

# Workshop: Topics in Applied Statistics

Cristobal Ridao-Cano  
Population Program  
Institute of Behavioral Science  
University of Colorado at Boulder

July 2000

## Abstract

This workshop focuses on a selected set of estimation issues that we generally encounter in the analysis of micro data. The topics that we will be addressing represent violations of one of the key assumption of the classical linear regression model (CLRM), namely, the independence between regressors and the error term. Contrary to other sources of specification error, the violation of the independence assumption results in biased and inconsistent OLS estimates, which has major implications for the interpretation of the estimation results.

The course will begin with a brief review of the CLRM assumptions, and the properties of OLS under those assumptions. Then I will introduce three sources of deviation from the independence assumption, namely, measurement error, omitted variables/endogenous regressors, and sample selection. Although I will be presenting some analytical results when necessary, the course will make special emphasis on examples and applications of the issues and techniques covered here. Finally, this course will mainly focused on the analysis of single equation models with cross-sectional data.

## Contents

1	The Classical Linear Regression Model . . . . .	3
1.1	Assumptions of the CLRM . . . . .	3
1.2	The OLS estimator in the CLRM . . . . .	4
1.2.1	Small (Finite) Sample Properties of $\hat{\beta}$ . . . . .	6
1.2.2	Large Sample (Asymptotic) Properties of $\hat{\beta}$ . . . . .	7
2	Measurement Error and Proxy Variables . . . . .	8
2.1	OLS Estimation with Badly Measured Independent Variables . . . . .	9
2.2	Instrumental Variable (IV) Estimation . . . . .	11
2.3	A Specification Test for Measurement Error . . . . .	15
3	Omitted Variables/Endogenous Regressors . . . . .	16
3.1	OLS Estimation with Omitted Relevant Variables . . . . .	17
3.2	Consistent Estimation and Specification Tests in the Presence of Omitted Variables/Endogenous Regressors . . . . .	21
4	Truncation, Censoring, and Sample Selection . . . . .	24
4.1	The Truncated Regression Model . . . . .	26
4.2	The Tobit (Censored Regression) Model . . . . .	28
4.3	Sample Selection Models . . . . .	33
4.3.1	The Heckman Model . . . . .	34
4.3.2	Program Evaluation . . . . .	38
4.3.3	Switching Models . . . . .	40

## 1. The Classical Linear Regression Model

Any empirical study in social sciences begins with a set of theoretical propositions about some aspect of the society or the economy. The theory specifies a set of relationships among variables, and guides the subsequent specification of the empirical model. The empirical investigation provides estimates of unknown parameters of the model, and attempts to measure the validity of the theoretical propositions against the behavior of the observable data.

In this context, a statistics textbook provides the researcher with a catalog of which estimators are most desirable in what estimating situations. This catalog is centered around a standard estimating situation referred to as the classical linear regression model (CLRM). It happens that in this standard situation the OLS estimator is considered the optimal estimator. The CLRM consists of a set of assumptions concerning the way in which the data are generated. Most estimation problems can be characterized as situations in which one (or more) of the CLRM assumptions is violated in a particular way. In many of these situations the OLS estimator is no longer considered to be the optimal estimator.

In this course we will focus on situations where the assumption of independence between regressors and the error term is violated, since this has major consequences on the properties of the OLS estimator, namely, biasedness and inconsistency, and thus the interpretation of the estimation results. How you ever think about these situations is a product of your knowledge of statistical theory (the textbook catalog), your theoretical framework, inspection of the data at hand, and your review of similar or related work.

### 1.1. Assumptions of the CLRM

The general form of the linear regression model is

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_K x_{iK} + \varepsilon_i, \quad i = 1, \dots, n \quad (1.1)$$

where  $y$  is the dependent variable;  $\{x_k\}$ ,  $k = 1, \dots, K$ , is the set of independent or explanatory variables;  $\varepsilon$  is the random component of the model or disturbance term; and  $i$  indexes  $n$  sample observations. Rewriting (1.1) in matrix notation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1.2)$$

where  $\mathbf{y} = (y_1, y_2, \dots, y_n)'$ ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ ,  $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_K)'$ , and  $\mathbf{X}$  is a  $n \times K$  matrix whose  $i$ th row is equal to  $\mathbf{x}'_i = (x_{i1}, x_{i2}, \dots, x_{iK})$ . In most contexts, the first column of  $\mathbf{X}$  is assumed to be a column of ones, so that  $\beta_1$  is the constant term of the model.

The basic assumptions of the CLRM are

1. *Functional form*:  $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ , i.e. linearity. Note that the linearity assumption refers to the manner in which the parameters enter the equation, not necessarily to the relationship between  $\mathbf{y}$  and  $\mathbf{X}$ . The estimation methods considered in this section are also applicable to a substantial variety of functional forms that can be transformed to linearity (e.g. log-linear).
2. *Zero mean of the error term*:  $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ , which implies that  $E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ .
3. *Spherical disturbances (i.e. homoscedasticity and nonautocorrelation)*:  $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'] = \sigma^2\mathbf{I}$ , which is equivalent to  $Var[\varepsilon_i] = \sigma^2$ , a constant for all  $i$ , and  $Cov[\varepsilon_i, \varepsilon_j] = 0$  for all  $i \neq j$ . Assumption (2) and (3) imply that the disturbances are independently and identically distributed, i.e.  $\varepsilon_i \sim i.i.d[0, \sigma^2]$ .
4. *Independence of regressors from the error term*:  $E[\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0}$ , which is equivalent to  $Cov[x_{ik}, \varepsilon_j] = 0$  for all sample observations  $i$  and  $j$ , and for all explanatory variables  $k$ . Note that since  $E[\boldsymbol{\varepsilon}] = \mathbf{0}$  then  $Cov(\mathbf{X}, \boldsymbol{\varepsilon}) = E[\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0}$ . Another way of stating this assumption is to say that the regressors in  $\mathbf{X}$  are exogenous (to  $\mathbf{y}$ ).
5.  $\mathbf{X}$  is a nonstochastic  $n \times K$  matrix with rank  $K$ . This means that the regressors are constant or fixed, and that there are no exact linear relationships among the regressors (i.e. no multicollinearity).

It is common to add to the above list the assumption that the disturbances are independently and identically normally distributed:  $\boldsymbol{\varepsilon} \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$ . The assumption of normality is not necessary to obtain the statistical results of the CLRM, but it is useful for making exact statements about the distribution of estimators and hypothesis testing procedures.

## 1.2. The OLS estimator in the CLRM

The OLS estimators of the regression coefficients  $\{\beta_k\}$ ,  $k = 1, \dots, K$ , in the multiple regression model (1.2) are defined as the values of  $\{\beta_k\}$  which minimize the sum of squared residuals

$$S(\boldsymbol{\beta}) = \sum_i (y_i - \boldsymbol{\beta}' \mathbf{x}_i)^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (1.3)$$

Expanding this, we have

$$S(\boldsymbol{\beta}) = \mathbf{y}' \mathbf{y} - \boldsymbol{\beta}' \mathbf{X}' \mathbf{y} - \mathbf{y}' \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\beta}' \mathbf{X}' \mathbf{X} \boldsymbol{\beta} \quad (1.4)$$

or

$$S(\boldsymbol{\beta}) = \mathbf{y}' \mathbf{y} - 2\boldsymbol{\beta}' \mathbf{X}' \mathbf{y} + \boldsymbol{\beta}' \mathbf{X}' \mathbf{X} \boldsymbol{\beta}$$

The necessary condition for a minimum is

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\mathbf{X}' \mathbf{y} + 2\mathbf{X}' \mathbf{X} \boldsymbol{\beta} = \mathbf{0} \quad (1.5)$$

Let  $\hat{\boldsymbol{\beta}}$  be the solution. Then  $\hat{\boldsymbol{\beta}}$  satisfies the normal equations <sup>1</sup>

$$\mathbf{X}' \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X}' \mathbf{y} \quad (1.6)$$

Assuming that the inverse of  $\mathbf{X}' \mathbf{X}$  exists, which follows from the full rank assumption,<sup>2</sup> we find that the solution to (1.6) is <sup>3</sup>

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \quad (1.7)$$

In the simple linear regression the corresponding solutions are

$$\hat{\alpha} = \bar{y} - \hat{\beta} \bar{x} \text{ and } \hat{\beta} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2} \quad (1.8)$$

---

<sup>1</sup>For the simple linear regression model,  $y_i = \alpha + \beta x_i + \varepsilon_i$ , the corresponding normal equations are

$$\sum_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0 \text{ and } \sum_i x_i (y_i - \hat{\alpha} - \hat{\beta} x_i) = 0$$

<sup>2</sup>In the simple linear regression model, the equivalent condition is that  $\sum_i (x_i - \bar{x})^2$  is positive.

<sup>3</sup>The assumption that  $\mathbf{X}$  has full rank guarantees that the second order condition for a minimum is indeed satisfied (Green, 1997).

### 1.2.1. Small (Finite) Sample Properties of $\hat{\beta}$

We shall show that the OLS estimator  $\hat{\beta}$  is the best linear unbiased estimator of  $\beta$  under the classical assumptions. Insert (1.2) in (1.7) to obtain

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon \quad (1.9)$$

If  $\mathbf{X}$  is fixed and  $E(\varepsilon) = 0$ , or if  $E(\mathbf{X}'\varepsilon) = 0$ , the expected value of the second term in (1.9) is zero, and so  $E(\hat{\beta}) = \beta$ . In other words, the OLS estimator  $\hat{\beta}$  is unbiased. The variance-covariance matrix is

$$\text{Var}(\hat{\beta}) = E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \quad (1.10)$$

Then, using (1.9),  $E(\hat{\beta}) = \beta$ , and  $E[\varepsilon\varepsilon'] = \sigma^2\mathbf{I}$ , we have <sup>4</sup>

$$\begin{aligned} \text{Var}(\hat{\beta}) &= E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon\varepsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \end{aligned} \quad (1.11)$$

Under the classical assumptions, the Gauss-Markov Theorem shows that  $\hat{\beta}$  is the best linear unbiased estimator of  $\beta$ , that is, the OLS estimator is the one with the smallest variance among the class of linear and unbiased estimators. Furthermore, if  $\varepsilon \sim N[\mathbf{0}, \sigma^2\mathbf{I}]$  then  $\hat{\beta} \sim N[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$  and the maximum likelihood estimator of  $\beta$  is equal to  $\hat{\beta}$ , and so it can be shown that  $\hat{\beta}$  is the best unbiased estimator (i.e. best among all unbiased estimators, not just linear unbiased estimators).

It is important to note that the assumption of nonstochastic regressors (i.e.  $\mathbf{X}$  is fixed in repeated samples) simplifies the derivation of the statistical properties of the OLS estimator. However, the important finite sample results (unbiasedness and the Gauss-Markov Theorem) do not depend on the assumption of nonstochastic regressors. The unbiasedness of OLS rests only on the assumption independence of regressors from the error term, i.e.  $E[\mathbf{X}'\varepsilon] = 0$ .

---

<sup>4</sup>In practice  $\sigma^2$  is unknown and thus replaced in (1.11) by an unbiased estimator,  $s^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-K} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-K}$ , where  $\hat{\varepsilon}$  are the OLS residuals (i.e. the OLS estimate of the error term  $\varepsilon$ ).

### 1.2.2. Large Sample (Asymptotic) Properties of $\hat{\beta}$

In this section we show the consistency (i.e. asymptotic unbiasedness), asymptotic normality, and asymptotic efficiency of  $\hat{\beta}$  under the classical assumptions. Multiply and divide (1.9) by  $n$  to obtain

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\varepsilon = \beta + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\varepsilon\right) \quad (1.12)$$

We now take the probability limit of (1.12)

$$\begin{aligned} p \lim \hat{\beta} &= p \lim \left[ \beta + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \left(\frac{1}{n}\mathbf{X}'\varepsilon\right) \right] \\ &= \beta + p \lim \left[ \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \left(\frac{1}{n}\mathbf{X}'\varepsilon\right) \right] \\ &= \beta + p \lim \left[ \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \right] p \lim \left[ \left(\frac{1}{n}\mathbf{X}'\varepsilon\right) \right] \end{aligned} \quad (1.13)$$

Now if  $\mathbf{X}$  is fixed and  $E(\varepsilon) = 0$ , or if  $E(\mathbf{X}'\varepsilon) = 0$  we have that

$$p \lim \left[ \left(\frac{1}{n}\mathbf{X}'\varepsilon\right) \right] = \mathbf{0} \quad (1.14)$$

Hence for  $p \lim \hat{\beta} = \beta$ , that is, for  $\hat{\beta}$  to be a consistent estimator of  $\beta$ , we need to further assume

$$p \lim \left[ \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \right] = \mathbf{Q} \quad (1.15)$$

where  $\mathbf{Q}$  is a positive definite matrix. Using the consistency of  $\hat{\beta}$  and the central limit theorem,  $\hat{\beta}$  can be shown to be asymptotically normally distributed

$$\hat{\beta} \overset{A}{\approx} N\left[\beta, \frac{\sigma^2}{n}\mathbf{Q}^{-1}\right] \quad (1.16)$$

where  $\frac{\sigma^2}{n}\mathbf{Q}^{-1}$  is the asymptotic variance-covariance matrix.<sup>5</sup> The important implication of (1.16) is that regardless of the finite sample distribution of  $\varepsilon$ , the conditions for  $\hat{\beta}$  to be a consistent estimator of  $\beta$  along with the central limit theorem guarantee that the limiting distribution of the OLS estimator is normal.<sup>6</sup>

<sup>5</sup>In practice, it is necessary to estimate  $(1/n)\mathbf{Q}^{-1}$  with  $(\mathbf{X}'\mathbf{X})^{-1}$  and  $\sigma^2$  with  $\hat{\varepsilon}'\hat{\varepsilon}/(n-K)$ .

