 = -7.695$	BIAS x10	1.71	*	-0.22	◇	-0.37	◇	-0.35
	MSE	3.14	*	0.42	*	0.37	*	0.37
$\beta_1 = 0.041$	BIAS x10^3	-14.10	*	0.13	*	0.26	*	0.28
	MSE x10^5	20.63	*	2.10	*	1.72		1.73
$\beta_2 = 0.014$	BIAS x10^4	56.74	*	-.84	◇	-1.57	◇	-2.07
	MSE x10^5	6.99	*	4.60	*	4.49	*	4.47

Table 3.2 Percent coverage of 95% Wald-type confidence intervals for the optimal simple estimator.

	CORR(Y,U)		
Parameter	.15	.25	.35
β_0	95.8	96.1	95.4
β_1	96.4	96.5	94.7
β_2	94.9	94.6	93.3

Table 3.3 Relative efficiency of the simple estimator to the optimal simple estimator. Table entries are $\frac{\text{MSE(OS)}}{\text{MSE(S)}}$.

	CORR(Y,U)		
Parameter	.15	.25	.35
β_0	.881	.860	.896
β_1	.804	.782	.822
β_2	.980	.985	.975

Table 3.4 Relative efficiency of the maximum likelihood estimator to the optimal simple estimator. Table entries are $\frac{\text{MSE(OS)}}{\text{MSE(MLE)}}$.

	CORR(Y,U)		
Parameter	.15	.25	.35
β_0	1.001	1.002	1.003
β_1	.997	.998	.998
β_2	1.001	1.001	1.003

Table 3.5 Correlations between estimates for β_1 when $\text{CORR}(Y, U) = .25$.

	Naive	Simple	Opt. Sim.	MLE
Naive	1	.662	.711	.709
Simple	.	1	.906	.906
Opt. Sim.	.	.	1	1*
MLE	.	.	.	1

* To five decimal places, correlation=.99979.

Table 3.6 Robustness of the optimal simple estimator. An asterisk between two values indicates the difference was significant at the 5% level, see the discussion in Section 3.8 for details. Percent coverage refers to the percentage of 95% Wald type confidence intervals that contained the true value.

Parameter		Naive		Opt. Simple
		CORR(Y,U) = .15		
$\beta_0 = -4.078$	BIAS x10	3.29	*	-0.52
	MSE	0.23	*	0.17
	% Coverage	81.0	*	93.8
$\beta_1 = 0.012$	BIAS x10^3	-3.05	*	0 .39
	MSE x10^5	1.40	*	0.83
	% Coverage	68.6	*	95.1
$\beta_2 = 0.025$	BIAS x10^5	171.71	*	11.54
	MSE x10^5	2.81	*	2.68
	% Coverage	92.9	*	94.4
		CORR(Y,U) = .25		
$\beta_0 = -5.595$	BIAS x10	6.64	*	-1.81
	MSE	0.57	*	0.23
	% Coverage	54.0	*	94.2
$\beta_1 = 0.025$	BIAS x10^3	-6.10	*	1.36
	MSE x10^5	4.16	*	1.10
	% Coverage	20.3	*	93.7
$\beta_2 = 0.019$	BIAS x10^5	347.34	*	-30.80
	MSE x10^5	3.77	*	2.87
	% Coverage	90.6	*	95.2

Table 3.6 (continued) Robustness of the optimal simple estimator.

Parameter		Naive		Opt. Simple
		CORR$(Y, U) = .35$		
$\beta_0 = -7.695$	BIAS x10	11.41	*	-4.34
	MSE	1.49	*	0.52
	% Coverage	24.8	*	90.0
$\beta_1 = 0.041$	BIAS x10^3	-10.14	*	3.06
	MSE x10^5	10.95	*	2.37
	% Coverage	3.9	*	89.4
$\beta_2 = 0.014$	BIAS x10^4	55.06	*	-4.23
	MSE x10^5	6.04	*	3.81
	% Coverage	82.8	*	95.4

CHAPTER 4

Optimal Estimating Functions for Generalized Linear Measurement Error Models with Instrumental Variables

4.1 Introduction

This chapter concerns adjusting for measurement error in generalized linear models in canonical form when instrumental variables are available. Optimal estimating functions for the functional measurement error model and optimal estimating functions for the structural model with the distribution of the true covariates unspecified are identified. The theory is a generalization of that in Stefanski and Carroll (1987). They considered efficient estimation in generalized linear measurement error models when the ratio of the response variance to the measurement error variance was known. Here the additional information allowing identification of model parameters is in the form of instrumental variables.

The model studied has the following form. Given a covariate p-vector $U = u$ and covariate q-vector $Z = z$, Y has density

$$f_{Y|UZ} = \exp\left\{\frac{y(\beta_0 + u^T\beta_1 + z^T\beta_2) - b(\beta_0 + u^T\beta_1 + z^T\beta_2)}{a(\phi)} + c(y,\phi)\right\} \quad (4.1)$$

with respect to a σ-finite measure ν. In the examples considered the dominating measure is either Lebesgue or counting measure. Compare to (4.1) to equation (1.1) in Chapter 1. The right hand side of (1.1) depends on a function $g(\cdot)$. Here $g(t) = t$, which is the canonical representation of the generalized linear model (1.1). The canonical form is necessary for application of the theory described below. Suppose U cannot be observed directly, but given $U = u$ one observes

X according to

$$f_{X|U} = (2\pi)^{-\frac{p}{2}} |\Sigma_{\delta\delta}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-u)^T \Sigma_{\delta\delta}^{-1}(x-u)\right\} \tag{4.2}$$

where $\Sigma_{\delta\delta} = \text{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. We also observe instrumental variables W such that, given $U = u$ and $Z = z$, the k-vector W has density

$$\begin{aligned} f_{W|UZ} = (2\pi)^{-\frac{k}{2}} |\Sigma_{WW|UZ}|^{-\frac{1}{2}} \\ \times \exp\Big\{ -\frac{1}{2}(w - \xi - \eta^T u - \gamma^T z)^T \\ \times \Sigma_{WW|UZ}^{-1}(w - \xi - \eta^T u - \gamma^T z)\Big\}. \end{aligned} \tag{4.3}$$

The observable data consist of n independent observations (Y, X, W) and the associated covariates z. If the covariates $\{u_i, z_i\}_{i=1}^n$ are considered fixed, a functional model is obtained. When $\{u_i, z_i\}_{i=1}^n$ are considered realizations from independent identically distributed random variables $\{U_i, Z_i\}_{i=1}^n$, a structural model is obtained.

This chapter has the following organization. Section 4.2 briefly discusses why ordinary maximum likelihood is not a viable method in the functional model and outlines an approach to estimation based on estimating functions. An optimal estimating function is defined, and for a particular class of models, the optimal estimating function is identified. In Section 4.3 the optimal estimating function for the functional measurement error model above is derived. Section 4.4 presents the theory for the structural model. An analysis of the optimal estimating function for linear and logistic regression is considered in Section 4.5. A simulation study for the logistic model is presented in Section 4.6 and a summary of the results is given in Section 4.7.

4.2 The optimal estimating function

In statistical estimation problems classical maximum likelihood theory does not generally apply when the dimension of the parameter vector is on the order of the sample size. Neymann and Scott (1948) were the first to illustrate this. They demonstrated maximum likelihood estimators that were inconsistent or inefficient. In the functional linear measurement error model, knowledge of the ratio of response variance to measurement error variance leads to consistent estimators for the regression parameters but inconsistent estimators for the variances. In logistic regression, knowledge of the measurement error variance identifies model parameters but maximum likelihood fails to produce consistent estimators, see Stefanski and Carroll (1985).

Indeed, maximum likelihood is not an option for the model given by (4.1)-(4.3). Take $u_i = X_i$ to see the likelihood is unbounded as $\text{diag}(\Sigma_{\delta\delta}) \to 0$. In fact, in the normal theory linear model, the solution to the derivative of the log-likelihood yields a saddlepoint, see Solari (1969). An alternative to maximum likelihood is needed.

In our effort to overcome the effects of nuisance parameters we will work with estimating functions rather than directly with estimators themselves. To motivate the ideas we consider a generalization of the model given by (4.1)-(4.3). Suppose the data consist of n independent observations $Y_1, \ldots, Y_n$ such that Y_i , possibly vector valued, has density

$$f_{Y_i}(y_i; \theta, u_i) = h(y_i; \theta) g(S(y_i; \theta); \theta, u_i) \tag{4.4}$$

where $\theta \in \Theta \subseteq \Re^d$ and $u_i \in \Re^p$. Interest lies in estimating the *structural* parameters θ. The $\underline{u} = (u_1^T, \ldots, u_n^T)^T$ are *nuisance* parameters.

Fixing $\underline{u} = \underline{\tilde{u}}$, an estimating function $\psi_+(\underline{y}, \theta, \underline{\tilde{u}})$ is a function of the data $\underline{Y} = (Y_1^T, \ldots, Y_n^T)^T$ and the structural parameters θ. An estimator $\hat{\theta}$ for θ solves $\psi_+(\underline{Y}, \hat{\theta}, \underline{\tilde{u}}) = 0$. Such estimators are often termed M- estimators. Define $V = \frac{\partial}{\partial \theta} \log f_Y$ and $\dot{l} = \sum_{i=1}^n V_i$. Attention is restricted to estimating functions ψ_+ that satisfy the following conditions.

i) $\psi_+ = \sum_{i=1}^n \psi(Y_i, \theta, \tilde{u}_i)$.

ii) $E_{\theta,u}\psi(Y, \theta, \tilde{u}) = 0$ for any $\theta \in \Theta$, u and $\tilde{u} \in \Re^p$.

iii) $E_{\theta,\underline{u}} \frac{\partial}{\partial \theta} \psi_+(\underline{Y}, \theta, \underline{\tilde{u}}) = -E_{\theta,\underline{u}}\{\psi_+(\underline{Y}, \theta, \underline{\tilde{u}}) \dot{l}^T(\underline{Y}, \theta, \underline{u})\}$.

iv) $E\|\psi(Y, \theta, \tilde{u})\| < \infty$.

