INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Measurement Error and the Effects of Illegal Drug Use on Birthweight: An Econometric Investigation

By

Yongjia Ye

A dissertation submitted to the Graduate Faculty in Economics in partial fulfillment of the requirements for the degree of Doctor of Philosophy, The City University of New York

1999

Copyright

This manuscript has been read and accepted for the Graduate Faculty in Economics in satisfaction of the dissertation requirement for the degree of Doctor of Philosophy.

12/16/98 [signature] Michael Grossman
Date Chair of Examining Committee

12/16/98 [signature] Michael Grossman
Date Executive Officer

Michael Grossman

Theodore Joyce

Robert Kaestner
Supervisory Committee

The City University of New York

Abstract

Measurement Error and the Effects of Illegal Drug Use on Birthweight: An Econometric Investigation

By

Yongjia Ye

Adviser: Distinguished Professor Michael Grossman

This paper mainly develops a theoretical criterion for assessing bias of OLS estimators when measurement errors in variables are ignored in the linear regression model. Several bias or variance correction indexes due to measurement errors are introduced in this paper. It also spells out some important empirical estimation issues on measurement errors. The author shows that the regression coefficients for continuous and binary measurement errors are not comparable. In estimating infant health production function, split sampling, bootstrap reshaping and simulating techniques, simultaneous equations model with the endogenous binary or latent variables, and two step estimation procedure combined with the corrections for standard errors in the second step, heteroscedasticity, multicollinearity, selectivity and specification errors are used in the analysis. The findings in this paper suggest that the OLS estimator is biased downwards due to binary measurement errors, and it is biased upwards except for Blacks due to continuous measurement errors. The direction and magnitude of change in the variance of OLS estimator for the variable given error can be determinate if one has information about measurement error. The restricted OLS estimator always has smaller variance than the unrestricted one if the omitted measurement error is non-stochastic in the finite sample.

The heavy user are more likely to report their use. The treatment effect on heavy users will be underestimated if one uses the coefficient on the actual to evaluate drug program aimed at heavy users. The study indicates that one can get consistent estimator in a less expensive way by using out-sample information about measurement error. Additionally, this paper convinces or extends some interesting findings in the literature about human behaviors of illicit drug use among pregnant women.

Acknowledgments

Thanks to Dr. Michael Grossman
Dr. Theodore Joyce
Dr. Robert Kaestner
Dr. Linda Edwards
Dr. Rachel F. Boaz
Kyumin Shim
Min Xu
Sang Pae

Thanks to My Wife, Parents and Son!

Contents

Tables

Chapter I
Introductions

Most economic data used in empirical analysis contain measurement errors. Some theoretical variables in the model may be unobservable because their measures are known to be imperfect or their scales of measurement do not exist. Under certain circumstances, the variables we measure are not really what we want to measure, proxy variables may be subject to large random measurement errors. Even for the observable variables, a variety of errors may occur due to different reasons. For instances, illicit drug use from birth certificate is measured with underreporting error in its frequency and participation. Urine test is a more objective tool in screening illicit drug users. But it may still involve some technical errors, test-timing errors and sampling errors etc. Usually, the researchers do not know the quantity of illicit drug use, only observe whether pregnant women use illicit drug or not from the public survey.

Based on large sample theory, measurement errors can be defined as stochastic variables in the model. They are expressed in two major forms: the absolute measurement errors which are equal to the actual minus the observed, and the deviated measurement errors which are equal to the observed minus its mean. One can also decompose measurement errors into two parts: systematic errors and random errors. The random errors follow the Gauss equation (Standard Normal Distribution). Which component will be dominating depends on measuring performance in reality. In experimental science, people may eliminate systematic errors by using new advanced devices or improving the devices' accuracy and preciseness, and reduce random errors by making repeated

experiments. As we know in social science, it is less possible for one to make repeated experiments, and it is difficult to control experimental conditions. Socioeconomic and cultural factors may affect human behaviors of response in the public survey, which will produce systematic errors in economic data. Measurement errors can be transmitted within the system given the relationship among the variables. They may have zero, positive or negative correlation with the observed and the actual variables especially for cross section data. Measurement errors in economic data are generally not symmetrical. In the context of illicit drug use, measurement errors include underreporting errors in frequency and participation. The marginal and heavy users among pregnant women may have different consumption preferences and reporting styles for illicit drug use. ZuKermen et al. (1989), Chasnoff et al. (1990), Vega et al. (1993) and Kaestner et al. (1996) show that the black women have the highest prevalence of illicit drug use (14.22%). They more frequently have evidence of cocaine use (7.5%), whereas the white women more frequently have evidence of marijuana use (14.4%). Among the races, white women have the highest underreporting rate for illicit drug use. On the average, heavy users are more likely to report illicit drug use in participation, but they are more likely to underreport illicit drug use in quantity.

It is a well-known finding for linear regression model in statistics that errors in explanatory variables lead to inconsistent OLS estimators unless the variables measured with errors are orthogonal to its errors and all other independent variables. When the knowledge of the variances of measurement errors is available, one can get consistent estimators by making an adjustment. If there is only one explanatory variable measured with error, its OLS estimator is biased downwards by assuming that the correlation

between the actual variable and measurement error is zero. Maddala (1977) shows that not only the coefficient of the variable measured with error is biased, but also all parameter estimates are biased. For the situation in which there are more than one variable measured with errors, the direction of bias can be either way.