<sup>6</sup>This is a useful result for hypothesis testing, since the three classical tests in statistics (Wald, Likelihood Ratio, and Lagrange Multiplier) are all based on asymptotics, and the t distribution approximates a normal in large samples.

As for the finite sample properties, the asymptotic properties of  $\hat{\beta}$  do not rely on the assumption of nonstochastic regressors. The key assumption for consistency is the uncorrelatedness between the error and the regressors.

It remains to be shown whether  $\hat{\beta}$  is asymptotically efficient, that is, whether its asymptotic variance-covariance matrix is no larger than the asymptotic variance-covariance matrix of any other consistent, asymptotically normally distributed estimator. If the errors are normally distributed,  $\hat{\beta}$  coincides with the maximum likelihood estimator (MLE) for  $\beta$ , and so it inherits all the desirable asymptotic properties of MLE, including asymptotic efficiency. However, if  $\varepsilon$  has a nonnormal distribution and it emerges that  $\hat{\beta}$  is not the MLE, it follows directly that  $\hat{\beta}$  is not efficient, although it is still consistent.

We have seen that the violation of the assumption of fixed regressors does not affect the biasedness or consistency properties of the OLS estimator as long as the regressors are distributed independently of the disturbances (i.e. orthogonal). However, if the regressors are correlated with the error term, the OLS estimator is biased even asymptotically (i.e. inconsistent). This is because the OLS procedure, in assigning 'credit' to regressors for explaining variation in the dependent variable, assigns, by mistake, some of the disturbance-generated variation of the dependent variable to the regressor with which that disturbance is correlated. All the sources of violation of the orthogonality assumption considered below can be examined within this framework.

## 2. Measurement Error and Proxy Variables

Thus far, it has been assumed (at least implicitly) that the data used to estimate the parameters of our models are true measurements on their theoretical counterparts. In practice, this only happens in the best of circumstances. In fact, many social scientists feel that the greatest drawback to statistics is the fact that the data with which they must work are so poor. A well-known quotation expressing this feeling is due to Josiah Stamp:

The Government are very keen on amassing statistics - they collect them, add them, raise them to the  $n$ th power, take the cube root and prepare wonderful diagrams. But you must never forget is that everyone of those figures comes in the first instance from the village watchman, who just puts down whatever he damn pleases. (1929, 00. 258-9)

Examples of poorly measured variables include, for example, wages (and income, in general). In addition, there are many situations in which the variable in a model has no observable counterpart and an observable proxy for this variable is used instead. Examples of the latter include, for example, education (an output) which is normally proxied by years of schooling (an input) as a determinant of earnings.

Errors in measuring the dependent variable are incorporated in the error term, leading to inefficient but still consistent estimates.<sup>7</sup> However, errors in measuring an independent variable make this variable stochastic, and, by construction, correlated with the disturbance term. As mentioned above, this causes biased and inconsistent OLS estimates. Hence, in this section, we examine OLS estimation with badly measured independent variables (the problem), instrumental variable (IV) estimation as an alternative estimation technique that yields consistent estimates (the solution), and the Hausman test as a general specification test of the orthogonality assumption (the detection).

## 2.1. OLS Estimation with Badly Measured Independent Variables

The following presentation will make use of asymptotic results for the classical regression model, since almost all of the results for the models with measurement error are asymptotic. By the same token, we assume that the sample moments such as  $\mathbf{X}'\boldsymbol{\varepsilon}/n$  converge in probability to population moments such as  $Cov[\mathbf{X}, \boldsymbol{\varepsilon}]$ , that is,  $p \lim[\mathbf{X}'\boldsymbol{\varepsilon}/n] = Cov[\mathbf{X}, \boldsymbol{\varepsilon}]$ .

To illustrate the problem, we shall use the standard example of estimating the effect of wages on hours of work (i.e. labor supply) using the sample of workers (i.e. those who participate in the labor market). For simplicity, assume that the only variable affecting hours of work  $h$  is wages  $w^t$ , where  $w^t$  represents true wages, and that the true model

$$h_i = \alpha + \beta w_i^t + \varepsilon_i \tag{2.1}$$

conforms to all the assumptions of the CLRM. However, the measure of wages that we have in the survey is that reported by the individuals ( $w^r$ ), which is an imprecise measure of  $w^t$ , that is,  $w^r$  measures  $w^t$  with some error (i.e. the measurement error). To this extent, assume that

---

<sup>7</sup>This is true for linear models. However, in the context of non-linear model (e.g. probit), measurement error in the dependent variable leads to inconsistent estimates (Hausman et al. 1998).

$$w_i^r = w_i^t + u_i \quad (2.2)$$

where  $u \sim i.i.d.(0, \sigma_u^2)$ , and  $w^t$ ,  $\varepsilon$ , and  $u$  are mutually independent. Hence the estimated model can be obtained by substituting (2.2) in (2.1)

$$h_i = \alpha + \beta w_i^t + \varepsilon_i = \alpha + \beta w_i^r + [\varepsilon_i - \beta u_i] = \alpha + \beta w_i^r + \nu_i \quad (2.3)$$

Since  $u$  appears both in  $w^r$  and  $\nu$ ,  $w^r$  and  $\nu$  are correlated

$$Cov[w_i^r, \nu_i] = Cov[w_i^t + u_i, \varepsilon_i - \beta u_i] = -\beta \sigma_u^2 \quad (2.4)$$

where the last equality follows from the assumption of mutual independence between  $w^t$ ,  $\varepsilon$ , and  $u$ . This violates the orthogonality assumption of the CLRM so we can expect the OLS estimator of  $\beta$ ,

$$\hat{\beta} = \frac{\sum_i (w_i^r - \bar{w}_i^r)(h_i - \bar{h})}{\sum_i (w_i^r - \bar{w}_i^r)^2} = \frac{s_{w^r h}}{s_{w^r}^2} \quad (2.5)$$

to be inconsistent. To find the probability limit of  $\hat{\beta}$ , assume that the sample moments  $s_{w^r h}$  and  $s_{w^r}^2$  converge to their respective population moments  $\sigma_{w^r h} = Cov(w^r, h)$  and  $\sigma_{w^r}^2 = Var(w^r)$ . Insert (2.1) and (2.2) in (2.5) to obtain

$$\begin{aligned} p \lim \hat{\beta} &= \frac{Cov(w^r, h)}{Var(w^r)} \\ &= \frac{Cov(w^t + u, \alpha + \beta w^t + \varepsilon)}{Var(w^t + u)} \\ &= \frac{Cov(w^t, \beta w^t)}{Var(w^t) + Var(u)} \\ &= \beta \left[ \frac{\sigma_{w^t}^2}{\sigma_{w^t}^2 + \sigma_u^2} \right] \end{aligned} \quad (2.6)$$

where  $\sigma_{w^t}^2 = Var(w^t)$ . As long as  $\sigma_u^2$  is positive, the term in brackets in (2.6) is less than one and so  $\hat{\beta}$  is inconsistent, with a persistent bias toward zero. Clearly, the greater the variability in the measurement error (i.e. the noise), the worse the bias. The effect of biasing the coefficient toward zero is called attenuation. In this context, the true (structural) effect of wages on hours of work,  $\beta$ , is not identified.

Within a multiple regression framework, matters only get worse. More precisely, a badly measured variable contaminates all of the OLS coefficient estimates not just that on the badly measured variable.

The measurement error discussed above is the so called classical measurement error. There are other types of measurement error that do not respond to the classic framework developed above (i.e. non-classical). Continuing with the example of the hours of work model, the reported wage is only available for those individuals that are working at the time of the survey.<sup>8</sup> Thus it is common practice to use the so called constructive wage  $w^c$  instead

$$w^c = \frac{\text{Earnings}}{\text{Hours of work}}$$

If hours of work is not measured with error then  $w^c$  would measure  $w^t$  without error since  $w^c = (w^t h)/h = w^t$ . However, if hours of work is measured with error,  $h^r = h^t + \eta$ ,  $w^c$  would be measured with error as well, which would result in spurious correlation between  $h$  and  $w^c$ , thus leading to an inconsistent OLS estimate of  $\beta$ . More precisely, an increase in  $\eta$  leads to an increase in  $h$ , which in turn reduces  $w^c$ , creating a negative spurious correlation between  $h$  and  $w^c$ . As a result of that,  $\hat{\beta}$  is biased downward with respect to  $\beta$  (i.e.  $p \lim \hat{\beta} < \beta$ ). Note that we said earlier that measurement error in the dependent variable does not cause inconsistency, but in this case it is the combination of measurement error in  $h$  and the presence of  $h$  in  $w^c$  what causes inconsistency.

## 2.2. Instrumental Variable (IV) Estimation

An alternative estimation method in the context of errors in variables (and numerous others such as endogenous regressors) is instrumental variables (IV). We shall consider some general results in the context of a multiple regression where only one regressor is correlated with the error term,<sup>9</sup> and then we will apply those results to our hours of work model with measurement error. Suppose the model to be estimated is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2.7}$$

---

<sup>8</sup>This example will actually be useful in distinguishing instruments used to correct for measurement error bias from those used to correct for endogeneity bias.

<sup>9</sup>These results can be easily generalized to the case where we have several 'problem' variables.

where  $\mathbf{X} = [\mathbf{X}_1 \mathbf{x}_2]$  where  $\mathbf{X}_1$  is matrix of regressors (including a column of ones) such that

$$p \lim[(1/n)\mathbf{X}'_1 \boldsymbol{\varepsilon}] = \mathbf{0}$$

and  $\mathbf{x}_2$  is a regressor such that

$$p \lim[(1/n)\mathbf{x}'_2 \boldsymbol{\varepsilon}] \neq 0$$

so that the OLS estimator of  $\boldsymbol{\beta}$  is not consistent. Suppose, however, that there exists an instrument for  $\mathbf{x}_2$ ,  $\mathbf{z}$ , such that  $\mathbf{z}$  is correlated with  $\mathbf{x}_2$  but not with  $\boldsymbol{\varepsilon}$ , that is,

$$p \lim \frac{\mathbf{z}' \boldsymbol{\varepsilon}}{n} = 0 \text{ and } p \lim \frac{\mathbf{z}' \mathbf{x}_2}{n} \neq 0 \quad (2.8)$$

Defining the matrix  $\mathbf{Z} = [\mathbf{X}_1 \mathbf{z}]$ , we have

$$p \lim[(1/n)\mathbf{Z}' \boldsymbol{\varepsilon}] = \mathbf{0} \text{ and } p \lim[(1/n)\mathbf{Z}' \mathbf{X}] \neq \mathbf{0}$$

Then the IV estimator is

$$\widehat{\boldsymbol{\beta}}_{IV} = [\mathbf{Z}' \mathbf{X}]^{-1} \mathbf{Z}' \mathbf{y} \quad (2.9)$$

which is a consistent estimator of  $\boldsymbol{\beta}$  since <sup>10</sup>

$$\begin{aligned} p \lim \widehat{\boldsymbol{\beta}}_{IV} &= p \lim \left[ \frac{1}{n} \mathbf{Z}' \mathbf{X} \right]^{-1} p \lim \left[ \frac{1}{n} \mathbf{Z}' (\mathbf{X} \boldsymbol{\beta} + \boldsymbol{\varepsilon}) \right] \\ &= \boldsymbol{\beta} + p \lim \left[ \frac{1}{n} \mathbf{Z}' \mathbf{X} \right]^{-1} p \lim \left[ \frac{1}{n} \mathbf{Z}' \boldsymbol{\varepsilon} \right] \\ &= \boldsymbol{\beta} \end{aligned}$$