Denote the class of functions satisfying i)-iv) by Ψ. Restriction i) seems natural since the data are assumed independent. Estimating functions that satisfy ii) are termed unbiased. As noted in Carroll and Ruppert (1988, p. 210), this condition is almost necessary to establish consistency. The idea is that, while we may not be able to consistently estimate $(\theta, \underline{u})$, by specifying a fixed sequence $\underline{\tilde{u}}$, we can consistently estimate θ. Note that in general $E_{\theta,\underline{u}} \dot{l}(\underline{Y}, \theta, \underline{\tilde{u}}) \neq 0$ when $\underline{u} \neq \underline{\tilde{u}}$ so that generally $\dot{l} \notin \Psi$. Restriction iii) is fairly standard and follows provided differentiation and expectation can be interchanged.

When ii) is satisfied, the variance of the estimating function is $E\psi_+\psi_+^T$. Godambe (1976) defines the sensitivity as $E\frac{\partial}{\partial \theta}\psi_+$. Assuming that the sensitivity is invertible, it is convenient to study standardized estimating functions $\psi_s = (E\frac{\partial}{\partial \theta}\psi_+)^{-1}\psi_+$. The following defines an optimal estimating function, see Godambe (1976).

DEFINITION 4.1: An estimating function ψ_+^* is said to be optimum if

$$E\psi_s\psi_s^T - E\psi_s^*\psi_s^{*T} \geq_{p.d.} 0$$

for all $\psi_+ \in \Psi$.

What relevance do properties of an estimating function have to the estimator itself? Theorem 3.5 suggests the answer. Under regularity conditions, an estimator $\hat{\theta}$ solving $\psi_+ = 0$ has the property

$$(\hat{\theta} - \theta) \quad \text{is} \quad AN\left\{0, \frac{1}{n}(E\psi_s\psi_s^T)^{-1}\right\}.$$

Thus, optimal estimating functions can yield consistent estimators achieving a variance lower bound.

The present task is to identify the optimal estimating function in the class Ψ for the model (4.4). Note that, considering θ known, $S(y;\theta)$ is sufficient for u_i by the Factorization Theorem. The idea is to construct a conditional likelihood from the distributions of Y_i given S_i $(1 \leq i \leq n)$. If S is complete and does not depend on θ, then the conditional likelihood yields an optimal estimating function for θ free of the nuisance parameters $\underline{u}$. When S depends on θ, optimal estimating functions for θ can again be derived from the conditional likelihood. However, the optimal estimating function will generally depend on the sequence $\underline{u}$ generating the data, see Lindsay (1982).

The following theorem identifies the optimal estimating function in the class Ψ for the model (4.4). For a more general result, see Lindsay (1982).

THEOREM 4.1. *Let $Y_1, \ldots, Y_n$ be independent each with density*

$$f_Y = h(y;\theta)g(S(y;\theta);\theta,u)$$

with respect to σ-finite measure ν. Suppose that, when θ is considered known, $S(Y_i;\theta)$ is complete and sufficient for u_i. Let $\psi_i^* = V_i - E(V_i|S_i)$. Note that V_i,

and hence ψ_i^*, depends on the null u_i generating Y_i. Define

$$\psi_+^* = \sum_{i=1}^{n} \psi_i^*.$$

Then

$$E\psi_s\psi_s^T - E\psi_s^*\psi_s^{*T} \geq_{p.d.} 0$$

for $\psi_+ \in \Psi$ with equality iff $\psi_s = \psi_s^*$ almost everywhere ν.

PROOF: First note that with θ considered fixed, the sufficiency of S for u implies ψ^* is unbiased. Next, for $\psi_+ \in \Psi$, completeness of S together with property ii) above imply $E(\psi|S) = 0$. Conditioning first on S, it then follows that $E_{\theta,u}\psi\psi^{*T} = E_{\theta,u}\psi V^T$ so that

$$E_{\theta,\underline{u}}\psi_+\psi_+^{*T} = E_{\theta,\underline{u}}\psi_+ \dot{l}^T. \tag{4.5}$$

Now, $\dot{l}$ is evaluated at the sequence $\underline{u}$ generating $Y_1, \ldots, Y_n$ so that condition iii) and (4.5) imply

$$\begin{aligned} E\psi_s^*\psi_s^{*T} &= -\left(E\frac{\partial}{\partial\theta}\psi_+^*\right)^{-1} \\ E\psi_s\psi_s^{*T} &= -\left(E\frac{\partial}{\partial\theta}\psi_+^*\right)^{-1}. \end{aligned} \tag{4.6}$$

Note that if $\dot{l}$ was evaluated at $\tilde{\underline{u}} \neq \underline{u}$, then (4.5) still holds but cannot be combined with condition iii) to get (4.6). Using (4.6) we have

$$\begin{aligned} E(\psi_s - \psi_s^*)(\psi_s - \psi_s^*)^T &= E\psi_s\psi_s^T - E\psi_s\psi_s^{*T} - E\psi_s^*\psi_s^T + E\psi_s^*\psi_s^{*T} \\ &= E\psi_s\psi_s^T - E\psi_s^*\psi_s^{*T}. \end{aligned}$$

The theorem is proved since the matrix on the left hand side is non-negative definite and equal to zero iff $\psi_s^* = \psi_s$ almost everywhere ν. •

Remark: The value of the theorem seems questionable. The optimal estimating function requires knowledge of the sequence $\underline{u}$ generating $\underline{Y}$. However if $\underline{u}$ were known, $\dot{l}(\theta;\underline{u})$, not ψ_+^*, would be the optimal estimating function. Of course it may be that $\dot{l}(\theta;\underline{u}) \notin \Psi$, but if $\underline{u}$ were known the restriction $E_{\theta,u}\psi(Y,\theta,\tilde{u}) = 0$ for $u,\tilde{u} \in \Re^p$ should be weakened to $E_{\theta,u}\psi(Y,\theta,u) = 0$ for $u \in \Re^p$ so that Ψ is irrelevant. On the other hand, in practice $\underline{u}$ is unknown and the theorem tells us we cannot find the globally optimum estimating function. This problem is addressed in the next two sections.

4.3 The optimal estimating function for the model (4.1)-(4.3)

The objective of this section is to derive ψ_+^* for the functional model given by (4.1) -(4.3). The specific cases of linear and logistic regression will be analyzed in Section 4.5.

To derive the optimal estimating function note

$$S = Y\Omega\beta_1 + X + \Sigma_{\delta\delta}\eta\Sigma_{WW}^{-1}W$$

where $\Omega = \{a(\phi)\}^{-1}\Sigma_{\delta\delta}$, is complete and sufficient for u when

$$\theta = (\beta_0, \beta_1^T, \beta_2^T, \phi, \xi^T, \mathrm{vec}^T\eta^T, \mathrm{vec}^T\gamma, \mathrm{vech}^T\Sigma_{WW}, \mathrm{diag}^T\Sigma_{\delta\delta})^T \tag{4.7}$$

is known. Sufficiency and completeness follow from the fact that, with θ fixed, the density corresponding to (4.1) -(4.3) belongs to the exponential family with parameter space an open rectangle in $\Re^p$. From Theorem 4.1 the optimal estimating function ψ_+^* corresponding to θ is constructed from the components

$$Y - E(Y \mid S)$$

$$\{Y - E(Y \mid S)\}\, u$$

$$\{Y - E(Y \mid S)\}\, z$$

$$\frac{\partial}{\partial\phi}c(Y,\phi) - E\left\{\frac{\partial}{\partial\phi}c(Y,\phi) \mid S\right\}$$
$$W - E(W \mid S)$$
$$\text{vec}\left[\{W - E(W \mid S)\}\, u^T\right]$$
$$\text{vec}\left[\{W - E(W \mid S)\}\, z^T\right]$$
$$\text{vech}\left\{WW^T - E(WW^T \mid S)\right\}$$
$$\text{diag}\left\{XX^T - E(XX^T \mid S)\right\}. \tag{4.8}$$

The optimal estimating function depends on the unobservable $\{u_i\}_{i=1}^n$ as a sequence of weights. Then at best, we can only hope to estimate the optimal estimating function. It is not hard to see that the optimal estimating function remains unbiased when u is replaced by a function of S. Lindsay (1982) identifies the optimal estimating function in a class of estimated estimating functions where u_i is replaced by a function of S_i. For the measurement error models considered by Stefanski and Carroll (1987), Lindsay's optimal estimated estimating function is quite complex and thus less than fully acceptable. The models here are even more complex. Reasonable alternatives are to simply replace u by S or $E(X \mid S)$. Note that

$$E(S) = E(Y)\Omega\beta_1 + u + \Sigma_{\delta\delta}\eta\Sigma_{WW}^{-1}(\xi + \eta^T u)$$

is biased for u, but if $|\Sigma_{\delta\delta}|$ is small, the bias is small. Since X is unbiased for u and S is sufficient for u, $E(X \mid S)$ is a uniformly minimum variance unbiased estimator of u.