The previous studies related to measurement error issues have focused on (1). Finding the bound for the true coefficient of the variable measured with error; (2). Getting consistent and less biased estimates by using IV method; (3). Treating measurement error as omitted variable in the model; (4). Using the proxy to substitute the true variable so as to reduce bias. The classical error in variable model imposes some heroic assumptions in order to narrow the bound for the true coefficient. There are no fruitful findings along this direction.

IV method can be used for approaching the measurement error problem without needing prior information about the variance of measurement error. But one needs to bear in mind that it is difficult to evaluate bias and efficiency for IV estimators. Sometimes no good instruments can be found. For example, given non-zero correlation among measurement error, the observed and the actual illicit drug use, it is really hard to find an instrument which is highly correlated with the observed, and not correlated with the error term. People may choose instruments from many alternatives by ad hoc methods. The first stage specification error may affect the result from the second stage. Bound et al. (1995) show that the weak partial correlation between instrument and the observed leads to inconsistency and bias of estimates. Maddala (1983) suggests that two step estimation methods can be used for simultaneous equations system with endogenous binary and latent variables. The coefficients for binary variable and its fitted values can be compared

in their magnitude in recursive simultaneous equations system. Some researchers propose to use high moment (HM) estimation in dealing with measurement error issues. However, it has long been recognized that HM estimators are notably more erratic than the corresponding least squares estimators.

Treating measurement error as an omitted variable is more general because the errors occurring in economic data are systematic rather than random in the sense that measurement error is highly correlated with true variable. If the partial correlation between measurement error and the observed is not zero, and measurement error has significant effect on birth outcome, the gross effect and direct effect for the observed measured with error will not be the same in the model. In the case of zero partial correlation between the observed and its error, the OLS estimator for the observed is unbiased, but the other parameter estimates are biased. The direction and magnitude of bias for the observed are mainly determined by the partial correlation between the observed and its error, and the coefficient on the measurement error. A very few studies in the literature discuss binary measurement error issues. Aigner (1973) shows that OLS estimator for the variable with 'classification error' is biased downwards due to measurement error in the model. Some studies suggest that OLS estimator for the mis-measured variable is biased upwards due to non-random binary measurement error.

This paper theoretically and empirically settles the issue of whether OLS estimator is biased downwards or upwards due to binary measurement error. The author raises the following questions: (1). What kinds of crucial assumptions are needed to make in order to reach Aigner (1973)'s conclusion? (2). Suppose one knows the actual variable, how should error in binary or continuous types be measured? (3). Are the coefficients for the

binary variable and its predicted values comparable? (4). Are the coefficients for the observed and the actual illicit drug use comparable in simple OLS regression model? (5). What are the reasonable indexes for assessing bias of the OLS estimator? (6). How can the information about measurement error be used in out-of-sample estimation? Both urine test sample in California (CA) and birth certificate sample in New York City (NYC) are used in this study. Under certain valid assumptions and by controlling the confounding variable, one can get predicted probabilities of actual illicit drug use in NYC by applying individual characteristics in NYC sample and Probit coefficients from CA sample. The predicted probability is a comprehensive measure for frequency and participation of illicit drug use, and it can be converted to a binary indicator by using equivalent mean method (EQM). Omitted variable model and simultaneous equations model with endogenous binary and latent variables are main statistical focus in this paper. The important findings in this analysis are that: (1). The classical error in variable model and Aigner (1973)'s model are special cases of the general model derived in this study; (2). The OLS estimator is biased downwards due to omitted binary measurement error, and it is biased upwards except for Blacks due to omitted continuous measurement error; (3). The direction and magnitude of change in the variance of OLS estimator for the variable given error can be determinate if one has information about measurement error; (4). The restricted OLS estimator always has smaller variance than the unrestricted one if omitted measurement error is non-stochastic in finite sample. (5). The effect of illicit drug use on birth weight for the underreported is smaller in absolute value than that for the reported. (6). The treatment effect on heavy users will be understated if one use the coefficient on the actual to evaluate drug program aimed at heavy users. Additionally, this paper

convinces and extends some interesting findings about human behaviors of illicit drug use.

The rest of the paper is laid out as follow. Chapter II briefly reviews the previous literatures. Chapter III derives theoretical results assessing bias, inconsistency and change in variance for the OLS estimator due to measurement error. Chapter IV discusses some important empirical estimation issues on measurement error. Chapter V sets up the theoretical and empirical models for health production function. Chapter VI describes data sets, key variables and some issues in creating measurement errors. Chapter VII analyzes some empirical findings related to binary measurement errors in health production functions. Chapter VIII presents other methods and findings in dealing with measurement error issues. Chapter IX shows how to re-utilize the out-of-sample information about measurement errors. Chapter X examines some additional findings related to human behaviors of illicit drug use among pregnant women. Chapter XI draws some important conclusions.

Chapter II
Literature Review

Measurement error in scientific experiment has been studied for long time. It can be decomposed into systematic and random components. The random component fits the standard normal distribution. According to the Central Limit Theorem, one can approach the true value by making repeated experiments as the sample size goes to infinity. Stochastic process analysis is a good tool, which can be used to filter out 'noise' in the data. Curve fitting and simulation techniques are also useful for measurement error analysis. Compared to data in nature science and engineering, both cross-section and time series data in social science involves more systematic errors. The problem of measurement error in the independent variables of a regression equation has caused renewed attention among econometricans.