The asymptotic covariance matrix for  $\widehat{\boldsymbol{\beta}}_{IV}$  is

$$Asy.Var[\widehat{\boldsymbol{\beta}}_{IV}] = \sigma^2 [\mathbf{Z}' \mathbf{X}]^{-1} [\mathbf{Z}' \mathbf{Z}] [\mathbf{X}' \mathbf{Z}]^{-1} \quad (2.10)$$

An equivalent approach is the so called two-stages least squares (2SLS) which consists of regressing  $\mathbf{x}_2$  on  $\mathbf{Z}$  (first stage regression)

---

<sup>10</sup>IV estimation guarantees consistency in linear models, but not in the context of non-linear model (e.g. probit, tobit). In the section dedicated to the tobit model, I will present a simple method for obtaining consistent estimators and a test for the orthogonality assumption.

$$\mathbf{x}_2 = \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u} \quad (2.11)$$

and then replace  $\mathbf{x}_2$  in (2.7) with  $\mathbf{x}_2 = \widehat{\mathbf{x}}_2 + \widehat{\mathbf{u}}$ , where  $\widehat{\mathbf{x}}_2 = \mathbf{Z}\widehat{\boldsymbol{\gamma}}$ ,

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= [\widehat{\mathbf{X}} + \widehat{\mathbf{u}}]\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ &= \widehat{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} + \boldsymbol{\beta}\widehat{\mathbf{u}} \\ &= \widehat{\mathbf{X}}\boldsymbol{\beta} + \boldsymbol{\omega} \end{aligned} \quad (2.12)$$

where  $\widehat{\mathbf{X}} = [\mathbf{X}_1 \widehat{\mathbf{x}}_2]$ , and apply OLS (second stage regression). Then the 2SLS estimator is

$$\widehat{\boldsymbol{\beta}}_{2SLS} = [\widehat{\mathbf{X}}' \widehat{\mathbf{X}}]^{-1} \widehat{\mathbf{X}}' \mathbf{y} \quad (2.13)$$

Straightforward algebra shows  $\widehat{\boldsymbol{\beta}}_{2SLS} = \widehat{\boldsymbol{\beta}}_{IV}$ . The instrument  $\mathbf{z}$  is also called excluded variable since it affects  $\mathbf{y}$  only through  $\mathbf{x}_2$  but not directly, since it is uncorrelated with  $\boldsymbol{\varepsilon}$ . We use this exclusionary restriction to identify the true or structural effect of  $\mathbf{x}_2$  on  $\mathbf{y}$ . The inclusion of  $\mathbf{X}_1$  as instruments in  $\mathbf{Z}$  guarantees the consistency of  $\widehat{\boldsymbol{\beta}}$ , since, by construction of OLS,  $\mathbf{X}_1$  is uncorrelated with  $\widehat{\mathbf{u}}$ , and by assumption  $\mathbf{X}_1$  is uncorrelated with  $\boldsymbol{\varepsilon}$ , so  $\mathbf{X}_1$  is uncorrelated with  $\boldsymbol{\omega}$ .

One word of caution about instruments. The asymptotic variance of  $\widehat{\boldsymbol{\beta}}_{IV}$  can be rather large if  $\mathbf{z}$  is weakly correlated with  $\mathbf{x}_2$  (i.e.  $\mathbf{z}$  is a weak instrument for  $\mathbf{x}_2$ ). Furthermore, Bound et al. (1995) show that there are serious problems with IV estimation when the instruments are weak. First, they show that even if there is a tiny problem of correlation between  $\mathbf{z}$  and  $\boldsymbol{\varepsilon}$ , the consequence of this correlation on consistency gets exacerbated if  $\mathbf{z}$  is a weak instrument for  $\mathbf{x}_2$ . Second, they show that even if  $\widehat{\boldsymbol{\beta}}_{IV}$  is consistent it may not be unbiased, and this finite sample bias gets larger the weaker the instrument, and hence  $\widehat{\boldsymbol{\beta}}_{IV}$  closer to  $\widehat{\boldsymbol{\beta}}_{OLS}$ .

In view of these problems, they recommend (and I actually recommend it to you!) to check the joint significance of the instruments in the first stage regression (and, of course, report the first stage regression results) and make sure that you get a  $F$  statistic greater than 10 (as a rule of thumb). They also recommend to do a  $F$  test for overidentification, that is, a test of joint significance of the instruments when we extend the model in (2.12) to include these instruments. Note that the latter is basically a test of  $Cov(\mathbf{z}, \boldsymbol{\varepsilon}) = 0$ , and hence the result of the test should be that the coefficients on the instruments should not be (statistically) different from zero. You would be surprised of how few studies actually do these checks!

The difficult task in applying IV estimation is to find suitable instruments as defined above. In the context of measurement error problems, the choice is rather limited.<sup>11</sup> If you have longitudinal data, you can use lagged values of the badly measured regressor as an instrument. In a pure cross-section, a few instruments have been devised based only on the data in hand (e.g. Green, 1997). Other than that, the choice of the instrument very much depends on the context and the information you have, and it is a real art!. For example, Ashenfelter and Krueger (1994) analyze the returns to schooling in a sample of twins and make use of sibling-reported schooling as an instrument for self-reported schooling. In the context of the hours of work model developed above, the model of interest is

$$h_i = \alpha + \beta w_i^t + \boldsymbol{\delta}' \mathbf{x}_i + \varepsilon_i \quad (2.14)$$

where  $\mathbf{x}$  represents other explanatory variables such as non-labor income and demographic shifters. The measure of wages available is  $w^c$  which measures  $w^t$  with error since  $h$  is measured with error (i.e. non-classical measurement error)

$$w^c = w^t + \nu \quad (2.15)$$

where  $w^t$ ,  $\nu$ , and  $\varepsilon$  are mutually independent. The model to be estimated is then

$$h_i = \alpha + \beta w_i^c + \boldsymbol{\delta}' \mathbf{x}_i + [\varepsilon_i - \beta \nu_i] \quad (2.16)$$

Now, abstracting from sample selection problems, we can focus on the sample of individuals who are currently working and use reported wages  $w^r$  as an instrument for  $w^c$ . More precisely, let

$$w^r = w^t + u \quad (2.17)$$

where  $u$  is uncorrelated with  $w^t$ ,  $\nu$ , and  $\varepsilon$ . Hence,  $w^c$  and  $w^r$  are only correlated with  $w^t$  (the signal), which is uncorrelated with  $\varepsilon$  by assumption. First, regress  $w^c$  on  $w^r$  and  $\mathbf{x}$  (first stage regression)

$$w_i^c = \gamma_0 + \gamma_1 w_i^r + \boldsymbol{\gamma}'_3 \mathbf{x}_i + \omega_i \quad (2.18)$$

---

<sup>11</sup>The following discussion focuses on instruments used to correct for measurement error only. In practical applications, as we will see, we often find variables that are measured with error and are endogenous. In these situations it is customary to use instruments that correct for both sources of specification error.

Next, replace  $w^c$  with  $w^c = \hat{w}^c + \hat{\omega}$  in (2.16) to obtain

$$h_i = \alpha + \beta \hat{w}_i^c + \boldsymbol{\delta}' \mathbf{x}_i + [\varepsilon_i - \beta \hat{\omega}_i] \quad (2.19)$$

Finally, apply OLS on (2.19), the second stage regression. Since  $\hat{w}^c$  only inflicts variation in  $w^t$ , which is uncorrelated with  $[\varepsilon_i - \beta \hat{\omega}_i]$ , the OLS estimate of  $\beta$  is consistent.

### 2.3. A Specification Test for Measurement Error

Hausman (1978) has devised a general test of the orthogonality assumption. Following the previous general development of IV estimation, the system of hypothesis is

$$\begin{aligned} H0 & : \text{Cov}[\mathbf{x}_2, \boldsymbol{\varepsilon}] = 0 \\ H1 & : \text{Cov}[\mathbf{x}_2, \boldsymbol{\varepsilon}] \neq 0 \end{aligned} \quad (2.20)$$

Under  $H0$ , both  $\hat{\boldsymbol{\beta}}_{IV}$  and  $\hat{\boldsymbol{\beta}}_{OLS}$  are consistent estimators of  $\boldsymbol{\beta}$ , although  $\hat{\boldsymbol{\beta}}_{OLS}$  is efficient, while  $\hat{\boldsymbol{\beta}}_{IV}$  is inefficient.<sup>12</sup> However, if  $H0$  is false, only  $\hat{\boldsymbol{\beta}}_{IV}$  is consistent. The test then examines the difference between  $\hat{\boldsymbol{\beta}}_{OLS}$  and  $\hat{\boldsymbol{\beta}}_{IV}$ . Under  $H0$ ,  $p \lim[\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}] = \mathbf{0}$ , while if  $H0$  is false  $p \lim[\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}] \neq \mathbf{0}$ . The Hausman test is a Wald test based on the asymptotic distribution of  $[\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}]$  under  $H0$

$$[\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}] \overset{A}{\sim} N(0, V[\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}]) \quad (2.21)$$

Using  $V[\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}] = V[\hat{\boldsymbol{\beta}}_{OLS}] - V[\hat{\boldsymbol{\beta}}_{IV}]$ , the Hausman statistic is the quadratic form

$$H = [\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}]' [V[\hat{\boldsymbol{\beta}}_{OLS}] - V[\hat{\boldsymbol{\beta}}_{IV}]]^{-1} [\hat{\boldsymbol{\beta}}_{OLS} - \hat{\boldsymbol{\beta}}_{IV}] \sim \chi_{[K]}^2 \quad (2.22)$$

As Hausman shows, this may be referred to a chi-squared table. In the context of a simple regression model the statistic is then

---

<sup>12</sup>This is a general result when using 'irrelevant' information for estimation. In this case, since  $H0$  is false then there is no need to instrument for  $\mathbf{x}_2$ , yet we use  $\hat{\mathbf{x}}_2$  instead of  $\mathbf{x}_2$  which leads to imprecise (inefficient) estimates.

$$H = \frac{[\widehat{\beta}_{OLS} - \widehat{\beta}_{IV}]^2}{\text{Var}[\widehat{\beta}_{OLS}] - \text{Var}[\widehat{\beta}_{IV}]} \sim \chi^2_{[1]}$$

In the context of the hours of work model developed above, Mroz (1987) makes use of the Hausman test to show evidence of a strong downward bias caused by measurement error. The coefficient estimate on wages actually goes from negative to positive after controlling for measurement error using the IV methodology described above. Ashenfelter and Krueger (1994) show that after controlling for unobserved ability, further controlling for measurement error in schooling -using sibling-reported schooling as an instrument for self-reported schooling- increases the estimated returns to schooling from 0.11 to 0.13. As you can see, measurement error can be a serious problem with significant implications for the interpretation of the results.

Other applied work dealing with measurement error problems include, for example, Bound and Krueger (1991), Altonji (1986), Lam and Schoeni (1993), Blackburn and Neumark (1994).

### 3. Omitted Variables/Endogenous Regressors

Our analysis has been based on the assumption that the correct specification of the regression model is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where the matrix  $\mathbf{X}$  captures observable (to the researcher) and specified (by the researcher) individual heterogeneity (i.e. characteristics) and  $\boldsymbol{\varepsilon}$  represents the 'trash' where observable but unspecified individual characteristics and unobserved heterogeneity live in. Now,  $\mathbf{X}$  may contain variables that are 'irrelevant' in explaining variation in  $\mathbf{y}$ , but this only leads to less efficient estimates with no further consequences. Likewise, we should not worry too much about the heterogeneity in  $\boldsymbol{\varepsilon}$  as long as it not correlated with any variable in  $\mathbf{X}$ , but obviously if there is something relevant in explaining  $\mathbf{y}$  that can be measured but we have not specified in  $\mathbf{X}$ , it makes sense to include it even if it is not correlated with any variable in  $\mathbf{X}$ .

The problem arises when there is something in  $\boldsymbol{\varepsilon}$  that is relevant in explaining  $\mathbf{y}$  and correlated with one (or more) regressor in  $\mathbf{X}$ , because in this case the OLS

estimator of  $\beta$  is biased even asymptotically. The reason again is that the OLS procedure, in assigning 'credit' to regressors for explaining variation in the dependent variable, assigns, by mistake, some of the disturbance-generated variation of the dependent variable to the regressor with which that disturbance is correlated. The direction of the bias depends on both the sign of the correlation between the unspecified characteristic and  $\mathbf{y}$ , and the sign of the correlation between the unspecified characteristic and the specified regressor.