4.4 Nonparametric structural case

In the preceding sections we considered estimating θ in the presence of nuisance parameters $\underline{u} = (u_1^T, \ldots, u_n^T)^T$. In the structural model the covariates are random so the the density of an observation (Y, X, W) given $Z = z$ is

$$f_{YXW|Z} = \int f_{Y|UZ}(u) f_{X|U}(u) f_{W|UZ}(u) f_{U|Z}(u)\, d\lambda(u) \tag{4.9}$$

where $f_{Y|UZ}$, $f_{X|U}$ and $f_{W|UZ}$ are defined by (4.1)- (4.3). The density of U given $Z = z$ is $f_{U|Z}$ and λ represents Lebesgue measure on $\Re^p$. The log-likelihood of n independent observations is

$$l = \sum_{i=1}^{n} \log f_{Y_i X_i W_i | Z_i}.$$

With $V = \frac{\partial}{\partial \theta} \log f_{YXW|Z}$ the likelihood score $\dot{l} = \sum V$ is then the optimal estimating function. Integration has eliminated the nuisance parameters. However, the integral requires knowledge of $f_{U|Z}$. Since $U = u$ is not observed, it may be difficult to determine an appropriate choice for $f_{U|Z}$. If $\dot{l}$ is computed using $f_{U|Z} = f^1_{U|Z}$ but in fact the data are observed according to $f_{U|Z} = f^0_{U|Z}$, the likelihood score is no longer unbiased in general and an inconsistent estimator may result. The density $f_{U|Z}$ is now the nuisance parameter. Recall that in the functional model we considered estimating functions that were unbiased regardless of the underlying u generating the data. Here, attention is restricted to estimating functions that are unbiased for any $f_{U|Z} \in \mathcal{F}$ where $\mathcal{F}$ is a collection of densities.

The theory in this section parallels that in the previous sections. An optimal estimating function is derived by conditioning on a type of complete, sufficient statistic. The sense that the statistic is complete and sufficient is made clear in what follows.

The first task is to clearly define an appropriate class of estimating functions. Care is necessary in defining the parameter space since for certain models $\omega = \beta_0 + \beta_1^T u + \beta_2^T z$ is restricted to a segment of the real line. For the normal, logistic and Poisson models there are no restrictions but the gamma model, for example, requires $\omega \in (0, \infty)$. Let ω be restricted to H and define θ as in (4.7). The restriction to θ is the product space $\Theta = \Re \times \Re^p \times \Re^q \times \Re^+ \times \cdots \times \Re^{p^+}$. With $\tau = (\theta, f_{U|Z})$ and $\text{supp}(f_{U|Z})\equiv$ support of $f_{U|Z}$, the parameter space for the structural model is

$$T_z = \{\tau : \theta \in \Theta, f_{U|Z} \in \mathcal{F},\ \omega \in H \text{ for } u \in \text{supp}(f_{U|Z})\}.$$

The subscript denotes that the parameter space can depend on the fixed value of z. Expectations throughout this section are conditional on $Z = z$. For simplicity, the conditioning is often not indicated; $E(Y \mid U = u)$ always means $E(Y \mid U = u, Z = z)$. When convenient, the conditioning is exhibited directly.

The estimating functions considered are those ψ_+ satisfying

i) $\psi_+ = \sum_{i=1}^n \psi(Y_i, X_i, W_i; \theta, z_i)$.

ii) $E_\tau \psi(Y, X, W; \theta, z) = 0 \quad \forall \tau \in T_z$.

iii) $E_\tau \frac{\partial}{\partial \theta} \psi(Y, X, W; \theta, z) = -E_\tau\{\psi(Y, X, W; \theta, z) V^T(Y, X, W; \theta, z)\} \quad \forall \tau \in T_z$.

iv) $E_\tau \|\psi(Y, X, W; \theta, z)\| < \infty \ \forall \tau \in T_z$.

Denote the class of estimating functions satisfying i)-iv) by Ψ. The objective is to find the optimal estimating function in this class.

As in the functional model, define

$$S = Y\Omega\beta_1 + X + \Sigma_{\delta\delta}\eta\Sigma_{WW}^{-1} W \tag{4.10}$$

where $\Omega = \{a(\phi)\}^{-1}\Sigma_{\delta\delta}$ and define θ by (4.7). In the functional model, with θ fixed, the sufficiency of S for u means the distribution of (Y, X, W) given $U = u, Z = z$ and $S = s$ does not depend on u. Then it may seem to be simply true that the distribution of (Y, X, W) given $Z = z$ and $S = s$ is independent of $f_{U|Z}$. This is the sense that S is sufficient for $f_{U|Z}$ in the structural model. A proof is given in the following proposition and corollary. We first remind the reader of the measure-theoretic definition of conditional expectation.

DEFINITION 4.2: Let $(\Omega, \mathcal{T}, P)$ be a probability space. Suppose X is an integrable random variable on this space and $\mathcal{G}$ a sub-σ-field of $\mathcal{T}$. Then there exists a random variable $E(X \mid \mathcal{G})$, unique almost surely, satisfying

i) $E(X \mid \mathcal{G})$ is measurable $\mathcal{G}$

ii) $\int_G E(X \mid \mathcal{G})\, dP = \int_G X\, dP$ for all $G \in \mathcal{G}$.

A random variable satisfying i) and ii) is called a version of $E(X \mid \mathcal{G})$. Conditional probability is a special case. Let $X = I_A$. Then $P(A \mid \mathcal{G})$ is defined as $E(X \mid \mathcal{G})$.

PROPOSITION 4.1. *Suppose a random variable Y on a measurable space $(\Omega, \mathcal{T})$ has density $f_Y = h(y) \int g(S(y), u) f_U(u)\, d\lambda(u)$ with respect to a σ-finite measure π. Suppose also that h is integrable π. Then the distribution of Y given $S = s$ does not depend on f_U.*

PROOF: The proof for the discrete case is straight forward but omitted. The proof for the general case follows closely that of the sufficiency of the factorization criterion in the Factorization Theorem; see, for example, Lemma 1 of Billingsley (1986, p. 473).

Since h is integrable π, it is no restriction to assume $\int h\, d\pi = 1$. Define a

probability measure P_{f_U} over $(\Omega, \mathcal{T})$ by

$$P_{f_U}(A) = \int_A f_Y \, d\pi = \int_A \left[\int g f_U \, d\lambda \right] h \, d\pi.$$

Define a probability measure P over $(\Omega, \mathcal{T})$ by $P(A) = \int_A h \, d\pi$. Then $P_{f_U}(A) = \int_A \left[\int g f_U \, d\lambda \right] dP$. Define $\mathcal{G}$ to be the sub-σ-field of $\mathcal{T}$ induced by S, i.e. $\mathcal{G} = \sigma(S)$. Note that $\int g f_U \, d\lambda$ is measurable $\mathcal{G}$. Then for $G \in \mathcal{G}$,

$$\int_G P(A \mid \mathcal{G}) \, dP_{f_U} = \int_G E(I_A \mid \mathcal{G}) \int g f_U \, d\lambda dP.$$

Since $\int g f_U \, d\lambda$ is measurable $\mathcal{G}$, the left hand side is

$$\int_G E(I_A \int g f_U \, d\lambda \mid \mathcal{G}) \, dP.$$

By the definition of conditional expectation this last expression is

$$\int_G I_A \int g f_U \, d\lambda \, dP = \int_{A \cap G} \int g f_U \, d\lambda \, dP = P_{f_U}(A \cap G).$$

We have shown $\int_G P(A \mid \mathcal{G}) \, dP_{f_U} = P_{f_U}(A \cap G)$, i.e. that $P(A \mid \mathcal{G})$, the conditional probability calculated with respect to P, is a version of $P_{f_U}(A \mid \mathcal{G})$. Since P does not depend on f_U, the proposition follows. •

COROLLARY 4.1. *Given $U = u$ and $Z = z$, suppose Y,X and W have densities given by (4.1)-(4.3). Then the distribution of (Y, X, W) given $S = s$ and $Z = z$ does not depend on $f_{U|Z}$.*

PROOF: Recall that the distribution of Y, given $U = u$ and $Z = z$, is absolutely continuous with respect to ν. Let π denote the product measure of ν and Lebesgue measure on $\Re^{p+k}$. Then the conditional distribution of (Y, X, W) is absolutely continuous with respect to π and the associated Radon-Nikodym derivative has the factorization

$$f_{YXW|UZ} = h(y, x, w; \theta, z) g(S(y, x, w; \theta); \theta, u, z),$$

where S is given by (4.10). The corollary follows from the proposition provided h is integrable π. The function h corresponding to (4.1)-(4.3) is

$$h = \exp\left[\frac{y(\beta_0+\beta_2^T z)}{a(\phi)}+c(y,\phi)-\frac{1}{2}x^T\Sigma_{\delta\delta}^{-1}x-\frac{1}{2}(w-\xi-\gamma^T z)^T\Sigma_{WW}^{-1}(w-\xi-\gamma^T z)\right].$$

Then provided $\beta_0+\beta_2^T z \in H$, h is integrable π. •

The next step is to determine what is meant by completeness of S and then determine conditions when this is so.

DEFINITION 4.3: Let $\mathcal{S}$ denote a collection of functions and λ a measure. If

$$\int t(s)g(s)\ d\lambda = 0$$

for all $g \in \mathcal{S}$ implies $t(\cdot) = 0$ almost surely λ, then we say $\mathcal{S}$ is complete.

The distribution of S is absolutely continuous with respect to Lebesgue measure λ on $\Re^p$. Write $f_{S|Z}$ for the density of S. For each fixed z and θ there is a family $\mathcal{S}_{z,\theta}$ of such densities; $f_{S|Z} \in \mathcal{S}_{z,\theta}$ corresponds to a $f_{U|Z} \in \mathcal{F}_{z,\theta}$ where $\mathcal{F}_{z,\theta} = \{f_{U|Z} : (\theta, f_{U|Z}) \in T_z\}$.

ASSUMPTION 4.1: For each fixed z and θ, $\mathcal{S}_{z,\theta}$ is complete with respect to λ.

The sense that S is complete, provided by Definition 4.2 and Assumption 4.1, is the same as the usual meaning of completeness except the family $\mathcal{S}_{z,\theta}$ is indexed by the collection of densities $\mathcal{F}_{z,\theta}$, whereas in the classical notion a complete family is indexed by a Euclidean parameter space. However, it is not immediately clear what restrictions are imposed on $\mathcal{F}_{z,\theta}$ by Assumption 4.1. For this reason we make the following assumption and show it implies the previous one.