2.1. Classical Error in Variable Model

In this model, if a single explanatory variable in a linear regression equation is subject to a form of stochastic measurement error, the model is not identified, but a set of estimates that asymptotically contains the true value is in the interval between the ordinary regression and the reverse regression. The OLS estimator is clearly biased downwards. In the multivariate case, given random measurement error in one variable, the coefficient of that variable is biased towards zero and the coefficient of any variable positively correlated with the variable measured with error is biased away from zero by assuming that the coefficient of the true variable is positive. If there are more than one independent variable measured with errors, the direction of bias can be either way.

Leamer (1983) starts asking how large measurement error is to make it impossible to put bounds on regression coefficients. He shows that even small measurement errors in some independent variables would open up the possibility of perfectly collinear explanatory variables and make the data useless for statistical inference. Klepper and Leamer (1984) ask under what kinds of conditions it is possible to draw some inference regarding the vector of unknown regression coefficients. They demonstrate that the true regression coefficient vector can be restricted to the convex hull of all possible regression if all these regressions yield coefficient vectors lying in the same orthant. Otherwise, the set of feasible coefficient vectors is unbounded. Klepper (1988) shows that the above bounding procedure can be adapted to bound the coefficients when the regression contains mismeasured binary variables. All these results require that measurement errors are random and uncorrelated with equation error term. If those assumptions are dropped, no bound exists any more. In order to narrow the bound, one needs to impose some restrictions. From the empirical aspect, there is no operational procedure being followed.

2.2. Instrumental Variable Method

When explanatory variables are endogenous, OLS gives biased and inconsistent estimators. One needs to find instruments for the observed, which are correlated to the true variables and uncorrelated to measurement errors and equation error term. Generally speaking, without some prior information about the variance of measurement error, IV can be used to get consistent and less biased estimates. Compared to OLS estimators, IV estimators have higher standard errors.

Nelson & Startz (1990) discuss the finite sample properties of IV estimation for the special case of exact identified by using simulation. Buse (1990) uses different

assumptions and approaches to analyze the finite sample properties of IV estimates. His finding suggests that (1). The bias of IV related to OLS is a function of τ^2/k, where τ^2 is the concentration parameter, k is the number of instruments; (2). The first stage F test statistic contains valuable information about the magnitude of the finite sample bias.

Bound et al. (1995) draw attention to the problems with IV estimation when the correlation between the instruments and the endogenous explanatory variables is weak. It is well recognized that these kinds of instruments are likely to produce estimates with large standard errors, and can lead to a large inconsistency in IV estimates. They demonstrate that in finite samples, IV estimates are biased in the same direction as OLS estimates as the R^2 between the instruments and the endogenous explanatory variable approaches zero. If one uses randomly generated instruments, estimating coefficients in the second stage are close to OLS estimates. When sample size increases, the bias of IV estimates decreases. They emphasize that first stage specification error may affect magnitude of bias of IV estimates in the second stage. Their results imply that with a priori in selecting instruments under study, it doesn't sufficiently mean that IV estimates will be less biased than OLS estimates.

Dagenais et al. (1997) propose consistent IV estimation for the linear regression model with error in variables that requires no extraneous information by using sample moments of order higher than two. The HM estimation is derived from the orthogonal conditions: $\underset{n\to\infty}{E}(w'\varepsilon/n)=0$, where w is a set of artificial instruments being a function of dependent variable and independent variables, ε is error term. Their finding suggests that HM estimators perform better than OLS estimators in terms of root mean squared errors and also in terms of size of type I errors of standard tests.

So far in dealing with measurement error or endogeneity issues, IV estimation is still a good candidate if one has no proxies for measurement errors or has no information about them. As is discussed before, IV estimation has certain limitation: (1). Maybe no instruments are available; (2). One may select instruments in an ad hoc manner; (3). There may exist finite sample bias; (4). It is difficult to adjust standard errors in some cases; (5). The endogenous binary variables in the model increase the complexity of the estimation procedure.

2.3. Omitted Variable Model

Specification errors may occur due to using wrong function forms, including unnecessary variables, omitting necessary variables and having measurement errors in the variables. If measurement errors are ignored in the model, OLS estimates will be biased unless measurement errors have no effect on the dependent variable or have zero partial correlation with the corresponding observed variables.

Aigner (1973) discusses a linear regression model with a binary independent variable subject to errors of observation. He shows that OLS estimator is biased downwards if one omits classification errors in the model. This conclusion is the same as that from classical error in variable model in which measurement error is random and continuous. Given knowledge of partial correlation between the observed and its error, one can get consistent estimator. But it is unclear whether OLS estimator is more efficient or not based on MSE criterion.

Kaestner et al. (1996) develops a method to correct for non-random measurement error in a binary indicator of illicit drug use. Measurement error is defined as the difference between the actual use and the observed use. It has no-zero mean and is

correlated to the actual illicit drug use. They assume that the likelihood of reporting illicit drug use is correlated with frequency of illicit drug consumption, which means that reported illicit drug use doesn't have the same effect on infant health as does underreported use. They find the impact of exposure to illicit drug or cocaine on infant health is substantially less among under-reporters. Their results suggest that estimate of the effect of self-reported prenatal drug use on birth weight are biased upwards by non-random binary measurement error. Their results also imply that relatively heavy users are more likely to report their use, whereas light users tend to underreport their use. In order to account for those individuals who underreported but were not identified as such by urine tests, they use predicted measurement error in estimating infant health production function and get much better empirical results in terms of degree of bias.