This problem receives a variety of names depending on the context: omitted variable bias, endogeneity bias, heterogeneity bias, simultaneity bias. The solution to this problem varies depending on whether the omitted characteristic is observable or not. If the variable can be measured then you should include it (even if it is measured with error). However, if the unspecified characteristic is not observable then the only thing you can do is to control for it using IV estimation.

In this section, we first examine OLS estimation with omitted relevant variables, and illustrate the problem with several examples. We then consider consistent estimation and specification tests in the presence of omitted variables/endogenous regressors.

### 3.1. OLS Estimation with Omitted Relevant Variables

Suppose that the true model is

$$\mathbf{y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \varepsilon \quad (3.1)$$

where all the assumptions of the CLRM are satisfied. The original  $\mathbf{X}$  has been decomposed into two submatrices  $\mathbf{X}_1$  and  $\mathbf{X}_2$  of  $K_1$  and  $K_2$  columns (variables), respectively. Now if we regress  $\mathbf{y}$  on  $\mathbf{X}_1$  without including  $\mathbf{X}_2$  we have

$$\mathbf{y} = \mathbf{X}_1\beta_1 + [\mathbf{X}_2\beta_2 + \varepsilon] = \mathbf{X}_1\beta_1 + \omega \quad (3.2)$$

where  $\omega$  is now the error term in this misspecified model which includes the unspecified  $\mathbf{X}_2$ . Since

$$Cov[\mathbf{X}_1, \omega] = Cov[\mathbf{X}_1, \mathbf{X}_2\beta_2 + \varepsilon] = \mathbf{X}_1'\mathbf{X}_2\beta_2 \quad (3.3)$$

unless  $\mathbf{X}_1'\mathbf{X}_2 = \mathbf{0}$  (i.e.  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are uncorrelated), this violates the orthogonality assumption of the CLRM, and hence we can expect  $\hat{\beta}_1$  based on (3.2) to be biased and inconsistent.

More precisely,

$$\begin{aligned}
\hat{\beta}_1 &= [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_1 \mathbf{y} \\
&= [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_1 [\mathbf{X}_1 \beta_1 + \mathbf{X}_2 \beta_2 + \boldsymbol{\varepsilon}] \\
&= \beta_1 + [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_1 \mathbf{X}_2 \beta_2 + [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_1 \boldsymbol{\varepsilon}
\end{aligned} \tag{3.4}$$

Taking the expectation, we see that unless  $\mathbf{X}'_1 \mathbf{X}_2 = \mathbf{0}$ ,  $\hat{\beta}_1$  is biased:

$$E[\hat{\beta}_1] = \beta_1 + \gamma \beta_2 \tag{3.5}$$

where  $\gamma$  is the parameter vector corresponding to the regression of the corresponding column of  $\mathbf{X}_2$  on the columns of  $\mathbf{X}_1$  where

$$\hat{\gamma} = [\mathbf{X}'_1 \mathbf{X}_1]^{-1} \mathbf{X}'_1 \mathbf{X}_2 \tag{3.6}$$

It is clear from above that for  $\hat{\beta}_1$  to be a biased estimator of  $\beta_1$  we need both  $\gamma \neq \mathbf{0}$  and  $\beta_2 \neq \mathbf{0}$ , that is, the variables in  $\mathbf{X}_2$  must be relevant in explaining  $\mathbf{y}$ , and  $\mathbf{X}_1$  and  $\mathbf{X}_2$  must be correlated. The direction of the bias depends on  $\gamma$  and  $\beta_2$  and it is easy to deduce when there is a single included variable and one omitted variable. However, just like in the measurement error problem, if more than one variable is included, the terms in (3.5) involve multiple regression coefficients, which themselves have the sign of partial, not simple, correlations.

If you look carefully enough, you would realize that the measurement error problem can be viewed as an omitted variable problem. To illustrate this point, we can write the estimated hours of work model

$$h_i = \alpha + \beta w_i^c + \boldsymbol{\delta}' \mathbf{x}_i + [\varepsilon_i - \beta \nu_i] \tag{3.7}$$

in a slightly different form as

$$h_i = \alpha + \beta w_i^c + \boldsymbol{\delta}' \mathbf{x}_i + \theta \nu_i + \varepsilon_i \tag{3.8}$$

where  $\theta = -\beta$ . If we could observe  $\nu$ , applying OLS on (3.8) would produce unbiased estimates.<sup>13</sup> In a model that omits  $\nu$ , the effect of variation of  $\nu$  on  $y$ ,  $-\beta$ , must be transmitted through variation in  $w^c$  (abstracting from  $\mathbf{x}$ ), and so  $\hat{\beta}$  in this model is a mixture of the coefficient on  $w^c$ , that is,  $\beta$ , and the coefficient on  $\nu$ , that is,  $-\beta$ .

---

<sup>13</sup>They would not be efficient, as they neglect the constraint  $\theta = -\beta$ .

Continuing with the hours of work model, we have so far assumed that the true wage  $w^t$  is uncorrelated with  $\varepsilon$ . However, there are unobserved individual characteristics that could potentially be important in explaining  $h$  and are correlated with  $w^t$ . One of these characteristics is motivation: more motivated individuals earn on average higher wages, and, given the same wage, they work harder on average than less motivated individuals. Then we can define a new error term  $\varepsilon' = m + \varepsilon$ , where  $m$  represents motivation. In this case, even after controlling for measurement error using  $w^r$  as an instrument for  $w^c$ , the OLS estimator of  $\beta$  would be biased. Again, this is because  $w^t$  is correlated with  $\varepsilon'$  through  $m$ . We can even predict that the bias is likely to be positive, that is, the OLS estimate of  $\beta$  reflects both the positive effect of motivation on hours of work through higher wages and the true (structural) effect of  $w^t$  on  $h$  (i.e.  $\beta$  itself) which is not identified.

Take another popular example, the earnings equation. Here we are mainly interested in the effect of education on earnings (i.e. returns to education). We mentioned in the measurement error section that since education is not observed, we normally use years of schooling as a proxy, which is measured with error. Suppose that we regress earnings ( $y$ ) on years of schooling ( $s$ ) only

$$y_i = \alpha + \beta s_i + \varepsilon_i \quad (3.9)$$

where  $\varepsilon$  contains, among other things, the schooling measurement error. Suppose we have other source of data about each individual's schooling level and use it as an instrument for schooling to correct for measurement error. Again, this would produce consistent estimates of the return to schooling ( $\beta$ ) as long as education is not correlated with some relevant and unspecified (observable or unobservable) individual characteristic.

Take the example of some observable characteristic such as labor market experience: one can easily argue that given the same level of education, those with more experience earn more money on average (if experience is rewarded in the labor market), but, at the same time, those with lower levels of schooling have on average more years of experience than those with higher levels of education (simply because they have been out of school and in the labor market for longer). Thus if we do not include experience in the model, we can predict that the estimated return to education will be biased downward.

Now take the example of a characteristic that is (pretty much) unobservable such as innate ability: given the same level of education, higher ability individuals earn more on average than lower ability individuals (simply because they are

more able), and also choose (i.e. self-select) and achieve higher levels of education because they are more able and have higher expected returns to education. This creates the so called omitted ability bias in estimating the returns to education, a bias that is likely to be positive since education is partly stealing the effect of ability (i.e. taking credit for it).

In all the above examples, there is some unspecified characteristic that makes some specified regressor endogenous. Literally speaking, the presence of measurement error also makes the variable that is measured with error endogenous, although we can consider this endogeneity as more of a 'spurious' type.

The so called simultaneity bias problem generally arises in the context of simultaneous system of equations, where a typical equation of this system has at least one endogenous variable as an independent variable. In this case, endogeneity takes its full meaning: the endogenous regressors has an equation on its own in the system, that is, this regressor is endogenously determined within the system. Here the theory tells us explicitly that this regressor is endogenous, but we are still interested in examining the so called structural effect of this regressors on the variable in question.<sup>14</sup>

Although simultaneous equation system are not covered here per se, we can use this tool to get a better understanding of the endogeneity problem in our simple one equation model. There are many situations where, based on our theoretical framework, we believe that, say, one regressor is jointly determined with the dependent variable in question, but we are only interested in examining the structural effect of this regressors on the dependent variable (and hence not interested per se in modelling this regressor explicitly). If we do not control for the endogeneity of this regressors our parameter estimates (particularly that on the endogenous regressor) are biased and inconsistent.

Examples of the latter flourish in the literature. For example, in examining female labor supply we may be interested in the effect of some fertility outcome, but we know that fertility and labor supply decisions are jointly determined. Using Nizam's work, health expenditures allocated to a household member is a function of the economic contribution of this member, but the productivity of this member depends among other things on his or her health, which is a function of, among other things, health expenditures. Using my own work, the time allocation of school-age children in rural Bangladesh is a function of the number of pre-school

---

<sup>14</sup>Note that this information is lost when we estimate the so called reduced form version of the model, that is, a model where each endogenous variable is a function of exogenous variables only.

children (Ridao-Cano, 2000). However, fertility and time allocations decisions are partly determined by the economic contribution of children, which is a function of observed and unobserved characteristics.

Assuming that have gotten something out of this workshop so far, whether you ever even think about potential endogeneity in your model depends on your theoretical framework, and how well you have done your homework of reviewing similar or related work. Remember one more thing: endogeneity is at the end of the day an empirical matter, and hence a matter of degree (statistical degree), that is, you may suspect some regressor is endogenous but your test may show otherwise, which does not imply that the regressor is not endogenous but simply that you should not worry about it as far as your estimation results are concerned.

### 3.2. Consistent Estimation and Specification Tests in the Presence of Omitted Variables/Endogenous Regressors

Let us start with the simple case. If some measurable relevant variable is omitted from the analysis, and the appropriate test indicates that it should, then include it, and end of the story!. So what test should we use ?. An already familiar candidate is the Hausman test. Recall the model in (3.1)

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon} \quad (3.10)$$

Now consider the system of hypothesis in play

$$H0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

$$H1 : \boldsymbol{\beta}_2 \neq \mathbf{0}$$

Let  $\hat{\boldsymbol{\beta}}_1$  be the OLS estimator of  $\boldsymbol{\beta}_1$  under  $H0$  (constrained model), and  $\tilde{\boldsymbol{\beta}}_1$  under  $H1$  (unconstrained model). If  $H0$  is true, then  $\hat{\boldsymbol{\beta}}_1$  and  $\tilde{\boldsymbol{\beta}}_1$  are both unbiased and consistent, although  $\hat{\boldsymbol{\beta}}_1$  is efficient and  $\tilde{\boldsymbol{\beta}}_1$  is not (since it is based on irrelevant information). If  $H0$  is false then only  $\tilde{\boldsymbol{\beta}}_1$  is unbiased and consistent. The test statistic under  $H0$  is

$$H = [\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1]' [V[\hat{\boldsymbol{\beta}}_1] - V[\tilde{\boldsymbol{\beta}}_1]]^{-1} [\hat{\boldsymbol{\beta}}_1 - \tilde{\boldsymbol{\beta}}_1] \sim \chi^2_{[p_1]} \quad (3.11)$$

where  $p_1$  is the number of regressors in the restricted model, that is, the number of columns in  $\mathbf{X}_1$ . An alternative test is the Wald test, which focuses on  $\boldsymbol{\beta}_2$  instead of  $\boldsymbol{\beta}_1$

$$W = [\tilde{\beta}_2]' [V[\tilde{\beta}_2]]^{-1} [\tilde{\beta}_2] \sim \chi_{[p_2]}^2 \quad (3.12)$$

where  $p_2$  is the number rows (parameters) in  $\beta_2$ . Some people argue that the Hausman test is less powerful than the Wald test when  $\mathbf{X}'_1 \mathbf{X}_2 \rightarrow \mathbf{0}$ .

Let us now consider the case where the omitted characteristics -with which some specified regressor is correlated- is not observable (again, to the researcher). In this case, the instrumental variables method is again the right way to go in the context of linear models.<sup>15</sup> Alternatively, if you have longitudinal data and are ready to assume that the individual unobserved characteristics are fixed over time, then you can get rid of those by first differencing or applying the fixed effects methodology. The same approach can be used if you have data on identical twins, since by looking at within twin variation in, say, earnings we get rid of their common unobserved characteristics such as ability.

Since IV estimation and the Hausman test have already been covered in the measurement error section, I will just focus here on the difficult task of finding valid instruments, the potentially difficulty of distinguishing between measurement error and endogeneity using the Hausman test, and possible interaction between measurement error and endogeneity.