ASSUMPTION 4.2: For each fixed z and θ, $\mathcal{F}_{z,\theta}$ is complete with respect to $\lambda_{z,\theta}$ where $\lambda_{z,\theta}$ is Lebesgue measure restricted to $\{u \in \Re^p : \beta_0+\beta_1^T u+\beta_2^T z \in H\}$.

PROPOSITION 4.2. *Assumption 4.2 implies Assumption 4.1.*

PROOF: Fix z and θ and note

$$\begin{aligned}\int \phi(s) f_{S|Z}\, d\lambda &= \int\int \phi(s) f_{S|UZ} f_{U|Z}\, d\lambda_{z,\theta}\, d\lambda \\ &= \int \left[\int \phi(s) f_{S|UZ}\, d\lambda\right] f_{U|Z}\, d\lambda_{z,\theta}.\end{aligned}$$

Then $\int \phi(s) f_{S|Z}\, d\lambda = 0$ implies, by completeness of $\mathcal{F}_{z,\theta}$,

$$\int \phi(s) f_{S|UZ}\, d\lambda = 0. \tag{4.11}$$

Now, for fixed θ, the distribution of S given $U = u$ and $Z = z$ is an exponential family in u with parameter space $\mathcal{U} = \{u \ : \ \beta_0 + \beta_1^T u + \beta_2^T z \in H\}$ an open subset of $\Re^p$. This implies the family $\{f_{S|UZ}(s; u, z)\}_{u \in \mathcal{U}}$ is complete so that (4.11) implies $\phi(s) = 0$ almost everywhere λ. •

Note that $\mathcal{F}_{\theta,z}$ is complete with respect to $\lambda_{\theta,z}$ provided $\mathcal{F}_{\theta,z}$ contains a complete family. For example, if $\{u \ : \ \beta_0 + \beta_1^T u + \beta_2^T z \in H\} = \Re^p$ then $\lambda_{\theta,z}$ is Lebesgue measure on $\Re^p$ and $\mathcal{F}_{\theta,z} = \mathcal{F}$ is complete if it contains all the p-dimensional normal densities.

We are now in a position to identify the optimal estimating function. Suppose that $f_{U|Z}$ is the density generating U given $Z = z$ and define

$$\begin{aligned}\psi_+^* = \sum_{i=1}^n \psi_i^* = \sum_{i=1}^n \Big[& \dot{l}(y_i, x_i, w_i; \theta, f_{U|Z_i}, z_i) \\ & - E\left\{\dot{l}(Y_i, X_i, W_i; \theta, f_{U|Z_i}, z_i) \mid S_i = s_i, Z_i = z_i\right\}\Big].\end{aligned} \tag{4.12}$$

The expectation in the expression above is taken with respect to the distribution of (Y, X, W) given $S = s$ and $Z = z$. By Corollary 4.1, this distribution is independent of $f_{U|Z}$, implying that ψ^* indeed satisfies $E_\tau \psi^* = 0$ for $\tau \in T_z$. As in the functional model, completeness of $\mathcal{S}_{z,\theta}$ implies the optimality and uniqueness of ψ_+^*. This is formally stated in the next theorem.

THEOREM 4.2. *Let* $\{(Y_i, X_i, W_i, Z_i)\}_{i=1}^n$ *be* n *i.i.d. random variables such that, given* $Z_i = z_i$, (Y_i, X_i, W_i) *has density (4.9). Then Assumption 4.1 or Assumption 4.2 implies*

$$E\psi_s\psi_s^T - E\psi_s^*\psi_s^{*T} \geq_{p.d.} 0$$

for $\psi_+ \in \Psi$ *with equality iff* $\psi_s = \psi_s^*$ *almost surely.*

PROOF: The proof is identical to that of Theorem 4.1. •

Before remarking on the Theorem we first derive the form of the optimal estimating function. To compute the optimal estimating function, note that

$$\begin{aligned} \dot{l}(y,x,w;\theta,z) &= \frac{\int \frac{\partial}{\partial\theta} f_{YXW|UZ} f_{U|Z}\, d\lambda}{f_{YXW|Z}} \\ &= E\Bigg[\frac{\partial}{\partial\theta}\log f_{YXW|UZ}(y,x,w;\theta,U,z) \\ &\qquad\qquad \mid Y=y, X=x, W=w, Z=z\Bigg]. \end{aligned}$$

Now, sufficiency of S for u implies the distribution of $U \mid Y, X, W, Z$ is equivalent to the distribution of $U \mid S, Z$. A formal argument for this is that $f_{YXW|UZS} = f_{YXW|ZS}$ implies

$$f_{U|SZ} = \frac{f_{USZ}}{f_{SZ}} = \frac{f_{YXWUZS}}{f_{YXWZS}} = f_{U|YXWZS} = f_{U|YXWZ}.$$

Then

$$\dot{l}(y,x,w;\theta,z) = E\left[\frac{\partial}{\partial\theta}\log f_{YXW|UZ}(y,x,w;\theta,U,z) \mid S=s, Z=z\right]. \quad (4.13)$$

Using (4.12) and (4.13) it follows that the optimal estimating function is of the

form

$$
\begin{aligned}
& Y - E(Y \mid S) \\
& \{Y - E(Y \mid S)\} E(U \mid S) \\
& \{Y - E(Y \mid S)\} z \\
& \frac{\partial}{\partial \phi} c(Y, \phi) - E\left\{ \frac{\partial}{\partial \phi} c(Y, \phi) \mid S \right\} \\
& W - E(W \mid S) \\
& \text{vec}\left[\{W - E(W \mid S)\} E(U^T \mid S)\right] \\
& \text{vec}\left[\{W - E(W \mid S)\} z^T\right] \\
& \text{vech}\left\{WW^T - E(WW^T \mid S)\right\} \\
& \text{diag}\left\{XX^T - E(XX^T \mid S)\right\}. \qquad (4.14)
\end{aligned}
$$

The discussion following Theorem 4.1 is applicable here with little change. Knowledge of $f_{U|Z}$ is required to find ψ_+^*. However, if $f_{U|Z}$ was known, $\dot{l}$ is the optimal estimating function. Constructing $\psi_{+,1}^*$ by specifying $f_{U|Z} = f_{U|Z}^1$ may not be optimum in Ψ if $f_{U|Z} = f_{U|Z}^0 \neq f_{U|Z}^1$. However, $\psi_{+,1}^*$ is still unbiased, and regularity conditions ensure the existence of a consistent estimator. Furthermore, the optimal estimating function depends on $f_{U|Z}$ only through $E(U \mid S, Z) \equiv E(U \mid S)$, and then as a sequence of weights.

As noted in the functional model, the optimal estimating function remains unbiased when $E(U \mid S)$ is replaced by any function of S. The first function that comes to mind is the identity function, that is replace $E(U \mid S)$ by S. The resulting estimating function is not optimal unless $E(U \mid S)$ is linear in S. The linear and logistic models are explored in the next section and conditions that imply $E(U \mid S)$ is linear in S are given. The implication is that, while replacing

$E(U \mid S)$ with S is not always optimal, it is at least optimal in certain cases.

4.5 The optimal estimating function in linear and logistic regression

Recall from Chapters 2 and 3 that analysis of the linear and probit models required $\dim(W) \geq \dim(X)$. This is also the case here. Suppose that $E(U \mid S)$ is replaced by S in (4.8). Write $X = S - Y\Omega\beta_1 - \Sigma_{\delta\delta}\eta\Sigma_{WW}^{-1}W$. Then for purposes of estimation in the logistic and normal theory linear models, the component $\text{diag}\{XX^T - E(XX^T \mid S)\}$ is equivalent to

$$\eta\Sigma_{WW}^{-1}\{YW - E(YW \mid S)\}.$$

If $\dim(W) < \dim(X)$, then

$$\eta\Sigma_{WW}^{-1}\sum_{i=1}^{n}\{Y_iW_i - E(Y_iW_i \mid S_i)\} \tag{4.15}$$

contains at most $\dim(W)$ linearly independent estimating equations. However, (4.15) is trying to estimate the $\dim(X)$ parameters in $\Sigma_{\delta\delta}$. Evidently the number of instruments must be at least as great as the number of covariates measured with error.

4.5.1 The linear model

The task is to analyze ψ_+^* corresponding to the following normal theory linear model;

$$Y = \beta_0 + \beta_1^T U + \epsilon$$

$$X = U + \delta$$

$$W = \xi + \eta^T U + \zeta$$

$$\begin{pmatrix} U \\ \delta \\ \epsilon \\ \zeta \end{pmatrix} \sim N\left\{\begin{pmatrix} \mu_U \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{UU} & 0 & 0 & 0 \\ 0 & \Sigma_{\delta\delta} & 0 & 0 \\ 0 & 0 & \sigma_{\epsilon\epsilon}^2 & 0 \\ 0 & 0 & 0 & \Sigma_{WW|U} \end{pmatrix}\right\}. \tag{4.16}$$

For simplicity, the covariates Z are omitted. Consistent with the notation in (4.3), $\Sigma_{WW|U}$ is used instead of $\Sigma_{\zeta\zeta}$. This model is of the form (4.9). Define

$$\theta = (\beta_0, \beta_1^T, \sigma_{\epsilon\epsilon}^2, \mathrm{diag}^T(\Sigma_{\delta\delta}), \xi^T, \mathrm{vec}^T\eta^T, \mathrm{vech}^T\Sigma_{WW|U})^T,$$
$$\theta_+ = (\theta^T, \mu_U^T, \mathrm{vech}^T\Sigma_{UU})^T,$$

and $f_{YXW}(\theta_+)$ the density of an observation (Y, X, W). Let

$$\dot{l}(\theta_+) = \frac{\partial}{\partial\theta_+}\log f_{YXW}(\theta_+).$$

PROPOSITION 4.3. *The maximum likelihood estimator for θ_+ under the model (4.16) is a solution to the optimal estimating function.*