Omitted variable model is very useful when one has good proxies for the actual variables or some information about measurement errors. The partial correlation between the observed and its measurement error can be used to adjust OLS estimator in order to get consistent one, and to reveal reporting styles of illicit drug use.

2.4. Proxy Variables

Welch (1975) shows that including some proxies for omitted variables doesn't necessarily give better estimates than the one that ignores the proxies. For example, in estimating the effect of schooling (s) and ability (a) on income (y), the model can be set up as follows.

$$y = \beta_1 \bar{s} + \beta_2 \bar{a} + u \qquad \textbf{(2-1)}$$

$$s = \bar{s} + v_1 \qquad \textbf{(2-2)}$$

$$a = \bar{a} + v_2 \qquad (2\text{-}3)$$

Where $\bar{s}, \bar{a}$ are true variables, u, v_1, v_2 are measurement errors, $b_{ys.a}$ is the coefficient from regressing y on s by holding a constant, b_{ys} is coefficient from regressing y on s without holding a constant. We can prove that plim $b_{ys.a}$ < plim b_{ys}. But this does not necessarily mean that the asymptotic bias in $b_{ys.a}$ is less than that in b_{ys}. If one puts more variables related to v_1, or v_2 against various possible biases due to omitted variables, the estimating results may be even worse.

2.5. Two Step Estimation (T-S)

When simultaneous equations include endogenous binary and latent variables, the coefficients we estimate measure association, and we can not give causal interpretations unless the model is full recursive. If the system fits logical consistency and is just or over identified, all parameters in the system are estimable. Technically speaking, if the models involve a large number of interdependent truncated variables, MLE is very cumbersome and in some cases even infeasible due to multiple integrals. In such cases, two step estimation methods can be used to produce consistent estimates. Maddala (1983) reviews two step estimation methods in a wide class of models involving censored and qualitative endogenous variables. He points out that two step methods have been found to create severe multicollinearity problems. He also suggests one to use the correct asymptotic covariance rather than the covariance from the second stage.

It is well known that T-S procedures yield consistent estimates of second step parameters under fairly general conditions, but estimated standard errors in second step are incorrect. So hypothesis tests based on the estimated covariance matrix of the second-

step estimation are biased, even in large sample. Murphy and Topel (1985) develop a method of correcting standard errors in T-S estimation by assuming that first step auxiliary model produces consistent estimates of parameters and their covariance, and those exogenous variables in second step are not contained in first step. They find that (1). The incorrect T-S standard errors are always asymptotically less than the Cramer-Rao low bound; (2). All estimated standard errors in the model are adjusted upward; (3). Proportional adjustment in estimated standard errors is largest for the imputed regressors; (4). In the case of 2SLS, all estimated standard errors are adjusted by the same factor.

As we know, IV method is a special case of T-S estimation. T-S estimation is appreciated in many cases in which IV is simply infeasible or can't be used. Heckman sample selection model, switch regression model and two steps Tobit etc. are examples of them. The advantage of the T-S estimation over IV estimation is that it exploits the model structure and associated constraints in estimating the reduced form. T-S estimation has sometimes been found to yield poor estimates of parameters due to multicollinearity between the independent variables in first step and that in second step. However, as compared to MLE, T-S estimation is more likely to be popular because of their computational simplicity.

In conclusions, for some recursive simultaneous equations system with endogenous latent and binary variables, we can get predicted probability in first step, and put it in second step in place of endogenous binary variable, then estimate second step equation by OLS. Of cause, standard errors in second step need to be corrected. From the view of simultaneous equations system, the coefficient on the binary observed variable in OLS and that on its predicted probability in structure system are comparable in their

magnitude. It should be cautioned that the interpretations for coefficients of both binary and its fitted values are different if one views the fitted variable as a proxy!

Chapter III
Bias, Inconsistency and Change in Variance Of the OLS Estimator Due to Measurement Error

Aigner (1973) considers the multiple regression model in which one independent variable may be subject to binary measurement error. He demonstrates that the direction of the proportionate bias is consistent with the usual result, and that $\hat{\beta}_{ols}$ is biased downward if one omits binary measurement error in the model. He can't draw a clear conclusion about efficiency of the estimator in the model by adopting a mean-squared-error (MSE) criterion. We should pursue more general conclusions regarding bias, inconsistency and change in variance for the coefficient of the variable measured with binary or continuous error.