In the context of the hours of work model, several variables have been used as instruments for wage, including human capital variables (e.g. education, age), family background variables (e.g. parental education), higher moments of human capital and family background variables (i.e. squares, cubes), and demand variables (e.g. industry and occupation specific unemployment rates).<sup>16</sup>

Mroz (1987) offers a good example of how to search for the best set of instruments. Mroz also notes that these instruments can be used to correct for both measurement error and endogeneity, and then the Hausman test cannot distinguish the source of misspecification. In fact, he makes use of the Hausman test to show that the hypothesis of exogeneity cannot be rejected. He then hypothesize that this due to the opposing forces of the two biases, which end up cancelling each other out. Recall that the non-classical measurement error in  $w^c$  produces downward bias, whereas the correlation between wages and motivation produces

---

<sup>15</sup>As mentioned earlier, IV estimation guarantees consistency in linear models, but not in the context of non-linear model (e.g. probit, tobit). In the section dedicated to the tobit model, I will present a simple method for obtaining consistent estimators and a exogeneity test.

<sup>16</sup>Note that if we use higher moments of included variables as instruments, identification relies only on the non-linear way in which these variables enter the first stage regression. In this case, several authors have shown that the results are normally unstable, so try to avoid that.

an upward bias. He tests this hypothesis by first examining measurement error only, using  $w^r$  as an instrument for  $w^c$ , and indeed finds evidence of a downward bias. He then further corrects for endogeneity using extra instruments, and indeed finds evidence of an upward bias. The moral of the story: two wrongs make a right, so little OLS is still OK!

In the context of a dynamic hours of work model, MaCurdy (1981) and Altonji (1986), show how the method of first differencing to get rid of time invariant unobserved heterogeneity may indeed exacerbate the measurement error problem, so we need to control for that to get consistent estimates. With the classical measurement error, this occurs when the measurement error is time variant and  $w^t$  is autocorrelated.

Now let us retake the earnings equation. Some people have tried to proxy directly for ability. Lam and Schoeni (1993) use family background variables as proxies for a man's ability, including wife's education. Behrman et al. (1995), in a similar approach, use information on marriage market outcomes (e.g. wife's characteristics) to obtain estimates of the men's unobserved human capital. Blackburn and Neumark (1994) use test scores (AFQT) as a proxy for ability, allowing for measurement error in these test scores. These pieces of work find substantial upward biases in the OLS estimate of the returns to education even after controlling for measurement error as well. In addition, Lam and Schoeni show that adding family background variables increases the measurement error bias for a given measurement error in schooling, since these variables are correlated with schooling.

Continuing with the earnings equation, other researchers have instrumented for schooling using, for example, family background variables. However, in one of the most influential papers in the literature, Angrist and Krueger (1991) came up with a true source of exogenous variation in schooling, namely, quarter of birth. Under the compulsory schooling legislation in the US children born in the first quarter are eligible to quit school earlier because they start schooling at an older relative age, thus completing less schooling. Cool instrument as it looks, it does not in fact satisfy one of the desirable properties of an instrument, namely, to be highly correlated with schooling, which leads to the sort of problems outlined in the measurement error section (see Bound et al., 1995).

Finally, Ashenfelter and Krueger (1994) make use of a sample of identical twins to estimate the economic returns to schooling, adjusting for unobserved ability and measurement error in schooling. By looking at within twins differences in earnings, they are able to get rid of their common unobserved ability but at

the expense of increasing the measurement error bias. After instrumenting for self-reported schooling using sibling reported schooling, the estimated returns to schooling lies between 12-16%, much higher than previously found. The results of this paper are in sharp contrast with the rest of the literature (e.g. Behrman et al., 1980), since the authors do not find evidence of an upward omitted ability bias, and find instead a sharp downward bias caused by measurement error.

Lacking valid instruments, longitudinal data (sometimes synthetic cohorts might do the trick) or twins data, you are pretty much left with three choices.<sup>17</sup> First, learn to live with the endogeneity problem. Second, impose suitable restrictions on the covariance structure of the regression errors (Pitt et al. 1998). Third, use the estimated residuals from the first stage regression corresponding to the endogenous regressor as an 'instrument' for the endogenous regressor in the second stage regression, that is, replace the endogenous regressor with the estimated residuals (Foster and Roy, 1997). I have used this approach in estimating the effect of the number of pre-school children on the time allocation of school-age children (Ridao-Cano, 2000).<sup>18</sup>

#### 4. Truncation, Censoring, and Sample Selection

Up until now we have assumed that the dependent variable of interest is continuously observed. However, dependent variables are sometimes limited in their range (i.e. they are observed in only some of the ranges). For example, data from the negative income tax experiment are such that income lies at or below some threshold level. Data on household expenditures on automobiles has a lot of observations at 0, corresponding to households who choose not to buy a car. Data on hours of work and earnings is only available for those working. Data on the health status of women from of a particular cohort at a given point in time is only available for those women in the cohort who are alive at that time.

Samples with limited dependent variables are classified into two general categories, depending on whether or not the values of  $\mathbf{X}$  for the missing  $\mathbf{y}$  data are known:

---

<sup>17</sup>The problem of finding valid indentifying instruments is specially so in the study of intra-household resource allocation issues. This is related to the well-known 'more goods than prices' problem.

<sup>18</sup>Even if these residuals are still correlated with the error term in the main equation, there is certainly more exogenous variation in those than in the endogenous regressor, so at least we are doing relatively better.

1. *Censored sample.* In this case some observations on  $\mathbf{y}$ , corresponding to known values of  $\mathbf{X}$ , are not observable. For example, in a study of the determinants of earnings, you may have data on the explanatory variables for people working and not working, but only data on earnings for those working.
2. *Truncated sample.* In this case values of  $\mathbf{X}$  are known only when  $\mathbf{y}$  is observed. In the negative income tax experiment, for example, no data of any kind are available for those above the income threshold, they were simply not part of the sample.

As we will show, if the dependent variable is limited in some way, the OLS estimation, based on the sample of individuals for whom the dependent variable is observed, produces biased and inconsistent estimates of the *population* parameters. The OLS bias can be interpreted within the omitted variable framework developed earlier.

In general, the key issue is that the sample of individuals for whom the dependent variable is observed is *not a random or representative sample* of the whole population, but a *selected sample*. More precisely, 'sample inclusion' is not determined exogenously (i.e. independently) to the behavioral process being studied, but endogenously. I want you to keep in mind that sample selection is an issue at all if what we want is to use the estimation results to make inferences about the whole population, more so the higher the proportion of excluded observations. If, on the contrary, we just want to make inferences about the selected sample then we are safe with OLS. I am telling you this so that you do not overuse the type of models that I am going to be presenting here.

For example, in our hours of work model we have so far focused on the sample of workers. However, nature does not randomly assign people in and out of the labor force. Those in the labor force may differ in some observable and/or unobservable respect from those out of the labor force, and it is precisely those differences that determine whether an individual is in or out of the labor force. Hence, individuals *self-select* into the labor force, and this selection is not independent of their choice of how many hours they want to work.

If the selection rule, in this case the labor force participation equation, is assumed to be equal to the hours of work equation we have the so called tobit model, whereas if we allow the two equations to be different but related we have the so called sample selection model. There are many sample selection models depending on the specification of the selection equation and the nature of the

dependent variable. The simplest sample selection model is the Heckman model (1979), where the selection equation is a probit, and the dependent variable is continuously observed above the observation threshold.<sup>19</sup>

Although most empirical work on this subject involves censoring rather than truncation, we will study the simpler truncated regression model first, since it provides most of the theoretical tools we need to analyze models of censoring and sample selection.

#### 4.1. The Truncated Regression Model

A good example of the truncated regression model is the earnings equation estimated from the data for the negative income tax experiment (Hausman and Wise, 1976, 1977). There, families with income levels above a certain limit (1.5 times the 1967 poverty line) were eliminated from the study. The truncation thus takes the form  $y_i < c$ .<sup>20</sup>

We assume that in the population the relationship between earnings and exogenous variables is of the form

$$y_i = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i \quad (4.1)$$

which satisfies all the assumptions of the CLRM and  $\varepsilon_i \sim N[0, \sigma^2]$ , so that

$$y_i \sim N[\boldsymbol{\beta}' \mathbf{x}_i, \sigma^2]$$

However, even though families were selected at random, only those with incomes higher than  $c$  were eliminated from the study. To fully specify the density function of  $y_i$  in this situation we just need to define the distribution of  $y_i$  given that  $y_i < c$  under the normality assumption

$$f(y_i | y_i < c) = \frac{f(y_i)}{\Pr(y_i < c)} = \frac{\frac{1}{\sigma} \phi[(y_i - \boldsymbol{\beta}' \mathbf{x}_i)/\sigma]}{\Phi[(c - \boldsymbol{\beta}' \mathbf{x}_i)/\sigma]} \quad (4.2)$$

Where  $\phi(\cdot)$  and  $\Phi(\cdot)$  are, respectively, the density function and the distribution function of the standard normal. Then the density function of  $y_i$  is the truncated normal. **(see graphical illustration on the board)**

---

<sup>19</sup>There is a very important class of models involving censored data, namely, duration models. Due to time constraints, we will not be covering them.

<sup>20</sup>In the actual study,  $c$  depends on family size. Furthermore, the threshold refers to family income but the earnings equation refers to individuals. For simplicity, we ignore these two issues here.

$$\begin{aligned}
f(y_i) &= \frac{\frac{1}{\sigma}\phi[(y_i - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma]}{\Phi[(c - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma]} \text{ if } y_i \leq c \\
f(y_i) &= 0 \text{ otherwise}
\end{aligned}$$

Now define

$$\lambda(\alpha_i) = -\frac{\frac{1}{\sigma}\phi[(y_i - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma]}{\Phi[(c - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma]}$$

where  $\alpha = (c - \boldsymbol{\beta}'\mathbf{x}_i)/\sigma$ , and the function  $\lambda_i$  is called the inverse Mills ratio. Since no observations are available for  $y_i > c$ , the expected value and variance of  $y_i$  conditional on  $y_i > c$  are equal to their unconditional counterparts

$$E[y_i \mid y_i < c] = E[y_i] = \boldsymbol{\beta}'\mathbf{x}_i + \sigma\lambda_i \quad (4.3)$$

$$Var[y_i \mid y_i < c] = Var[y_i] = \sigma^2[1 - \delta(\alpha_i)] \quad (4.4)$$

where  $\delta(\alpha_i) = \lambda(\alpha_i)[\lambda(\alpha_i) - \alpha_i]$ . We now consider the OLS estimation of the parameters of the truncated regression. For the subpopulation of individuals such that  $y_i < c$  we could write (4.3) in the form

$$y_i \mid y_i < c = y_i = E[y_i] + u_i = \boldsymbol{\beta}'\mathbf{x}_i + \sigma\lambda_i + u_i \quad (4.5)$$

By construction  $u_i$  has zero mean, but it is heteroscedastic. If we estimate (4.5) by OLS excluding the non-linear term  $\lambda_i$  (which is then part of the error term) we then incur in a standard omitted variable bias, since  $\lambda$  is by construction a (non-linear) function of  $\mathbf{x}$ , and thus correlated with the specified  $\mathbf{x}$ . However, without knowledge of the distribution of  $\mathbf{x}$ , it is not possible to determine how serious the bias is likely to be. In this and other applications, it has been usually found that the OLS estimates are biased toward zero.<sup>21</sup> A possible interpretation of this result is low income individuals have on average lower ability and lower motivation but also achieve lower levels of education. The latter would bias the estimated return to education downward. (see **graphical illustration on the board**).

---

<sup>21</sup>Note that even if we included  $\lambda_i$  OLS would still produce inefficient estimates since  $\mu$  is heteroscedastic.