PROOF: Using the same argument leading to (4.13) it follows that

$$\frac{\partial}{\partial(\mu_U^T, \mathrm{vech}^T\Sigma_{UU})^T}\log f_{YXW} = E\left[\frac{\partial}{\partial(\mu_U^T, \mathrm{vech}^T\Sigma_{UU})^T}\log f_U(U) \mid S\right]$$

where f_U is the normal density corresponding to (4.16). The components of the maximum likelihood score corresponding to μ_U and Σ_{UU} are then

$$\begin{aligned} \mu_U: &\quad E(U \mid S) - \mu_U \\ \Sigma_{UU}: &\quad E\left\{(U-\mu_U)(U-\mu_U)^T \mid S\right\} - \Sigma_{UU}. \end{aligned} \tag{4.17}$$

In the following, repeated use is made of the joint normality of $(Y, X, W, U, \epsilon, \delta, \zeta)$. Note

$$E(U \mid S) = \mu_U + \Sigma_{SU}\Sigma_{SS}^{-1}(S - \mu_S)$$
$$E(UU^T \mid S) = \Sigma_{UU} - \Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{SU} + E(U \mid S)E(U \mid S)^T,$$

so that the components for μ_U and Σ_{UU} can be written

$$\begin{aligned} \mu_U: &\quad S - \mu_S \\ \Sigma_{UU}: &\quad (S-\mu_S)(S-\mu_S)^T - \Sigma_{SS}. \end{aligned} \tag{4.18}$$

The optimal estimating function is constructed by summing the terms $\psi_i^* = \dot{l}_i - E(\dot{l}_i \mid S_i)$, $i = 1, \ldots, n$. Since the maximum likelihood estimator solves $\sum_{i=1}^n \dot{l}_i = 0$, the proposition will follow provided the maximum likelihood estimator also solves $\sum_{i=1}^n E(\dot{l}_i \mid S_i) = 0$. We show that $E(\dot{l} \mid S)$ is a matrix multiple of (4.18), or more accurately that $\text{vec}\{E(\dot{l} \mid S)\}$ is a matrix multiple of $\left[(S-\mu_S)^T, \text{vec}^T\{(S-\mu_S)(S-\mu_S)^T - \Sigma_{SS}\}\right]^T$.

Starting with (4.13), the components of $E(\dot{l} \mid S)$ are easily seen to be

$$
\begin{aligned}
\beta_0: &\quad E(Y \mid S) - \beta_0 - \beta_1^T E(U \mid S) \\
\beta_1: &\quad \{E(Y \mid S) - \beta_0\}E(U \mid S) - E(UU^T \mid S)\beta_1 \\
\sigma^2_{\epsilon\epsilon}: &\quad E(Y^2 \mid S) - 2E(Y \mid S)\beta_0 + \beta_0^2 - \beta_1^T E(UU^T \mid S)\beta_1 - \sigma^2_{\epsilon\epsilon} \\
\text{diag}(\Sigma_{\delta\delta}): &\quad \text{diag}\Big[E(XX^T \mid S) - 2E(X \mid S)E(U^T \mid S) \\
&\qquad\qquad + E(UU^T \mid S) - \Sigma_{\delta\delta}\Big] \\
\xi: &\quad E(W \mid S) - \xi - \eta^T E(U \mid S) \\
\eta^T: &\quad \{E(W \mid S) - \xi\}E(U^T \mid S) - \eta^T E(UU^T \mid S) \\
\Sigma_{WW|U}: &\quad E(WW^T \mid S) - \xi E(W^T \mid S) - E(W \mid S)\xi^T - \\
&\qquad\qquad \eta^T E(UU^T \mid S)\eta - \Sigma_{WW|U}. \qquad (4.19)
\end{aligned}
$$

Note that $E(\dot{l} \mid S)$ depends on S only through $(S-\mu_S)$ and $(S-\mu_S)(S-\mu_S)^T$.

First consider the β_0 component;

$$
E(Y \mid S) - \beta_0 - \beta_1^T E(U \mid S) = E(\epsilon \mid S) = \Sigma_{\epsilon S}\Sigma_{SS}^{-1}(S-\mu_S),
$$

so that the β_0 component is a matrix multiple of (4.18). For the β_1 component,

$$\begin{aligned}
&\{E(Y \mid S) - \beta_0\}E(U \mid S) - E(UU^T \mid S)\beta_1 \\
&\quad = E(U \mid S)E(\epsilon \mid S) - \Sigma_{UU}\beta_1 + \Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{SU}\beta_1 \\
&\quad = \mu_U\Sigma_{\epsilon S}\Sigma_{SS}^{-1}(S - \mu_S) + \Sigma_{US}\Sigma_{SS}^{-1}\left\{(S - \mu_S)(S - \mu_S)^T - \Sigma_{SS}\right\}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} \\
&\qquad + \Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} - \Sigma_{UU}\beta_1 + \Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{SU}\beta_1.
\end{aligned}$$

Recall that $f_{Y|SU} = f_{Y|S}$ so that $\beta_{Y|S\underline{U}} = 0$, where $\beta_{Y|S\underline{U}}$ is the coefficient of U in the linear regression $E(Y \mid S, U)$. It is not hard to show

$$\beta_{Y|S\underline{U}} = -\Sigma_{UU|S}^{-1}\Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{SY} + \Sigma_{UU|S}^{-1}\Sigma_{UY}.$$

Since $\Sigma_{SY} = \Sigma_{SU}\beta_1 + \Sigma_{S\epsilon}$ and $\Sigma_{UY} = \Sigma_{UU}\beta_1$ it follows

$$\Sigma_{US}\Sigma_{SS}^{-1}(\Sigma_{S\epsilon} + \Sigma_{SU}\beta_1) - \Sigma_{UU}\beta_1 = 0$$

so that the β_1 component is a matrix multiple of (4.18).

Next consider the $\sigma_{\epsilon\epsilon}^2$ component;

$$\begin{aligned}
&E(Y^2 \mid S) - 2E(Y \mid S)\beta_0 + \beta_0^2 - \beta_1^T E(UU^T \mid S)\beta_1 - \sigma_{\epsilon\epsilon}^2 \\
&\quad = \sigma_{YY|S}^2 + (\beta_0 + \beta_1^T E(U \mid S) + E(\epsilon \mid S))^2 \\
&\qquad - 2(\beta_0 + \beta_1^T E(U \mid S) + E(\epsilon \mid S))\beta_0 + \beta_0^2 \\
&\qquad - \beta_1^T(\Sigma_{UU|S} + E(U \mid S)E(U \mid S)^T)\beta_1 - \sigma_{\epsilon\epsilon}^2 \\
&\quad = \beta_1^T\Sigma_{UU}\beta_1 - \Sigma_{YS}\Sigma_{SS}^{-1}\Sigma_{SY} + E(\epsilon \mid S)^2 \\
&\qquad + 2\beta_1^T E(U \mid S)E(\epsilon \mid S) - \beta_1^T\Sigma_{UU|S}\beta_1 \\
&\quad = \Sigma_{\epsilon S}\Sigma_{SS}^{-1}\left\{(S - \mu_S)(S - \mu_S)^T - \Sigma_{SS}\right\}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} \\
&\qquad + 2\beta_1^T\mu_U(S - \mu_S)\Sigma_{SS}^{-1}\Sigma_{S\epsilon} + 2\beta_1^T\Sigma_{US}\Sigma_{SS}^{-1} \\
&\qquad \times \left\{(S - \mu_S)(S - \mu_S)^T - \Sigma_{SS}\right\}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} + \Sigma_{\epsilon S}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} \\
&\qquad + 2\beta_1^T\Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} + \beta_1^T\Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{SU}\beta_1 - \Sigma_{YS}\Sigma_{SS}^{-1}\Sigma_{SY}.
\end{aligned}$$

Since $\Sigma_{YS} = \beta_1^T \Sigma_{US} + \Sigma_{\epsilon S}$, it easily follows

$$\Sigma_{\epsilon S}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} + 2\beta_1^T \Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{S\epsilon} - \Sigma_{YS}\Sigma_{SS}^{-1}\Sigma_{SY} + \beta_1^T \Sigma_{US}\Sigma_{SS}^{-1}\Sigma_{SU}\beta_1 = 0.$$

The arguments for the remaining components are very similar and therefore are omitted. •

In the course of proving the proposition it was argued that $E(U \mid S)$ is linear in S for the model (4.16). Thus, for the normal theory linear model, that is when $f_{Y|U}$, $f_{X|U}$ and $f_{W|U}$ are normal densities, replacing $E(U \mid S)$ with S is optimal when f_U is also a normal density.

4.5.2 Logistic regression

The first task is to derive the optimal estimating function for logistic regression. Proposition 4.4 gives conditions that ensure $E(U \mid S)$ is linear in S for the logistic model. A simulation study for the logistic model is presented in Section 4.6.