3.1. Bias & Inconsistency

We set one identity and two competing population regression models as follows:

$$X = x + u \qquad (3\text{-}1)$$

$$y = z\gamma + X\beta + \varepsilon \qquad (3\text{-}2)$$

$$= z\gamma + x\beta + u\beta + \varepsilon$$

$$y = z\gamma + x\beta + u\alpha + \varepsilon \qquad (3\text{-}3)$$

Where y is a $n\times 1$ vector of dependent variable, z is a $n\times k_1$ matrix of independent variables without measurement errors, X, x are the $n\times k_2$ matrixes of the true variables and the observed variables with measurement errors respectively, u is a $n\times k_2$ matrix of measurement errors with $E(u_i)=0$ and $E(u_i u_i)=\sigma_{ui}^2 I$, and ε is a $n\times 1$ vector of error term

with $E(\varepsilon)=0$ and $E(\varepsilon'\varepsilon)=\sigma_\varepsilon^2 I$. We also assume that (1). All variables are measured as deviations from their respective mean values; (2). $Cov(u, \varepsilon)=0$; (3). The coefficients for x and u are not the same [1]. (4). ε is independent of all regressors [2]. It should be known that mean of measurement errors in non-deviation form and $cov(X,u)$ may not be zero.

We start with Equation (3-3) which is slightly deviated from Equation (3-2). Equation (3-3) is one more general population model mentioned by Greene (1993) [3].

From Equation (3-3) by ignoring measurement errors, OLS estimators will be:

$$\hat{\theta}=(w'w)^{-1}w'y \qquad (3\text{-}4)$$

Where

$$w'w=\begin{pmatrix} z'z & z'x \\ x'z & x'x \end{pmatrix} \quad w'y=\begin{pmatrix} z'y \\ x'y \end{pmatrix} \qquad \theta=\begin{pmatrix} \gamma \\ \beta \end{pmatrix}.$$

The corresponding sampling error in θ will be of the form:

$$\hat{\theta}-\theta=(w'w)^{-1}w'\varepsilon+(w'w)^{-1}w'u\,\alpha \qquad (3\text{-}5)$$

Assuming $\text{plim}\,(\frac{1}{n}w'\varepsilon)=0$, $\text{plim}\,(\frac{1}{n}w'w)=\Sigma_w$,

$$\text{plim}\,(\hat{\theta}-\theta)=\text{plim}\begin{pmatrix} c_{11}z'u+c_{12}x'u \\ c_{21}z'u+c_{22}x'u \end{pmatrix}\text{plim}\,\alpha \qquad (3\text{-}6)$$

Where $c_{11}=(z'z)^{-1}-c_{12}(x'z)(z'z)^{-1}$

$c_{12}=-(z'z)^{-1}(z'x)c_{22}$

$c_{21}=-c_{22}(x'z)(z'z)^{-1}$

$c_{22}=[(x'x)-(x'z)(z'z)^{-1}(z'x)]^{-1}$.

---

1. One can also assume that the coefficients for x and u are the same.
2. We use this assumption to eliminate contemporaneous correlation between ε and all regressors.
3. See W. Greene, Econometric Analysis, 2nd Edition, 1993, pp281, Formula (9-24').

After partitioning the matrix, the sample error in β will be:

$$\text{Plim}(\hat{\beta}-\beta)=\text{plim}\,[(x'x)-(x'z)(z'z)^{-1}(z'x)]^{-1}\ \text{plim}\,[(x'u)-(x'z)(z'z)^{-1}(z'u)]$$

$$\text{plim}\ \alpha$$

$$=\delta\alpha \qquad (3\text{-}7)$$

Where δ is a $k_2 \times k_2$ matrix of partial correlation between u and x.

For any of β, we have

$$\text{Plim}\,(\hat{\beta}_i-\beta_i)=\sum_{j=1}^{k_2}\delta_{ij}\alpha_j\,. \qquad (3\text{-}8)$$

Generally speaking,

if $\sum_{j=1}^{k_2}\delta_{ij}\alpha_j>0$ $\hat{\beta}_i$ is biased upwards and inconsistent [4];

if $\sum_{j=1}^{k_2}\delta_{ij}\alpha_j<0$ $\hat{\beta}_i$ is biased downwards and inconsistent;

if $\sum_{j=1}^{k_2}\delta_{ij}\alpha_j=0$ $\hat{\beta}_i$ is unbiased and consistent.

Now we begin to discuss several different cases:

1. **Kaestner et al. (1996)'s case**

If $k_2=1$, $\text{plim}\ \hat{\beta}=\beta+\alpha\delta_{ux}=\beta(1+\frac{\alpha}{\beta}\delta_{ux})$ (3-9)

If $\alpha=\beta$ $\text{plim}\ \hat{\beta}=\beta(1+\delta_{ux})$. (3-10)

2. **Aigner (1973)'s case**

[4] For unbiasedness, $E(\hat{\beta})=\beta$; For consistency, plim $\hat{\beta}=\beta$. Those conditions may not be held at the same time. But in this study, they are held at the same time by assumption (4).

From Equation (3-9) and (3-10), If $X = x - u$ and cov(x, z)=0,

$$\text{Plim } \hat{\beta} = \beta(1 - \frac{\text{cov}(x,u)}{\text{var}(x)}) \qquad (3\text{-}11)$$

Obviously, if cov(x,u) > 0 or cov(X,u) > -var(u), $\hat{\beta}$ is biased downwards. In his example, cov(x,u)=$(\nu + \eta)\widetilde{pq} > 0$, so the conclusion holds. Aigner (1973) only assumes that cov(z,u)=0, which is not sufficient to reach his conclusion. From Equation (3-7), one must assume that cov(x,z)=0. Otherwise his formula can't be derived.