A natural candidate for producing consistent estimates in this context is the maximum likelihood (ML) method. Recall that the maximum likelihood estimator (MLE) is the value of the parameter that maximizes the probability of jointly observing out data realization  $(y_i, \mathbf{x}_i)$  for all  $i$ . Since observations are independent, the likelihood function is just the product of individual contributions to the likelihood function. Thus using (4.2) the likelihood of the problem is

$$\begin{aligned}
 L &= \prod_i \Pr[Y_i = y_i \mid y_i < c] \\
 &= \prod_i \Pr[\varepsilon_i = y_i - \boldsymbol{\beta}' \mathbf{x}_i \mid \varepsilon_i < c - \boldsymbol{\beta}' \mathbf{x}_i] \\
 &= \prod_i \frac{f(y_i - \boldsymbol{\beta}' \mathbf{x}_i)}{\Pr[\varepsilon_i < c - \boldsymbol{\beta}' \mathbf{x}_i]} \\
 &= \prod_i \frac{\frac{1}{\sigma} \phi[(y_i - \boldsymbol{\beta}' \mathbf{x}_i)/\sigma]}{\Phi[(c - \boldsymbol{\beta}' \mathbf{x}_i)/\sigma]} \tag{4.6}
 \end{aligned}$$

The maximum likelihood estimators for this problem are those values of  $\boldsymbol{\beta}$  and  $\sigma$  that maximize  $\log L$ . The resulting MLE estimators are consistent and asymptotically efficient.<sup>22</sup>

As mentioned earlier, there are not many applications of the truncated regression model, but here is another example that I can think of. Jane's project is concerned with the examination of the determinants of health among women of a certain cohort. Think of health as a latent continuous variable, which below a certain threshold means death of the woman. At a particular point in the life of this cohort some women of this cohort have already died so you do not have information on their current health (of course) nor you have information on their current characteristic that are relevant for explaining health. If we apply OLS to the regression of health on characteristics for the sample of surviving women, you will probably get biased and inconsistent estimates. For example, health related behavior can make a lot of difference in determining life or death but have a minor effect on the health status of surviving women.

#### 4.2. The Tobit (Censored Regression) Model

A very common problem in micro data is censoring of the dependent variable. When the dependent variable is censored, values in a certain range are all trans-

---

<sup>22</sup>This is of course conditional on the other assumptions of the CLRM holding.

formed (or reported as). a single value. For example, suppose we are interested in the number of tickets *demanded* for events at a certain arena (the *latent* dependent variable). Our only measure is the number of tickets actually *sold* (the *observed* dependent variable). However, whenever an event sells out, we know that the actual number demanded is larger than the number sold. The number of tickets demanded is censored at zero when it is transformed to obtain the number sold.

Some other examples that have appeared in the empirical literature include household expenditures on various commodities, time allocated to various activities by household members, the amount of borrowing from financial institutions, the number of extramarital affairs. Each of these studies analyzes a dependent variable that is zero for a significant fraction of observations. Conventional regression methods fail to account for the qualitative difference between *limit* (zero) and *nonlimit* (continuous) observations.

The relevant distribution theory for a censored variable is similar to that for a truncated one. Once again, we focus on the normal distribution. We also assume that the censoring point is zero, though this is only a convenient normalization. In a truncated distribution, only the part of the distribution above  $y = 0$  is relevant to our computations. To make the density function integrate to one, we scale it up by the probability than an observation in the untruncated population fall in the range that interests us. When data are censored, the distribution that *applies to the sample data* is a mixture of discrete and continuous distributions. (see graphical representation).<sup>23</sup>

The regression model based on censored data is referred to as the censored regression model or the tobit model. Without loss of generality, we can discuss the tobit model with reference to a an intra-household time allocation model developed by Datt and Ravallion (1994). The main goal of this paper is to examine the effect of public-works employment on the time allocation of males and females to different activities. In this model, hours spent in public-works employment is suspected to be endogenous, so a method for obtaining consistent estimates in the context of a tobit model (a non-linear model) is developed. Let us ignore this endogeneity problem now, and focus on one activity only, say, wage employment.

Suppose the model explaining desired/optimal hours of market work (the latent variable) is as follows

---

<sup>23</sup>Note that censored observations are part of the saample, whereas truncated observation are not part of the sample.

$$h_i^* = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i \quad (4.7)$$

where the assumptions of the CLRM are met and  $\varepsilon_i \sim N[0, \sigma^2]$  so that  $h_i^* \sim N[\boldsymbol{\beta}' \mathbf{x}_i, \sigma^2]$ , and  $E[h_i^*] = \boldsymbol{\beta}' \mathbf{x}_i$ . However we only observe hours of market work actually performed

$$\begin{aligned} h_i &= h_i^* \text{ if } h_i^* > 0 \\ h_i &= 0 \text{ if } h_i^* \leq 0 \end{aligned} \quad (4.8)$$

That is, we observe individuals' optimal choice of hours if they actually choose to work at all, otherwise all we observe is that they do not work, that is, all we know is that  $h_i^* \leq 0$ . The contribution to the likelihood function of those individuals such that  $h_i^* \leq 0$  is

$$\Pr[h_i = 0] = \Pr[h_i^* \leq 0] = \Pr[\varepsilon_i \leq -\boldsymbol{\beta}' \mathbf{x}_i] = \Phi\left[-\frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma}\right] = 1 - \Phi\left[\frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma}\right] \quad (4.9)$$

while the contribution to the likelihood function for those such that  $h_i^* > 0$  is

$$\begin{aligned} \Pr[h_i = h_i^*] &= \Pr[h_i^* > 0] \Pr[h_i = h_i^* | h_i^* > 0] \\ &= [1 - \Pr[h_i^* \leq 0]] f[h_i^* | h_i^* > 0] \\ &= [1 - \Pr[h_i^* \leq 0]] \frac{f[h_i - \boldsymbol{\beta}' \mathbf{x}_i]}{[1 - \Pr[h_i^* \leq 0]]} \\ &= \frac{1}{\sigma} \phi\left[\frac{h_i - \boldsymbol{\beta}' \mathbf{x}_i}{\sigma}\right] \end{aligned} \quad (4.10)$$

Consider first the OLS estimation of the parameters of the model in (4.7) using the sample of workers only. The expected value of  $h_i^*$  conditional on  $h_i^* > 0$  is<sup>24</sup>

$$E[h_i^* | h_i^* > 0] = \boldsymbol{\beta}' \mathbf{x}_i + E[\varepsilon_i | \varepsilon_i > -\boldsymbol{\beta}' \mathbf{x}_i] = \boldsymbol{\beta}' \mathbf{x}_i + \sigma \lambda_i \quad (4.11)$$

where

$$\lambda_i = \frac{\phi\left[\frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma}\right]}{\Phi\left[\frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma}\right]} = \frac{\phi_i}{\Phi_i}$$

---

<sup>24</sup>Note that since  $h_i = h_i^*$  for the sample of workers,  $E[h_i^* | h_i^* > 0] = E[h_i | h_i^* > 0]$ .

is the inverse Mills ratio. Hence the model that should be estimated if we just consider the sample of workers is

$$h_i = E[h_i^* | h_i^* > 0] + u_i = \boldsymbol{\beta}' \mathbf{x}_i + \sigma \lambda_i + u_i \quad (4.12)$$

where, as in the truncation case,  $u_i$  has zero mean but it is heteroscedastic. Again, if we fit the above model by OLS excluding the non-linear term  $\lambda_i$  (which is then part of the error term) we then incur in a standard omitted variable bias, since  $\lambda$  is by construction a (non-linear) function of  $\mathbf{x}$ , and thus correlated with the specified  $\mathbf{x}$ . To be more precise, the key issue here is to understand that a change in  $\mathbf{x}$  has two effects. It affects  $E[h_i^*] = \boldsymbol{\beta}' \mathbf{x}_i$  in the positive part of the distribution, and it affects  $\lambda_i$  by affecting the probability that the observation will fall in the positive part of the distribution. The OLS estimator of  $\boldsymbol{\beta}$  of the model excluding  $\lambda$  would just be the sum of these two effects.

An alternative estimation method that produces consistent estimates is, again, maximum likelihood. The likelihood function is just the product of the individual contributions (4.9) and (4.10)

$$L = \prod_i \left[ \frac{1}{\sigma} \phi \left[ \frac{h_i - \boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right] \right]^{P_i} \left[ 1 - \Phi \left[ \frac{\boldsymbol{\beta}' \mathbf{x}_i}{\sigma} \right] \right]^{1-P_i} \quad (4.13)$$

where  $P_i$  is an indicator variable equal to one if  $i$  is doing market work, 0 otherwise. This likelihood function is a mixture of discrete and continuous distributions. The MLE for  $\boldsymbol{\beta}$  and  $\sigma$  are obtained by maximizing the log of  $L$  with respect to  $\boldsymbol{\beta}$  and  $\sigma$ . The bias of OLS is generally negative. The inconsistency of OLS with respect to MLE increases as the proportion of censored observations in the sample increases.

Note that even if we use all the observations (limit and nonlimit), OLS estimator would still be inconsistent. More precisely, in the case of a truncated sample we cannot use information on truncated observations, while information is available for censored observations. What this means is that, contrary to the truncated model, the conditional mean in (4.12) is different from the expected value of observed hours of work (for all observations)

$$\begin{aligned} E[h_i] &= \Pr[h_i^* > 0]E[h_i | h_i^* > 0] + \Pr[h_i^* \leq 0]E[h_i | h_i^* \leq 0] \\ &= \Phi_i \left[ \boldsymbol{\beta}' \mathbf{x}_i + \sigma \lambda_i \right] + [1 - \Phi_i]0 \\ &= \Phi_i \boldsymbol{\beta}' \mathbf{x}_i + \sigma \phi_i \end{aligned} \quad (4.14)$$

Note that in order to get a consistent OLS estimator of  $\beta$  based on all observations we still need to include  $\phi_i$  and  $\Phi_i$  in the regression. Finally, note that in the tobit model we have then three predictions:  $E[h_i^*]$ ,  $E[h_i | h_i^* > 0]$ , and  $E[h_i]$ . If what we want is to make inferences for the whole population then  $E[h_i]$  is the relevant prediction.

Now let us examine the endogeneity problem in this context. We mentioned earlier that the main goal of Datt and Ravallion (1994) is to examine the effect of public-works employment on intra-household time allocation, but they suspect that public-work employment is endogenous. Let  $l$  represent time devoted to public-works and rewrite the model in (4.7) and (4.8) as

$$\begin{aligned} h_i^* &= \beta' \mathbf{x}_i + \gamma l_i + \varepsilon_i \\ h_i &= h_i^* \text{ if } h_i^* > 0 \\ h_i &= 0 \text{ if } h_i^* \leq 0 \end{aligned} \tag{4.15}$$

where  $l$  is suspected to be endogenous (i.e. correlated with  $\varepsilon$ ). Now, a standard result in statistics is that the straight application of the 2SLS method does not guarantee consistency in a non-linear context (just take my word for it). The method proposed to get consistent estimates in this context was first developed by Smith and Blundell (1986) for a continuous endogenous regressor. Here we have a censored endogenous regressors but the principle is the same. In fact the new method looks very much like the standard 2SLS, at least in its principle. Let  $l$  be explained by the following tobit model

$$\begin{aligned} l_i^* &= \boldsymbol{\pi}' \mathbf{z}_i + u_i \\ l_i &= l_i^* \text{ if } l_i^* > 0 \\ l_i &= 0 \text{ if } l_i^* \leq 0 \end{aligned} \tag{4.16}$$

where, just like in the IV or 2SLS method,  $\mathbf{z}$  contains  $\mathbf{x}$  and some exclusionary restriction (i.e. identifying instrument).<sup>25</sup> Now the key assumption of this method is that  $\varepsilon$  and  $u$  are linearly related

$$\varepsilon_i = \delta u_i + \omega_i \tag{4.17}$$

---

<sup>25</sup>The parameter  $\gamma$  in (4.15) is still identified if  $z$  just includes  $x$  due to the nonlinearity of the Tobit model. However, identification relies solely on non-linearity, and the results in this case have been shown to be very unstable, so try to find some indentifying instrument.

Then plugging (4.17) in (4.15) we obtain

$$h_i^* = \beta' \mathbf{x}_i + \gamma l_i + \delta u_i + \omega_i \quad (4.18)$$

A sufficient condition for  $l$  to be weakly exogenous is that  $\delta = 0$ . To implement this technique we need to replace  $u_i$  with a consistent estimator. A consistent estimator of  $u_i$  is the residual from the estimated tobit model in (4.16), that is, the difference between predicted values (using  $E[l_i]$ ) and observed values ( $l_i$ ). Then the tobit model to be estimated is

$$\begin{aligned} h_i^* &= \beta' \mathbf{x}_i + \gamma l_i + \delta \hat{u}_i + \omega_i \\ h_i &= h_i^* \text{ if } h_i^* > 0 \\ h_i &= 0 \text{ if } h_i^* \leq 0 \end{aligned} \quad (4.19)$$

If  $\delta$  is significantly different from the zero then  $l$  is endogenous, and the inclusion of  $\hat{u}$  in (4.19) guarantess the consistency of the MLE of  $\gamma$ . I have used this method to correct to the endogeneity of parental credit from NGOs in a bivariate probit model of the school and work choices of school-age children in Bangladesh, although I modelled the credit equation as a sample selection model and not as a tobit model. Mother's credit turns out to be endogenous, and after controlling for endogeneity, mother's credit goes from being insignificant to having a significantly positive effect on child schooling. Ravallion and Wodon (1999) used this method to examine the effect of the Food-for-School Program in a school attendance probit.