The Jacobian of the transformation that takes (Y, X, W) into (Y, S, W) has determinant one. Define $F(x) = \{1 + \exp(-x)\}^{-1}$, $F = F(\beta_0 + \beta_1^T u + \beta_2^T z)$ and $\mu = \xi + \eta^T u + \gamma^T z$. Then

$$\begin{aligned} f_{Y,S,W} = {} & F^Y (1-F)^{1-Y} (2\pi)^{-\frac{p}{2}} |\Sigma_{\delta\delta}|^{\frac{1}{2}} \\ & \times \exp\left\{ -\frac{1}{2}(S - \Sigma_{\delta\delta}\beta_1 Y - \Sigma_{\delta\delta}\eta\Sigma_{WW|UZ}^{-1} W - u)^T \right. \\ & \left. \times \Sigma_{\delta\delta}^{-1}(S - \Sigma_{\delta\delta}\beta_1 Y - \Sigma_{\delta\delta}\eta\Sigma_{WW|UZ}^{-1} W - u) \right\} \\ & \times (2\pi)^{-\frac{k}{2}} |\Sigma_{WW|UZ}|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(W-\mu)^T \Sigma_{WW|UZ}^{-1}(W-\mu) \right\}. \end{aligned}$$

Define

$$\mu_Y = S - \Sigma_{\delta\delta}\beta_1 Y - u,$$
$$B_Y = \Sigma^{-1}_{WW|UZ}(\mu + \eta^T\mu_Y) = \Sigma^{-1}_{WW|UZ}\{\eta^T(S - \Sigma_{\delta\delta}\beta_1 Y) + \xi + \gamma^T z\},$$
$$A^{-1} = \Sigma^{-1}_{WW|UZ}(\Sigma_{WW|UZ} + \eta^T\Sigma_{\delta\delta}\eta)\Sigma^{-1}_{WW|UZ},$$

and $c = (2\pi)^{-\frac{p}{2}}|\Sigma_{\delta\delta}|^{-\frac{1}{2}}|\Sigma_{WW|UZ}|^{-\frac{1}{2}}|A|^{\frac{1}{2}}$. Then

$$f_{Y,S,W} = F^Y(1-F)^{1-Y}c|A|^{-\frac{1}{2}}(2\pi)^{-\frac{k}{2}} \times \exp\left[-\frac{1}{2}\left\{(W - AB_Y)^T A^{-1}(W - AB_Y) - B_Y^T AB_Y + \mu_Y^T\Sigma^{-1}_{\delta\delta}\mu_Y + \mu^T\Sigma^{-1}_{WW|UZ}\mu\right\}\right]. \tag{4.20}$$

Define μ_1 as μ_Y evaluated at $Y = 1$, that is $\mu_1 = S - \Sigma_{\delta\delta}\beta_1 - u$. Define μ_0, B_0, and B_1 in the same manner. Let $\Delta_1 = -B_1^T AB_1 + \mu_1^T\Sigma^{-1}_{\delta\delta}\mu_1 + \mu^T\Sigma^{-1}_{WW|UZ}\mu$ and $\Delta_0 = -B_0^T AB_0 + \mu_0^T\Sigma^{-1}_{\delta\delta}\mu_0 + \mu^T\Sigma^{-1}_{WW|UZ}\mu$. With this notation it readily follows that

$$f_{Y,S} = F^Y(1-F)^{1-Y}c\exp\left\{-\frac{1}{2}(-B_Y^T AB_Y + \mu_Y^T\Sigma^{-1}_{\delta\delta}\mu_Y + \mu^T\Sigma^{-1}_{WW|UZ}\mu)\right\}$$

and

$$f_S = Fc\exp\left(-\frac{1}{2}\Delta_1\right) + (1-F)c\exp\left(-\frac{1}{2}\Delta_0\right). \tag{4.21}$$

Then

$$\begin{aligned}\Pr(Y = 1 \mid S) &= \frac{Fc\exp(-\frac{1}{2}\Delta_1)}{Fc\exp(-\frac{1}{2}\Delta_1) + (1-F)c\exp(-\frac{1}{2}\Delta_0)} \\ &= \frac{1}{1 + \exp(-\beta_0 - \beta_1^T u - \beta_2^T z - \frac{1}{2}(\Delta_0 - \Delta_1))} \\ &= F(\beta_0^* + \beta_1^{*T} S + \beta_2^{*T} z)\end{aligned}$$

where

$$\beta_0^* = \beta_0 - \xi^T(\Sigma_{WW|UZ} + \eta^T\Sigma_{\delta\delta}\eta)^{-1}\eta^T\Sigma_{\delta\delta}\beta_1 + \frac{1}{2}\beta_1^T\Sigma_{\delta\delta}\left\{\eta(\Sigma_{WW|UZ} + \eta^T\Sigma_{\delta\delta}\eta)^{-1}\eta^T - \Sigma_{\delta\delta}^{-1}\right\}\Sigma_{\delta\delta}\beta_1$$

$$\beta_1^* = \{I - \eta(\Sigma_{WW|UZ} + \eta^T\Sigma_{\delta\delta}\eta)^{-1}\eta^T\Sigma_{\delta\delta}\}\beta_1$$

and

$$\beta_2^* = \beta_2 - \gamma(\Sigma_{WW|UZ} + \eta^T\Sigma_{\delta\delta}\eta)^{-1}\eta^T\Sigma_{\delta\delta}\beta_1.$$

Next, note

$$E(W \mid S) = \int w f_{W|S}\, dw = \int w \sum_{y=0}^{1} \frac{f_{Y,S,W}}{f_S}\, dw.$$

Using (4.20) and (4.21) if readily follows that

$$\begin{aligned} E(W \mid S) &= \frac{Fc\exp(-\frac{1}{2}\Delta_1)AB_1 + (1-F)c\exp(-\frac{1}{2}\Delta_0)AB_0}{Fc\exp(-\frac{1}{2}\Delta_1) + (1-F)c\exp(-\frac{1}{2}\Delta_0)} \\ &= AB_1\Pr(Y=1 \mid S) + AB_0\Pr(Y=0 \mid S). \end{aligned}$$

In a similar manner we compute

$$E(WW^T \mid S) = (A + AB_1B_1^TA)\Pr(Y=1 \mid S) + (A + AB_0B_0^TA)\Pr(Y=0 \mid S).$$

Finally,

$$\begin{aligned} E(YW \mid S) &= \sum_{y=0}^{1} \int yw \frac{f_{Y,W,S}}{f_S}\, dw \\ &= \sum_{y=0}^{1} \left[\frac{yAB_y}{f_S} F^y(1-F)^{1-y}c \right. \\ &\qquad \left. \times \exp\left\{ -\frac{1}{2}(-B_y^TAB_y + \mu_y^T\Sigma_{\delta\delta}^{-1}\mu_y + \mu^T\Sigma_{WW|UZ}^{-1}\mu) \right\} \right] \\ &= AB_1\Pr(Y=1 \mid S). \end{aligned}$$

We now have all the components to construct the optimal estimating function.

The next proposition gives conditions that imply $E(U \mid S)$ is linear in S.

PROPOSITION 4.4. *Suppose* $U \mid Y \sim N(\mu_Y, \Sigma)$ *and* $Pr(Y = 1) = \pi_1 = 1 - Pr(Y = 0)$. *Also suppose* $\delta \sim N(0, \Sigma_{\delta\delta})$, $\zeta \sim N(0, \Sigma_{WW|U})$ *and*

$$X = U + \delta$$
$$W = \xi + \eta^T U + \zeta$$
$$S = \Sigma_{\delta\delta}\beta_1 Y + X + \Sigma_{\delta\delta}\eta\Sigma^{-1}_{WW|U}W.$$

Then

$$Pr(Y = 1 \mid U = u) = F(\beta_0 + \beta_1^T u) = \{1 + \exp(-\beta_0 - \beta_1^T u)\}^{-1}$$

where

$$\beta_0 = \log\left(\frac{\pi_1}{1-\pi_1}\right) + \frac{1}{2}(\mu_0^T\Sigma^{-1}\mu_0 - \mu_1^T\Sigma^{-1}\mu_1),$$
$$\beta_1 = \Sigma^{-1}(\mu_1 - \mu_0).$$

Also,

$$E(U \mid S) = (\Sigma^{-1} + \Sigma_{\delta\delta}^{-1} + \eta\Sigma^{-1}_{WW|U}\eta^T)^{-1}(\Sigma_{\delta\delta}^{-1}S + \Sigma_{SS}^{-1}\mu_0 - \eta\Sigma^{-1}_{WW|U}\xi).$$

PROOF: It is well known that the conditions of the theorem imply

$$\Pr(Y = 1 \mid U = u) = F(\beta_0 + \beta_1^T u)$$

where β_0 and β_1 are defined as above. The interest lies in showing that $E(U \mid S)$ is linear in S. The Jacobian of the transformation that takes

$$\begin{pmatrix} Y \\ U \\ \delta \\ \zeta \end{pmatrix} \quad \text{into} \quad \begin{pmatrix} Y \\ U \\ S \\ \zeta \end{pmatrix}$$

has determinant one. A routine, though tedious, calculation shows that

$$f_{Y,U,S} = (1-\pi_1)^{1-Y}\pi_1^Y|\Sigma_{\delta\delta}|^{-\frac{1}{2}}|\Sigma_{WW|U}|^{-\frac{1}{2}}|E|^{\frac{1}{2}}(2\pi)^{-p} \times \exp\left[-\frac{1}{2}\left\{(u-\mu_Y)^T\Sigma^{-1}(u-\mu_Y) + (\bar{C}u-\bar{\mu})^T D(\bar{C}u-\bar{\mu})\right\}\right]$$

where

$$\begin{aligned}
\bar{\mu} &= S - \Sigma_{\delta\delta}\beta_1 Y - \Sigma_{\delta\delta}\eta\Sigma_{WW|U}^{-1}\xi, \\
\bar{C} &= I + \Sigma_{\delta\delta}\eta\Sigma_{WW|U}^{-1}\eta^T, \\
D &= \Sigma_{\delta\delta}^{-1} - \eta(\Sigma_{WW|U} + \eta^T\Sigma_{\delta\delta}\eta)^{-1}\eta^T, \\
E &= \Sigma_{WW|U}(\Sigma_{WW|U} + \eta^T\Sigma_{\delta\delta}\eta)^{-1}\Sigma_{WW|U}.
\end{aligned}$$

This expression is equivalent to

$$f_{Y,U,S} = (1-\pi_1)^{1-Y}\pi_1^Y|\Sigma_{\delta\delta}|^{-\frac{1}{2}}|\Sigma_{WW|U}|^{-\frac{1}{2}}|E|^{\frac{1}{2}}(2\pi)^{-p} \times \exp\left[-\frac{1}{2}\left\{(u-A^{-1}B)^T A(u-A^{-1}B) + \mu_Y^T\Sigma^{-1}\mu_Y + \bar{\mu}^T D\bar{\mu} - B^T AB\right\}\right]$$

where

$$\begin{aligned}
A &= \Sigma^{-1} + \bar{C}^T D\bar{C}, \\
B &= \bar{C}^T D\bar{\mu} + \Sigma^{-1}\mu_Y.
\end{aligned}$$