3. Classical case

From Equation (3-7) and (3-10), If cov(x,z)=0 and cov(X,u)=0 [5],

$$\text{Plim } \hat{\beta} = \beta(\frac{\text{var}(X)}{\text{var}(x)})$$

$$= \beta(1 - \frac{\sigma_u^2}{\sigma_x^2}). \qquad (3\text{-}12)$$

When assuming cov(X,u)=0 in classical model, var(X)/var(x) < 1, which means that $\hat{\beta}$ is biased downwards. Aigner (1973) implies that if cov(X,u) > -var(u), $\hat{\beta}$ is biased downwards. The assumption of classical model automatically satisfies Aigner (1973)'s condition.

3.2. Bias Indexes

We suggest the following indexes of bias for the case of only one variable measured with error.

1. **General cases**:

5. X and u are highly correlated in NYC sample.

$$BI = \frac{\alpha}{\beta}\delta_{ux} \qquad (3\text{-}13)$$

$$BI = \frac{\hat{\beta}}{\beta} - 1 \qquad (3\text{-}14)$$

2. **Special cases:**

$$BI = \delta_{ux} \qquad (3\text{-}15)$$

$$BI = \frac{\mathrm{cov}(x,u)}{\mathrm{var}(x)} \qquad (3\text{-}16)$$

$$BI = \frac{\sigma_u^2}{\sigma_x^2} \qquad (3\text{-}17)$$

Where BI is bias indexes, The Formula (3-13) and (3-14) are for general cases. The rest are for special cases with more constrains. Basically, we can define the bias index as (gross effect/direct effect) –1. When we use Formula (3-14) at the empirical level, we need to caution it.

3.3. The Change in Variance

We have derived the formula for assessing bias and inconsistency. Bias direction depends on the sign of partial correlation between the observed and its error by assuming $\alpha = \beta$. If one ignores measurement error, what happens to the variance of the coefficient on the variable measured with error? Can we make correction for it? We assume $\mathrm{cov}(X,z) = 0, \mathrm{cov}(x,z) = 0$ in following derivations. n is the number of observations in the sample.

1. Case I (Excluding Measurement Error & without Restriction: $\alpha = \beta$)

Basing on Equation (3-5), we have

$$\hat{\theta}-\theta=(w'w)^{-1}w'(\varepsilon+u\alpha).$$

$$\text{plim Var}(\hat{\theta})=\text{plim E}[(\hat{\theta}-\theta)(\hat{\theta}-\theta)'] \tag{3-18}$$

$$=\text{plim}(w'w)^{-1}w'\text{E}[(\varepsilon+u\alpha)(\varepsilon'+u'\alpha')]w(w'w)^{-1}$$

$$\text{E}[(\varepsilon+u\alpha)(\varepsilon+u\alpha)']$$

$$=(\sigma_\varepsilon^2+\sum_{j=1}^{k_2}\alpha_j^2\sigma_{uj}^2)\,\text{I}_{n\times n} \tag{3-19}$$

Where

$$E(u_i'u_i)=\sigma_{ui}^2$$

$$E(u_i'u_j)=0,\ \text{for } i\neq j.$$

$$E(\varepsilon\varepsilon')=\sigma_\varepsilon^2\,\text{I}_{n\times n}$$

$$E(\varepsilon\alpha'u')=0_{n\times n}$$

$$E(u\alpha\varepsilon')=0_{n\times n}$$

$$E(u\alpha\alpha'u')=(\sum_{j=1}^{k_2}\alpha_j^2\sigma_{uj}^2)\text{I}_{n\times n}.$$

$$\text{Plim Var}(\hat{\theta})=(\sigma_\varepsilon^2+\sum_{j=1}^{k_2}\alpha_j^2\sigma_{uj}^2)\ \text{plim}\,(w'w)^{-1} \tag{3-20}$$

$$\text{Plim Var}(\hat{\beta})=(\sigma_\varepsilon^2+\sum_{j=1}^{k_2}\alpha_j^2\sigma_{uj}^2)\ \text{plim}\,[(x'x)-(x'z)(z'z)^{-1}(z'x)]^{-1} \tag{3-21}$$

If we assume $k_2=1$ and $\text{cov}(x,z)=0$,

$$\text{plim Var}(\hat{\beta})=[1+\frac{\alpha^2\sigma_u^2}{\sigma_\varepsilon^2}]\sigma_\varepsilon^2\,\text{plim}\,(x'x)^{-1} \tag{3-22}$$

$$=\frac{\sigma_\varepsilon^2+\alpha^2\sigma_u^2}{n\sigma_x^2}$$

2. Case II (Including Measurement Error & without Restriction: $\beta = \alpha$)

From the population model (3-3),

$$\text{Plim Var } (\beta) = \sigma_\varepsilon^2[(x'x) - (x'z)(z'z)^{-1}(z'x) - (x'u)(u'u)^{-1}(u'x)]^{-1}$$

$$= \frac{\sigma_\varepsilon^2}{n(\sigma_x^2 - \sigma_{ux}^2/\sigma_u^2)} \text{ if } \text{cov}(z,u) = 0, \text{cov}(x,z) = 0, k_2 = 1. \quad (3\text{-}23)$$