### 4.3. Sample Selection Models

Technically, any model that controls for sample selection is a sample selection model. Sample selection occurs when the sample of individuals for whom the dependent variable is observed is *not a random or representative sample* of the whole population, but a *selected sample*. More precisely, sample inclusion or censoring is not determined exogenously (i.e. independently) to the behavioral process being studied, but endogenously.

Using this criteria, the tobit model is also a selection model. The main difference between the tobit model and the so called sample selection models is that the former assumes that the selection equation is the same as the equation of interest, whereas the latter group allows the selection equation to be different from the main equation but related.

Using the hours of work model as an example, the tobit model assumes not only that the variables determining participation and hours of work conditional on participation are the same, but also the coefficients on those variables are the same, so you end up estimating one set of parameters only. Sample selection models allow the coefficients on those variables to be different in both equations, so that you end up estimating two sets of parameters, one corresponding to the participation equation and the other corresponding to the hours of work equation conditional on participation. Whether the parameters in the two equations are equal or not is an empirical matter, but if they indeed are then the tobit estimates are a mixture of the true effect of each variable on hours of work and their effect on labor force participation, and thus biased and inconsistent.

There is enormous recent literature on sample selection models. Sample selection models basically vary according to the nature of the untruncated or uncensored part of the dependent variable (continuous or qualitative), and the selection rule (basically, binomial or multinomial).<sup>26</sup> Since the principle is the same in all of them, I will present the simplest model, namely, the Heckman model (1979), where the dependent variable is continuously observed if uncensored or untruncated, and the selection equation is a probit. Finally, I will show two very important applications/extensions of the basic sample selection model, namely program evaluation and switching models.

#### 4.3.1. The Heckman Model

To illustrate the mechanics of the model, let me use a project that I am working on. I am interested in the determinants of child schooling in Matlab (Bangladesh). To this extent, I can use a simple probit for school attendance, but I would be throwing away very useful information in the MHSS on the hours a child spends in school per day. However, this variable is obviously censored at zero for those children not currently attending school. Now the sample of children for whom information on school hours is available (i.e. those attending school) is not a random sample from the school-age children population, but a selected one. More precisely, children, or parents on their behalf, choose whether they want to attend school or not. What this means is that the school attendance decision is not independent of (exogenous to) the school hours decision conditional on school attendance.

Now the tobit model assumes that these two decisions are *the same*, but I

---

<sup>26</sup>See Green (1997) and Maddala (1983) for general surveys on these models. LIMDEP 7.0 has most of these models already built in.

suspect that the effect of variables such as parental education on school attendance is different from their effect on school hours conditional on school attendance. Hence, I want allow the two decisions to be somehow different but related.

Now suppose that my underlying *selection equation* is as follows

$$s_i^* = \boldsymbol{\gamma}' \mathbf{z}_i + u_i \quad (4.20)$$

Where  $s_i^*$  represents our latent selection variable, namely, propensity to go to school (or the net benefit of going to school), which is not observed. What we observe is whether or not the child attends school

$$s_i = 1 \text{ if } s_i^* > 0 \quad (4.21)$$

$$s_i = 0 \text{ if } s_i^* \leq 0 \quad (4.22)$$

Now our *model for desired or optimal school hours* is given by

$$h_i^* = \boldsymbol{\beta}' \mathbf{x}_i + \varepsilon_i \quad (4.23)$$

However, what we observe is

$$h_i = h_i^* \text{ if } s_i = 1 \quad (4.24)$$

$$h_i = 0 \text{ if } s_i = 0 \quad (4.25)$$

To complete the specification of the model,  $\varepsilon$  and  $u$  are assumed to be distributed as

$$\begin{bmatrix} \varepsilon \\ u \end{bmatrix} \sim BN \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon u} \\ \sigma_{\varepsilon u} & 1 \end{pmatrix} \right] \quad (4.26)$$

Thus, even though I am allowing the school attendance and school hours equations to be different, I am also allowing them to be related, that is,  $\sigma_{\varepsilon u}$  can be different from zero. That is, there are unobserved child or household characteristics that affect both the school attendance decision and the school hours decision, such as child motivation and expectations about future returns to education. The reason why we impose  $\sigma_u^2 = 1$  is that there is no information on the scale of  $s^*$  so  $\sigma_u^2$  cannot be estimated anyway, so to achieve identification of all parameters in the model we just normalize  $\sigma_u^2$  to one. Note that since  $\sigma_u^2 = 1$  then  $\sigma_{\varepsilon u} = \rho\sigma_\varepsilon$ .

More importantly, note that if  $\gamma' \mathbf{z}_i = \beta' \mathbf{x}_i$  and  $u_i = \varepsilon_i$  we are back to the tobit model, so that the tobit model is just a special case of the Heckman model.

Now suppose we estimate the school hours model for the sample of children attending school. Then,

$$\begin{aligned}
E[h_i \mid s_i = 1] &= \beta' \mathbf{x}_i + E[\varepsilon_i \mid \gamma' \mathbf{z}_i + u_i > 0] \\
&= \beta' \mathbf{x}_i + E[\varepsilon_i \mid u_i > -\gamma' \mathbf{z}_i] \\
&= \beta' \mathbf{x}_i + \sigma_{\varepsilon u} \frac{f_u[-\gamma' \mathbf{z}_i]}{1 - F_u[-\gamma' \mathbf{z}_i]} \\
&= \beta' \mathbf{x}_i + \sigma_{\varepsilon u} \frac{\phi[-\gamma' \mathbf{z}_i]}{1 - \Phi[-\gamma' \mathbf{z}_i]} \\
&= \beta' \mathbf{x}_i + \sigma_{\varepsilon u} \frac{\phi[\gamma' \mathbf{z}_i]}{\Phi[\gamma' \mathbf{z}_i]} \\
&= \beta' \mathbf{x}_i + \rho \sigma_{\varepsilon} \lambda(\gamma' \mathbf{z}_i)
\end{aligned} \tag{4.27}$$

Note, again, that if  $\gamma' \mathbf{z}_i = \beta' \mathbf{x}_i$  and  $\rho = 1$  we are back to the tobit model. Thus the regression model corresponding to those children attending school is

$$h_i = \beta' \mathbf{x}_i + \rho \sigma_{\varepsilon} \lambda(\gamma' \mathbf{z}_i) + \omega_i \tag{4.28}$$

where  $\omega$  has zero mean but it is heteroscedastic. Again, if we apply OLS to (4.28) excluding  $\lambda$  we get inconsistent estimates of  $\beta$ . The resulting bias is called selection bias. The marginal effect of any regressor  $x_k$  that appears in both  $\mathbf{x}$  and  $\mathbf{z}$  on  $h$  consists of two components. There is a direct effect on the mean of  $h$ , which is  $\beta_k$ . In addition,  $x_k$  it also influence  $h$  indirectly through its effect on  $\lambda$ . Thus if the selectivity correction,  $\lambda$ , is omitted for the least squares regression, the OLS estimate of  $\beta_k$  would include these two effects.

The parameters of the sample selection model of equations (4.20) through (4.26) can be consistently (and efficiently) estimated using maximum likelihood (full information maximum likelihood). The likelihood function is a combination of the likelihood function corresponding to the probit model (for the limit observations) and that corresponding to the truncated regression model (for the nonlimit observations). The estimation is quite cumbersome, and an alternative procedure due to Heckman (1979) has been used in the past instead, namely, the two-step procedure. The two-step procedure is consistent but not efficient. Nowadays, however, statistical packages such as STATA 6.0 or LIMDEP 7.0 have the full information

likelihood estimation method built in, as well as the two-step procedure. Having said that, the two-step procedure is very instructive and intuitive, so it is worth presenting it.

The two-step procedure is as follows:

- **First Step**

Estimate the probit equation for school attendance by maximum likelihood to obtain estimates of  $\gamma$ . The contribution of school attenders and non-attenders to the likelihood function are, respectively,

$$\Pr[s_i = 1] = \Phi[\gamma' \mathbf{z}_i] \text{ and } \Pr[s_i = 0] = 1 - \Phi[\gamma' \mathbf{z}_i] \quad (4.29)$$

so the likelihood function for the whole sample is

$$L = \prod_i [\Phi[\gamma' \mathbf{z}_i]]^{s_i} [1 - \Phi[\gamma' \mathbf{z}_i]]^{1-s_i} \quad (4.30)$$

Next, for each observation in the selected sample compute

$$\hat{\lambda}_i = \frac{\phi[\hat{\gamma}' \mathbf{z}_i]}{\Phi[\hat{\gamma}' \mathbf{z}_i]} \quad (4.31)$$

- **Second Step**

Estimate  $\beta$  and  $\sigma_{\varepsilon u}$  in (4.28) by OLS replacing  $\lambda$  with  $\hat{\lambda}$ . These estimates are consistent since we are controlling for  $\lambda_i$  (the selectivity correction) which says something about a child's propensity to attend to school. Hence by including (an estimate of)  $\lambda$  we are controlling for the fact that those attending school have, on average, a higher propensity to go to school than the population as a whole. These estimates, however, are not efficient since we are using  $\hat{\lambda}$  instead of  $\lambda$ , and the error in the school hours equation is heteroscedastic (although this source of inefficiency can be corrected).<sup>27</sup> It is also possible to obtain consistent estimates of  $\rho$  and  $\sigma_\varepsilon$  based on this model (see Green, 1997).

To test whether there is indeed selection bias, we just need to test the hypothesis  $\rho = 0$  using a t-test or a Wald test. To test whether the school attendance and school hours equations are the same we just need to test whether the parameters associated to the variables that are common in both equations are the same.

---

<sup>27</sup>This is contrary to the maximum likelihood method where the school attendance and the school hours equations are simultaneously estimated.

Now to test this you have to estimate the parameters of the model by maximum likelihood -since this method estimates the parameters of both equations simultaneously, that is, within the same model- and then use the Wald test (since the likelihood function is non-linear function of the parameters).

Up until now we have not said a word about the variables contained in  $\mathbf{x}$  and  $\mathbf{z}$ . Suppose that  $\mathbf{x}$  and  $\mathbf{z}$  contain the same variables. Then we can rewrite the second step regression as

$$h_i = \boldsymbol{\beta}' \mathbf{x}_i + \sigma_{\varepsilon u} \lambda(\hat{\boldsymbol{\gamma}}' \mathbf{x}_i) + \omega'_i \quad (4.32)$$

Now if  $\lambda$  were a linear function of  $\mathbf{x}$  then we could not estimate  $\boldsymbol{\beta}$  (i.e.  $\boldsymbol{\beta}$  is not identified) due to perfect multicollinearity since  $\lambda$  is a linear combination of  $\mathbf{x}$ . Now in the Heckman model, under the normality assumption in (4.26),  $\lambda$  is a non-linear function of  $\mathbf{x}$

$$\lambda(\hat{\boldsymbol{\gamma}}' \mathbf{x}_i) = \frac{\phi[\hat{\boldsymbol{\gamma}}' \mathbf{x}_i]}{\Phi[\hat{\boldsymbol{\gamma}}' \mathbf{x}_i]}$$

In this case  $\boldsymbol{\beta}$  would be identified, but identification would just rely on the non-linearity of  $\lambda$ . However, as mentioned in other sections, the parameter estimates in this case have been shown to be very sensitive to small changes in specification. The moral of the story is that we need an exclusionary restriction, that is, something that affects school hours only through school attendance (sound familiar?). What we are really doing is instrumenting for  $\lambda$  in the second step regression. The discussion about finding valid instruments to correct for measurement error and endogeneity applies here as well. The only difference is that here we can at least call the results without the exclusionary restriction as a last resort.

Before we move any further, I just want to remind you that the selection bias is only relevant if what you want is to use your estimation results to infer something about the population as a whole. However, if you just want to focus your inference on the selected sample, you are safe using OLS.

### 4.3.2. Program Evaluation

As I mentioned earlier, the basic sample selection model can be applied to a relatively new and very hot topic, namely, the evaluation of programs. If we had a true random experiment, we would randomly assign individuals to the treatment and control groups, so that the only difference between the two groups would be that the treatment group receives treatment (the program) and the control

group does not, so that the measure of the program effectiveness in terms of, say, earnings would be the difference in average earnings between the two groups. True random experiments are rare in social sciences, so we normally have to care about individual self-selection into programs when evaluating their effect.