Straightforward algebra shows $\bar{C}^T D = \Sigma_{\delta\delta}^{-1}$. Recall $\beta_1 = \Sigma^{-1}(\mu_1 - \mu_0)$ so that

$$\begin{aligned}
A &= \Sigma^{-1} + \Sigma_{\delta\delta}^{-1} + \eta\Sigma_{WW|U}^{-1}\eta^T, \\
B &= \Sigma_{\delta\delta}^{-1}S + \Sigma^{-1}\mu_0 - \eta\Sigma_{WW|U}^{-1}\xi.
\end{aligned}$$

Now, sufficiency of S for u means $f_{Y|SU} = f_{Y|S}$ which implies that $f_{U|S} = f_{U|SY}$. Then it readily follows that

$$f_{U|S} = (2\pi)^{-\frac{p}{2}}|A|^{\frac{1}{2}} \exp\left[-\frac{1}{2}\{(u - A^{-1}B)^T A(u - A^{-1}B)\}\right]$$

so that

$$\begin{aligned} E(U \mid S) &= A^{-1}B \\ &= (\Sigma^{-1} + \Sigma_{\delta\delta}^{-1} + \eta\Sigma_{WW|U}^{-1}\eta^T)^{-1}(\Sigma_{\delta\delta}^{-1}S + \Sigma^{-1}\mu_0 - \eta\Sigma_{WW|U}^{-1}\xi). \quad \bullet \end{aligned}$$

4.6 A simulation study

As in Section 3.8, the specifications for the simulation study here are based on a rough match to the Framingham Heart Study data, see Gordon and Kannel (1968). The model is such that one covariate is measured with error and two instruments are available. See Section 3.8 for details.

The specification of the mis-measured covariate U and the covariate Z is

$$\begin{pmatrix} U \\ Z \end{pmatrix} \sim N\left\{\begin{pmatrix} 133 \\ 46 \end{pmatrix}, \begin{pmatrix} 302 & 42 \\ 42 & 73 \end{pmatrix}\right\}. \tag{4.22}$$

Conditional on $U = u$ and $Z = z$, the specification for the instruments is

$$\begin{aligned} \begin{pmatrix} W_1 \\ W_2 \end{pmatrix} &= \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} + \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} u + \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix} z + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \\ &= \begin{pmatrix} -126.4 \\ .5 \end{pmatrix} + \begin{pmatrix} .5 \\ .1 \end{pmatrix} u + \begin{pmatrix} 3.9 \\ .5 \end{pmatrix} z + \begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \end{aligned}$$

where

$$\begin{pmatrix} \zeta_1 \\ \zeta_2 \end{pmatrix} \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 103 & 0 \\ 0 & 111 \end{pmatrix}\right\}.$$

The specification for X is

$$X = U + \delta$$

where $\delta \sim N(0, 121)$. Values for $(\beta_0, \beta_1, \beta_2)$ were obtained by exploiting the close relationship between the logistic and probit models. In particular, the logistic distribution can be written as a scale mixture of the normal cumulative distribution function, see Stefanski (1990). Recall that in the probit model values for $(\beta_0, \beta_1, \beta_2)$ were obtained by specifying values for $E(Y)$,CORR(Y, U) and CORR(Y, Z) and then solving the resulting equations for $(\beta_0, \beta_1, \beta_2)$, see Section 3.8. Values for the logistic model were obtained by scaling the values computed for the probit model by $(.58763)^{-1}$, see Balakrishnan (1992, p. 532). For example, when U and Z are distributed as in (4.22), the values $(\beta_0, \beta_1, \beta_2) = (-6.939, .021, .042)$ correspond closely to setting $E(Y) = .1$, CORR$(Y, U) = .15$ and CORR$(Z, U) = .15$ and solving the resulting equations for $(\beta_0, \beta_1, \beta_2)$ when the logistic model holds. The exact computations could be done, but the approximation is quite good and also convenient since values were already computed for the probit model. As in the probit model, CORR(Y, Z) was set at .15 and CORR(Y, U) varied at .15, .25 and .35.

Three estimators were computed: An estimator that ignores measurement error, termed the naive estimator; the estimator corresponding to the optimal estimating function where $E(U \mid S)$ is replaced by S, termed the conditional estimator; and a one-step version of the conditional estimator as described below.

The conditional estimator and its one-step version were computed by application of the Gauss-Newton algorithm to the optimal estimating function. The algorithm requires starting values, computed as follows. Starting values for the parameters $(\xi_1, \eta_1, \gamma_1)$ of the linear regression of W_1 on (U, Z) were obtained using the two-stage least squares instrumental variable estimator using (W_2, Y)

as instruments., see Judge et. al. (1985). Starting values for $(\xi_2, \eta_2, \gamma_2)$ were obtained using (W_1, Y) as instruments for the linear regression of W_2 on (U, Z). Since

$$\eta^T \sigma_{UU}^2 = \Sigma_{WX} - \gamma^T \sigma_{ZX},$$

σ_{UU}^2 was estimated using

$$\hat{\sigma}_{UU}^2 = (\hat{\eta}\hat{\eta}^T)^{-1}\hat{\eta}(\hat{\Sigma}_{WX} - \hat{\sigma}_{ZX})$$

where $\hat{\Sigma}_{WX}$ and $\hat{\sigma}_{ZX}$ are the sample moment estimators of Σ_{WX} and σ_{ZX} respectively. The measurement error variance was estimated by subtraction;

$$\hat{\sigma}_{\delta\delta}^2 = \hat{\sigma}_{XX}^2 - \hat{\sigma}_{UU}^2.$$

Starting values for $(\beta_0, \beta_1, \beta_2)$ were the naive estimates, that is the estimates obtained from the logistic regression of Y on (X, Z). Finally, the starting value for $\Sigma_{WW|UZ}$ was obtained via the relationship

$$\Sigma_{WW|UZ} = \Sigma_{WW} - \eta^T \sigma_{UU}^2 \eta - \gamma^T \sigma_{ZZ}^2 \gamma - 2\eta^T \sigma_{XZ} \gamma.$$

The conditional estimator proved difficult to compute. Part of the difficulty may be attributed to the use of numerical rather than analytical derivatives in the Gauss-Newton algorithm. Numerical derivatives were used because the analytical derivatives are quite complicated. The proportions of times the algorithm computing the conditional estimator did not converge were .118, .064 and .029 when CORR(Y, U) was .15, .25 and .35, respectively. Apparently convergence properties of the algorithm depended on CORR(Y, U). The reason is because as CORR(Y, U) increases, Y becomes a better instrument for U so that better starting values are obtained.

The component of the estimating function corresponding to $\sigma^2_{\delta\delta}$ seemed the problem when convergence did not occur. To illustrate, partition the optimal estimating function as

$$\psi^*_+ = \begin{pmatrix} \psi^*_1(\theta) \\ \psi^*_2(\theta) \end{pmatrix}$$

and θ by $\theta = (\underline{\theta}^T, \sigma^2_{\delta\delta})^T$ where

$$\underline{\theta} = (\beta_0, \beta_1^T, \beta_2^T, \xi^T, \text{vec}^T\eta, \text{vec}^T\gamma, \text{vech}^T\Sigma_{WW|UZ})^T.$$

The component ψ^*_1 contains the estimating equations for $\underline{\theta}$ and ψ^*_2 is the estimating equation for $\sigma^2_{\delta\delta}$. The estimate $\hat{\underline{\theta}}$ solving

$$\psi^*_1(\underline{\theta}, \hat{\sigma}^2_{\delta\delta}) = 0$$

was easily obtained when $\sigma^2_{\delta\delta} = \hat{\sigma}^2_{\delta\delta}$ is fixed at its starting value. Usually only four iterations of the algorithm were required, so that there was no difficulty computing the one-step estimator which is defined as follows.

i) Obtain $\hat{\underline{\theta}}$ by solving $\psi^*_1(\underline{\theta}, \hat{\sigma}^2_{\delta\delta}) = 0$.

ii) Compute the one-step estimator $\tilde{\theta}_{1S} = (\tilde{\underline{\theta}}^T, \tilde{\sigma}^2_{\delta\delta})^T$ by applying one step of the Gauss-Newton algorithm to the complete set of estimating equations ψ^*_+, using $(\hat{\underline{\theta}}, \hat{\sigma}^2_{\delta\delta})$ as the starting value.

Table 4.1 presents the estimated bias and mean squared error of the naive, conditional and one-step estimators. The results are based on 500 data sets for which the conditional estimator converged. A $\diamond$ between two values indicates the difference was significant at the 5% level but both individual values were not significantly different form zero at the 5% level. Thus, a $\diamond$ indicates statistically significant differences that are probably not practically significant. A * between two values indicates the difference was significant at the 5% level and at least

one individual value was significantly different from zero. Thus, a * indicates statistically significant differences that may be practically significant. The conditional and one-step estimators behaved similarly and did substantially better than the naive estimator in terms of bias and mean squared error.

Table 4.2 presents the relative efficiency of the one-step estimator to the conditional estimator. Table 4.3 compares bias and mean squared error of the one-step estimator between data sets where the conditional estimator converged to when it did not converge.

The conditional and one-step estimators are M-estimators. Then regularity conditions imply the asymptotic variances of these estimators can be computed via the formula given in Theorem 3.5. For example, an estimate of the variance of the conditional estimator $\hat{\theta}$ is

$$\left(\dot{\psi}_+^*(\hat{\theta})\right)^{-1}\left(\sum_{i=1}^{n}\psi_i^*(\hat{\theta})\psi_i^{*^T}(\hat{\theta})\right)\left(\dot{\psi}_+^*(\hat{\theta})\right)^{-T}.$$

The variances computed in this fashion were much too large. The substitution of numerical derivatives for analytical derivatives may be the problem and perhaps bootstrap or jackknife estimates would do better. Further research is needed.