3. Compare Variances in Case I & II

According to Gauss-Markov Theorem, if there is no specification error in linear regression model, OLS estimators are BLUE among many unbiased estimators for the finite sample. Generally speaking, whether the probability limit of variance of the coefficient on the variable measured with error in Case I is bigger or smaller than that on the variable in the model without measurement error in Case II is indeterminate in the large sample. However, if one has the information (α, σ_u^2, ρ_{ux}) about measurement error, the direction and magnitude of the variance for biased OLS estimator can be determined. Let's compare the variances of β in Case I & II.

$$\text{Plim}\left(\frac{\text{var}_1(\hat{\beta})}{\text{var}_2(\hat{\beta})}\right)$$

$$= 1 + f(\rho_{ux}) \quad (3\text{-}24)$$

Where $a = \frac{\alpha^2\sigma_u^2}{\sigma_\varepsilon^2}$, $\rho_{ux} = \frac{\sigma_{ux}}{\sigma_u\sigma_x}$, $f(\rho_{ux}) = a - (a+1)\rho_{ux}^2$, $0 \le |\rho_{ux}| \le 1$. $\text{var}_1(\hat{\beta}), \text{var}_2(\hat{\beta})$ are from Case I & II respectively.

If $|\rho_{ux}| > \sqrt{\frac{a}{1+a}}$, then $f(\rho_{ux}) < 0$, the biased estimator in Case I has smaller variance; If $|\rho_{ux}| = \sqrt{\frac{a}{1+a}}$, then $f(\rho_{ux}) = 0$, the estimators in Case I & II have the same

variance; If $|\rho_{ux}| < \sqrt{\frac{a}{1+a}}$, then $f(\rho_{ux}) > 0$, the biased estimator in case I has larger variance. It is important to note that $f(\rho_{ux})$ can be used to correct variance of $\hat{\beta}$ by taking account of measurement error. Some evidences from simulations and economic data show that $f(\rho_{ux})$ is more likely negative, as shown on Table 7-5.

4. Compare MSE in Case I & II

Based on Central Limit Theorem, one can obtain asymptotic distribution of parameters and make statistical inferences. There is a trade-off between bias and efficiency under study. On average, the biased estimator will be closer to the true parameter than will the unbiased estimator. Some statisticians suggest using mean squared error (MSE) as a criterion in adjudging efficiency. MSE can be defined as follows:

$$\begin{aligned} \text{MSE}(\hat{\beta}) &= E(\hat{\beta}-\beta)^2 \\ &= E[(E(\hat{\beta})-\beta)^2] + E[(\hat{\beta}-E(\hat{\beta}))^2] \\ &= [\text{bias}(\hat{\beta})]^2 + \text{var}(\hat{\beta}) \end{aligned} \tag{3-25}$$

For Case I,

$$\begin{aligned} \text{plim } MSE_1(\hat{\beta}) &= \frac{\alpha^2\sigma_{ux}^2}{\sigma_x^4} + \frac{\sigma_\varepsilon^2 + \alpha^2\sigma_u^2}{n\sigma_x^2} \\ &= \frac{\sigma_\varepsilon^2 + \alpha^2\sigma_u^2(1+n\rho_{ux}^2)}{n\sigma_x^2} \end{aligned} \tag{3-26}$$

For Case II,

$$\text{plim } MSE_2(\hat{\beta}) = 0 + \frac{\sigma_\varepsilon^2}{n\sigma_x^2(1-\rho_{ux}^2)} \tag{3-27}$$

$$\text{plim}\left(\frac{MSE_1(\hat{\beta})}{MSE_2(\hat{\beta})}\right)=1+g(\rho_{ux}) \qquad (3\text{-}28)$$

Where $a=\frac{\alpha^2\sigma_u^2}{\sigma_\varepsilon^2}$, $\rho_{ux}=\frac{\sigma_{ux}}{\sigma_u\sigma_x}$, $g(\rho_{ux})=a+(an-a-1)\rho_{ux}^2-an\rho_{ux}^4, 0\leq|\rho_{ux}|\leq 1$.

There are two real roots for equation $g(\rho_{ux})=0$.

$$Root_{1,2}=\pm\sqrt{\frac{(an-a-1)+\sqrt{(an-a-1)^2+4a^2n}}{2an}} \qquad (3\text{-}29)$$

If $|\rho_{ux}|>|Root_{1,2}|$, then $g(\rho_{ux})<0$, the OLS estimator in Case I is more efficient; If $|\rho_{ux}|<|Root_{1,2}|$, then $g(\rho_{ux})>0$, the OLS estimator in Case II is more efficient; Otherwise, the estimators in Case I & II have the same efficiency. Which estimator will be more efficient depends on our knowledge of $\alpha,\sigma_u^2,\rho_{ux}$ from the sample. Griffiths et al. (1992) make a sampling experiment for correct-, under-, and over-specified models based on 1000 samples of data. Their finding suggests the OLS estimator in correct-specified model has smallest mean squared error; If a relevant right-hand side variable is excluded, then the estimator is biased but has improved sampling precision; If the right-hand side variables are over-specified, the estimator is unbiased but the sampling variances are inflated.

Unfortunately, the MSE is less demanding in reality because it depends on the unknown parameter. The more favorable criterion is minimum variance unbiasedness (or consistency). We usually use Cramer-Rao Lower Bound as a benchmark in evaluating asymptotic efficiency of the estimator for the large sample.