Suppose we are interested in investigating the work disincentive effect of participation in the AFDC/TANF program. Ignoring other sample selection issues, let us focus on the sample of workers that could potentially qualify for AFDC. The model to be estimated is

$$h_i = \beta' \mathbf{x}_i + \gamma P_i + \varepsilon_i \quad (4.33)$$

where  $P$  is a dummy variable that is equal to one if the individual participates, zero otherwise. The key question here is whether the estimate of  $\gamma$  captures the true work disincentive effect of AFDC participation. The answer is probably not since individuals self-select into AFDC, that is, AFDC participation is endogenous (sound familiar?). Another way of saying this is that, contrary to a random experiment, AFDC participants and non-participants differ in observable and unobservable ways, and these differences determine whether an individual actually participates or not. It is likely that the OLS estimate of  $\gamma$  overestimates the work disincentive effect to the extent that, say, low motivated individuals (and hence individuals working less hours) are more likely to participate in the AFDC program. We have already seen a way to deal with the endogeneity of  $P$ , namely, instrumental variables/2SLS. An alternative approach is to use the selection model.

We model the propensity to participate in the AFDC program as

$$P_i^* = \gamma' \mathbf{z}_i + u_i \quad (4.34)$$

However, we only observe whether the individual participates or not, that is,

$$\begin{aligned} P_i &= 1 \text{ if } P_i^* > 0 \\ P_i &= 0 \text{ if } P_i^* \leq 0 \end{aligned}$$

The errors  $\varepsilon$  and  $u$  are allowed to be correlated as before. This correlation is probably negative since unobserved individual characteristics such as motivation reduce the likelihood of AFDC participation but increase hours of work. The conditional expectations for the sample of participants and the sample of non-participants is, respectively,

$$E[h_i \mid P_i = 1] = \beta' \mathbf{x}_i + \gamma + \sigma_{\varepsilon u} \frac{\phi[\gamma' \mathbf{z}_i]}{\Phi[\gamma' \mathbf{z}_i]} \quad (4.35)$$

$$E[h_i \mid P_i = 0] = \beta' \mathbf{x}_i + \sigma_{\varepsilon u} \frac{-\phi[\gamma' \mathbf{z}_i]}{1 - \Phi[\gamma' \mathbf{z}_i]} \quad (4.36)$$

The difference in expected hours of work between participants and non-participants is then

$$E[h_i \mid P_i = 1] - E[h_i \mid P_i = 0] = \gamma + \sigma_{\varepsilon u} \left[ \frac{\phi_i}{\Phi_i[1 - \Phi_i]} \right] \quad (4.37)$$

If the selectivity correction,  $\lambda_i$ , is omitted from the least squares regression, this difference is what is estimated by OLS coefficient estimate on  $P_i$ . Assuming that  $\gamma$  and  $\sigma_{\varepsilon u}$  are negative, the OLS estimate of  $\gamma$  overestimates the negative effect of AFDC participation on hours of work.

### 4.3.3. Switching Models

Another very important application/extension of the sample selection model is the switching model. The program evaluation model developed above is restrictive in the sense that  $P$  is not allowed to interact with the variables in  $\mathbf{x}$ . A more flexible approach is the so called switching model. Without loss of generality, let us consider (in honor of James) a mover/stayer model.

Suppose you are interested in estimating the effect of migration on earnings allowing migration status to interact with all variables explaining earnings. The earnings of individual  $i$  if he or she decides to stay in the present location is

$$y_{is}^* = \beta'_s \mathbf{x}_i + \varepsilon_{is} \quad (4.38)$$

and the earnings if he or she decides to move

$$y_{im}^* = \beta'_m \mathbf{x}_i + \varepsilon_{im} \quad (4.39)$$

Migration entails costs defined as

$$C_i^* = \alpha' \mathbf{w}_i + \nu_i \quad (4.40)$$

The individual migrates if the net benefit,  $y_{im}^* - y_{is}^*$  is greater than the cost  $C_i^*$ . The net benefit of moving is

$$\begin{aligned}
M_i^* &= y_{im}^* - y_{is}^* - C_i^* \\
&= \beta'_m \mathbf{x}_i - \beta'_s \mathbf{x}_i - \alpha' \mathbf{z}_i + (\varepsilon_{im} - \varepsilon_{is} - \nu_i) \\
&= \gamma' \mathbf{z}_i + u_i
\end{aligned} \tag{4.41}$$

However, we only observe whether the individual actually moves or stays, that is,

$$M_i = 1 \text{ if } M_i^* > 0 \tag{4.42}$$

$$M_i = 0 \text{ if } M_i^* \leq 0 \tag{4.43}$$

Furthermore, we only observe migrants' earnings if the individual actually migrates and stayers' earnings if the individual actually stays, that is,

$$y_i = y_{im}^* \text{ if } M_i = 1, 0 \text{ otherwise} \tag{4.44}$$

$$y_i = y_{is}^* \text{ if } M_i = 0, 0 \text{ otherwise} \tag{4.45}$$

To complete the specification of the model assume that the three errors of the model and distributed as a trivariate normal with zero mean and covariance

$$Cov(\varepsilon_{im}, \varepsilon_{is}, u_i) = \begin{bmatrix} \sigma_m^2 & \sigma_{ms} & \sigma_{mu} \\ \sigma_{ms} & \sigma_s^2 & \sigma_{su} \\ \sigma_{mu} & \sigma_{su} & 1 \end{bmatrix} \tag{4.46}$$

Note that in this case we have two regression models (i.e. the two earnings equations), and one selection equation (i.e. the migration equation). The model can be estimated by full information maximum likelihood or the Heckman two-step procedure.

Now for each migrant we can compare the moving outcome  $y_{im}^*$  and the expected potential outcome without migration, that is,  $E[y_{is}^* | M_i = 1]$ . Under the normality assumption, the gross benefit for migrant  $i$  is

$$y_{im}^* - E[y_{is}^* | M_i = 1] = y_{im}^* - \beta'_s \mathbf{x}_i + \sigma_{su} \frac{\phi[\gamma' \mathbf{z}_i]}{\Phi[\gamma' \mathbf{z}_i]} \tag{4.47}$$

The expected gross benefit for migrant  $i$  is

$$E[y_{im}^* | M_i = 1] - E[y_{is}^* | M_i = 1] = (\beta'_m - \beta'_s)\mathbf{x}_i + (\sigma_{su} - \sigma_{mu}) \frac{\phi[\gamma' \mathbf{z}_i]}{\Phi[\gamma' \mathbf{z}_i]} \quad (4.48)$$

If self-selection is based on comparative advantage,  $(\sigma_{su} - \sigma_{mu})$  is greater than zero. Then estimated benefit of migration obtained by OLS without control for self-selection actually overestimates the benefit of migration.

## References

- [1] Altonji, J. (1986), "Intertemporal Substitution in Labor Supply: Evidence from Micro Data", *Journal of Political Economy*, June 1986.
- [2] Amemiya, T. (1984), "Tobit Models: A Survey", *Journal of Econometrics*, 24, 3-63.
- [3] Amemiya, T. (1985), *Advanced Econometrics*, Harvard University Press.
- [4] Amemiya, T. (1994), *Introduction to Statistics and Econometrics*, Harvard University Press.
- [5] Angrist, J. and A. Krueger (1991), "Does Compulsory Schooling Attendance Affect Schooling and Earnings", *Quarterly Journal of Economics*, November 1991, 979-1014.
- [6] Ashenfelter, O. and A. Krueger (1994), "Estimates of the Economic Return to Schooling from a New Sample of Twins", *American Economic Review*, December 1994, 1157-1173.
- [7] Bound, J. and A. Krueger (1991), "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?", *Journal of Labor Economics*.
- [8] Bound, J. D. Jaeger, and R. Baker (1995), "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak", *Journal of the American Statistical Association*, June 1995, 443-450.
- [9] Baltagi, B. (1995), *Econometrics Analysis of Panel Data*, Wiley ed.
- [10] Behrman, J., N. Birdsall and A. Deolalikar (1995), "Marriage Markets, Labor Markets, and Unobserved Human Capital: An Empirical Exploration for South-Central India", *Economic Development and Cultural Change*.
- [11] Blackburn, M. L. and D. Neumark (1995), "Are OLS Estimates and the Return to Schooling Biased Downward?", *Review of Economics and Statistics*, 77, 217-230.

- [12] Borjas, G. (1987), "Self-Selection and the Earnings of Immigrants", *American Economic Review*, 77, 531-553.
- [13] Cameron, A. and P. Trivedi (1998), *Regression Analysis of Count Data*, Econometric Society Monographs No. 30, Cambridge University Press.
- [14] Datt, G. and M. Ravallion (1994), "Income Gains for the Poor from Public Works Employment: Evidence from Two Indian Villages, *LSMS Paper 100*, World Bank.
- [15] Engle, R. and D. MacFadden, eds. (1994), *Handbook of Econometrics*, Volume 3 and 4, North Holland.
- [16] Falaris, E. (1987), "A nested Logit Migration Model with Selectivity", *International Economic Review*, 28, 429-443.
- [17] Foster, A. and N. Roy (1997), "The Dynamics of Education and Fertility: Evidence from a Family Planning Experiment", Mimeo, University of Pennsylvania.
- [18] Green, W. H. (1997), *Econometric Analysis*, Upper Saddle River, N. J., Prentice Hall.
- [19] Griliches, Z. (1977), "Estimating the Returns to Schooling: Some Econometric Problems", *Econometrica*, January 1977, 1-22.
- [20] Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998), "Misclassification of the Dependent Variable in a Discrete-Response Setting", *Journal of Econometrics*, 87, 239-269.
- [21] Heckman, J. (1979), "Sample Selection as a Specification Error", *Econometrica*, 47, 153-161.
- [22] Hsiao, C. (1986), *Analysis of Panel Data*, Cambridge University Press.
- [23] Kennedy, P. (1992), *A Guide to Econometrics*, Blackwell.
- [24] LaLonde, R. (1986), "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", *American Economic Review*, 76, 604-620.

- [25] Lam, D. and R. F. Schoeni (1993), "Effects of Family Background on Earnings and Returns to Schooling: Evidence from Brazil", *Journal of Political Economy*, 101, 710-740.
- [26] Lancaster, T. (1990), *The Analysis of Transition Data*, New York, Cambridge University Press.
- [27] Lee, L. F. (1978), "Unionism and Wage Determination", *International Economic Review*, 19.
- [28] MaCurdy, T. (1981), "An Empirical Model of Labor Supply in a Life Cycle Setting", *Journal of Political Economy*, 89, 1059-1089.
- [29] Maddala, G. S. (1983), *Limited-Dependent and Qualitative Variables in Econometrics*, New York, Cambridge University Press.
- [30] Maddala, G. S. (1992), *Introduction to Econometrics*, New York, Macmillan, 1992.
- [31] Montmarquette, C., S. Mahseredjian and R. Houle (1996), "The Determinants of University Dropouts: A Sequential Decision Model with Selectivity Bias", *Working Paper 96-04*, CIRANO.
- [32] Mroz, T. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions", *Econometrica*.
- [33] Newey, W. (1987), "Efficient Estimation of Limited Dependent Variable Model with Endogenous Explanatory Variables", *Journal of Econometrics*, 36, 231-250.
- [34] Pitt, M. M., S. R. Khandker, O. H. Chowdhury and D. L. Millimet (1998), "Credit Programs for the Poor and the Nutritional Status of Children in Rural Bangladesh", *Working Paper 98-01*, PSTC, Brown University.
- [35] Pitt, M. M. (1999), "Credit Programs for the Poor and the Productive Behavior in Low-Income Countries: Are the Reported Causal Relations the Result of Heterogeneity Bias?", *Demography*, 36, 1-21.
- [36] Ravallion, M. and Q. Wodon (1999), "Does Child Labor Displace Schooling? Evidence on Behavioral Responses to an Enrollment Subsidy", Mimeo, the World Bank.

- [37] Ridao-Cano, C. (2000), "Child Labor and Schooling in a Low Income Rural Economy", unpublished manuscript, Department of Economics, University of Colorado at Boulder.
- [38] Smith, R. J. and R. W. Blundell (1986), "An Exogeneity Test for a Simultaneous Equation Tobit Model with Application to Labor Supply, *Econometrica*, 54, 679-686.
- [39] Stamp, J. (1929), *Some Economic Factors in Modern Life*, London, King and Son.
- [40] Wallis, R. and S. Rosen (1979), "Education and Self-Selection", *Journal of Political Economy*, 87, S7-S36.