4.7 Summary

This chapter identified optimal estimating functions in a class of generalized linear measurement error models when instrumental variables were available. The form of the optimal estimating function was nearly identical for the functional and non-parametric structural measurement error models. In both models, the optimal estimating function depends on quantities not available to the statistician; the functional model required knowledge of the sequence $\underline{u}$ and the non-parametric structural model required knowledge of $E(U \mid S)$.

Arguments for replacing $\underline{u}$ or $E(U \mid S)$ with S were given. In the structural model, cases where $E(U \mid S)$ is linear in S were identified when the regression was linear or logistic, implying that no information is lost by replacing $E(U \mid S)$ with S in these cases.

In the structural normal theory linear model, the maximum likelihood estimator is a solution to the optimal estimating function. Thus, our approach extends instrumental variable estimation to generalized linear models in a natural and coherent fashion.

Table 4.1 Comparison of the naive, conditional and one-step estimators. Comparison is for data sets where the conditional estimator converged. Asterisks or diamonds between two values indicates the difference was significant at the 5% level, see the discussion in Section 4.6 for details.

Parameter		Naive		Conditional		One-step
			CORR(Y,U) = .15			
$\beta_0 = -6.939$	BIAS x10	7.18	*	-0.48	◇	-0.56
	MSE	0.95	*	0.82	*	0.83
$\beta_1 = 0.021$	BIAS x10^3	-6.40	*	0.52	◇	0.60
	MSE x10^5	5.99	*	4.79	*	4.86
$\beta_2 = 0.042$	BIAS x10^3	3.12	*	-0.60	◇	-0.64
	MSE x10^4	1.11		1.09		1.09
			CORR(Y,U) = .25			
$\beta_0 = -9.521$	BIAS x10	14.84	*	-1.26	*	-1.44
	MSE	2.78	*	1.29		1.28
$\beta_1 = 0.042$	BIAS x10^3	-13.30	*	0.78	*	0.93
	MSE x10^5	19.65	*	6.74		6.67
$\beta_2 = 0.032$	BIAS x10^3	7.30	*	0.21	◇	0.14
	MSE x10^4	1.62	*	1.22		1.22
			CORR(Y,U) = .35			
$\beta_0 = -13.095$	BIAS x10	28.43	*	-0.56	◇	-0.84
	MSE	8.78	*	1.83		1.83
$\beta_1 = 0.070$	BIAS x10^3	-23.55	*	0.64	◇	0.87
	MSE x10^5	57.85	*	9.96		9.96
$\beta_2 = 0.023$	BIAS x10^3	9.82	*	-0.84	◇	-0.94
	MSE x10^4	2.18	*	1.58		1.59

Table 4.2 Relative efficiency of the one-step estimator to the conditional estimator. Table entries are $\frac{\text{MSE(OS)}}{\text{MSE(S)}}$.

	CORR(Y,U)		
Parameter	.15	.25	.35
β_0	.987	1.007	1.000
β_1	.986	1.000	1.000
β_2	.999	1.000	.998

Table 4.3 Performance of the one-step estimator when the conditional estimator did not converge. Note that numerical values in the *Converged* column are from Table 4.1. An asterisk between two values indicates the difference was significant at the 5% level, see the discussion in Section 4.6 for details. CORR$(Y,U) = .15$.

Parameter		Converged (n=500)		Did not converge (n=118)
$\beta_0 = -6.939$	BIAS x10	-0.56		2.59
	MSE	0.83	*	1.17
$\beta_1 = 0.021$	BIAS x10^3	0.60	*	-2.38
	MSE x10^5	4.86		6.20
$\beta_2 = 0.042$	BIAS x10^3	-0.64		1.18
	MSE x10^4	1.09		1.41

APPENDIX

RESULT A.1. (see Rao, p.33) Let A be a non-singular matrix, and U and V column vectors. Then

$$(A + UV^T)^{-1} = A^{-1} - \frac{A^{-1}UV^TA^{-1}}{1 + V^TA^{-1}U}.$$

RESULT A.2. Let A and C be non-singular symmetric matrices. Then

$$\begin{pmatrix} A & B \\ B^T & C \end{pmatrix}^{-1} = \begin{pmatrix} W & -WBC^{-1} \\ -C^{-1}B^TW & C^{-1} + C^{-1}B^TWBC^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} A^{-1} + A^{-1}BVB^TA^{-1} & -A^{-1}BV \\ -VB^TA^{-1} & V \end{pmatrix}.$$

where $W = (A - BC^{-1}B^T)^{-1}$ and $V = (C - B^TA^{-1}B)^{-1}$, provided the indicated inverses exist.

RESULT A.3. *Let $\hat{\theta}_U$ be the unique maximum of $L_n(\theta)$. Let $\hat{\theta}_R$ maximize $L_n(\theta)$ subject to $\chi(\theta) = 0$. Assume*

$$Pr(\chi(\hat{\theta}_U) = 0) \longrightarrow 1,$$

$$(\hat{\theta}_U - \theta_0) = O_p(n^{-\frac{1}{2}})$$

and

$$(\hat{\theta}_R - \theta_0) = O_p(n^{-\frac{1}{2}}).$$

Then

$$\hat{\theta}_R - \hat{\theta}_U = o_p(n^{-\frac{1}{2}}).$$

PROOF: Define $\hat{\delta} = I(\chi(\hat{\theta}_U) = 0)$ where $I(\cdot)$ is the indicator function. Note that $\hat{\delta} \xrightarrow{P} 1$ and that $\hat{\theta}_R$ can be written

$$\hat{\theta}_R = \hat{\theta}_U \hat{\delta} + \hat{\theta}_R(1 - \hat{\delta}).$$

Then

$$(\hat{\theta}_R - \theta_0) - (\hat{\theta}_U - \theta_0) = (\hat{\theta}_U - \theta_0)(\hat{\delta} - 1) + (\hat{\theta}_R - \theta_0)(1 - \hat{\delta}).$$

An application of Slutsky's Theorem gives the result. •

BIBLIOGRAPHY

Amemiya, Y. (1985), "Instrumental variable estimator for the nonlinear errors-in-variables model," *J. Econometrics*, 28, 273–289.

Amemiya, Y. (1990a), "Two-stage instrumental variable estimator for the nonlinear errors-in-variables model," *J. Econometrics*, 44, 311–332.

Amemiya, Y. (1990b), "Instrumental variable estimation of the nonlinear measurement error model," *Statistical Analysis of Measurement Error Models and Applications*, 147–156, P. J. Brown and W. A. Fuller, eds., American Mathematical Society, Providence.

Andersen, P. K. and Gill, R. D. (1982), "Cox's regression model for counting processes: A large sample study," *Annals of Statistics*, 10, 1100-1120.

Balakrishnan, N. (editor), (1992), *Handbook of the Logistic Distribution,* New York: Marcel Dekker.

Billingsley, P. (1986), *Probability and Measure*, New York: John Wiley.

Bowden, R. J. and Turkington, D. A. (1984), *Instrumental Variables* , Cambridge: Cambridge University Press.

Burr, D. (1988), "On errors-in-variables in binary regression – Berkson case," *Journal of the American Statistical Association*, 83, 739-43.

Carroll, R. J. , Spiegelman, C. H., Lan, K. K., Bailey, K. T. & Abbott, R. D. (1984),"On errors-in-variables for binary regression models," *Biometrika*, 71, 19-25.

Carroll, R. J. and Ruppert, D. (1988), *Transformation and Weighting in Regression*, New York: Chapman and Hall.

Fuller, W. A. (1987), *Measurement Error Models*, New York: John Wiley.

Godambe, V. P. (1976), "Conditional likelihood and unconditional optimum estimating equations," *Biometrika*, 63, 277-284.

Gordon, T. and Kannel, W. E. (1968), *The Framingham Study*, introduction and general background in the Framingham study, §§ 1, 2. Bethesda, Maryland: National Heart, Lung, and Blood Institute.

Jennrich, R. I. (1969), "Asymptotic properties of nonlinear least squares estimation," *Annals of Mathematical Statistics*, 40, 633-643.

Judge, G. G. et. al. (1985), *The Theory and Practice of Econometrics*, New York: John Wiley.

Kendall, M. G. and Stuart, A. (1977), *The Advanced Theory of Statistics*, Vol. 1, 4th ed. New York: Hafner.

Kendall, M. G. and Stuart, A. (1979), *The Advanced Theory of Statistics*, Vol. 2, 4th ed. New York: Hafner.

Lindsay, B. G. (1982), "Conditional score functions: some optimality results," *Biometrika*, 69, 503-512.

Neymann, J. and Scott, E. L. (1948), "Consistent estimates based on partially consistent observations," *Econometrica*, 16, 1-32.

Rao, C. R. (1973), *Linear Statistical Inference and its Applications,* New York: John Wiley.

Rockafellar, R. T. (1970), *Convex Analysis*, Princeton: Princeton University Press.

Stefanski, L. A. (1985), "The effects of measurement error on parameter estimation," *Biometrika*, 74, 385-391.

Stefanski, L. A. (1990), "A normal scale mixture representation of the logistic distribution," *Statistics & Probability Letters*, 11, 69-70.

Stefanski, L. A. and Buzas, J. S. (1992), "Instrumental variable estimation in binary measurement error models," Institute of Statistics Mimeograph Series No. 2239, North Carolina State University.

Stefanski, L. A. and Carroll, R. J. (1985), "Covariate measurement error in logistic regression," *Annals of Statistics*, 13, 1335-51.

Stefanski, L. A. and Carroll, R. J. (1987), "Conditional scores and optimal scores for generalized linear measurement error models," *Biometrika*, 74, 703-716.

Stein, C. M. (1981), "Estimation of the mean of a multivariate normal distribution," *Annals of Statistics*, 9, 1135-51.

Wedderburn, R. W. M. (1976), "On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models," *Biometrika*, 63, 27- 32.

Whittemore, A. S. and Keller, J. B. (1988), " Approximations for regression with covariate measurement error," *Journal of the American Statistical Association*, 83, 1057-66.