3.4. Statistical Inference

We usually use following t^2 statistics which is the same as $F(1,n-k)$ to make statistical inferences. Only one variable (x) is measured with error (u) and z is measured without error in following models. We assume $\text{cov}(x,z)=0, \text{cov}(X,z)=0$. n is the number of observations in the sample.

1. **Model I (Excluding Measurement Error and without Restriction: $\beta=\alpha$)**

$$y=z\gamma+x\beta+u\alpha+\varepsilon=z\gamma+x\beta+v$$

For null hypothesis: $\beta=0$,

$$\text{plim}\,T_b^2=\frac{n\beta^2\sigma_x^2(1+\frac{\alpha\sigma_{ux}}{\beta\sigma_u^2})^2}{\sigma_\varepsilon^2+\alpha^2\sigma_u^2} \qquad (3\text{-}30)$$

2. Model II (Including Measurement Error and without Restriction: $\beta=\alpha$)

$$y=z\gamma+x\beta+u\alpha+\varepsilon$$

For null hypothesis: $\beta=0$,

$$\text{plim}\,T_\beta^2=\frac{n\beta^2(\sigma_x^2-\frac{\sigma_{ux}^2}{\sigma_u^2})}{\sigma_\varepsilon^2} \qquad (3\text{-}31)$$

When we compare T_β^2 in Model II to T_b^2 in Model I, which one will be more statistically significant in testing null hypothesis: $\beta=0$ is unknown. If we make two more key assumptions: (1). $\text{cov}(X,u)=0$; (2). $\beta=\alpha$, We will have the following expression from Mode I & II:

$$\text{plim}\left(\frac{T_\beta^2}{T_b^2}\right)=\left[\frac{\sigma_\varepsilon^2+\beta^2\sigma_u^2}{\sigma_\varepsilon^2}\right]\left[\frac{1}{1-r^2}\right]>1 \qquad (3\text{-}32)$$

Where $r^2 = \frac{\sigma_u^2}{\sigma_x^2}$. Equation (3-32) [6] indicates that the null hypothesis that the coefficient of the variable measured with error equals zero is more likely to be accepted than the null hypothesis that the coefficient of that variable in the model without omitted measurement error equals zero.

3.5. The Bias and Change in Variance of the Restricted OLS Estimator in the Finite Sample

In dealing with consistency and asymptotic variance of the OLS estimator in the large sample, all regressors including measurement error are assumed to be stochastic. This Section is only focused on the finite sample properties of the restricted OLS estimator. We make one classical assumption that all regressors in the model are non-stochastic. From Bayesian views, the correct prior information may increase efficiency of OLS estimator in terms of MSE criterion. If one imposes some restrictions on the regression parameters, the estimates will be biased, and have smaller variances. The biased estimator may be closer to the true parameter. If we omit non-stochastic measurement error in the model, what happen to the bias and variance of OLS estimator? Does the restricted OLS estimator always have smaller variance than the unrestricted one? Now lets view those questions from a different angle.

Suppose that we have a correct-specified model with certain restrictions:

$$\begin{aligned} y &= z\gamma + x\beta + u\alpha + \varepsilon \\ &= w\theta + \varepsilon \qquad (3\text{-}33) \\ \text{ST:} & \; r\theta = q . \end{aligned}$$

[6] The original formula is derived by M. Grossman in "A Note on t^2 Ratio" in Mar. 1998.

Where $w=(z,x,u), \theta=(\gamma,\beta,\alpha)'$, $r\theta=q$ is a set of restrictions, r is a $j\times k$ matrix, y is the dependent variable measured without error, z is other regressors measured without errors, x is the variable measured with error u. One should note that excluding u is equivalently to set restriction $\alpha=0$ in Equation (3-33).

If we do not impose restrictions,

$$\hat{\theta}=(w'w)^{-1}w'y$$

$$E(\hat{\theta})=\theta \qquad (3\text{-}34)$$

$$\text{var}(\hat{\theta})=\sigma_\varepsilon^2(w'w)^{-1} \qquad (3\text{-}35)$$

If we do impose restrictions: $r\theta=q$,

$$\hat{\theta}_\bullet=\hat{\theta}-(w'w)^{-1}r'(r(w'w)^{-1}r')^{-1}(r\hat{\theta}-q)$$

$$E(\hat{\theta}_\bullet)=\theta-A(r\theta-q) \qquad (3\text{-}36)$$

Where $A=(w'w)^{-1}r'(r(w'w)^{-1}r')^{-1}$. From Equation (3-36), the restricted OLS estimator is biased. If $A(r\theta-q)>0$, $\hat{\theta}_\bullet$ is biased downwards. If $A(r\theta-q)<0$, $\hat{\theta}_\bullet$ is biased upwards. Otherwise, $\hat{\theta}_\bullet$ is unbiased. The direction of bias depends on the data under study. These conclusions are the same as what we get in Section 3.1. For example, if we impose $\alpha=0$, we have the restricted OLS estimator: $\hat{\beta}_\bullet=\beta+\alpha\delta_{wx}$. Now we consider the difference in variance due to the restriction.

$$\text{var}(\hat{\theta}_\bullet)=\text{var}(\hat{\theta})-\sigma_\varepsilon^2 Ar(w'w)^{-1} \qquad (3\text{-}37)$$

$Ar(w'w)^{-1}$ is a positive and definite matrix. So the variances of restricted OLS estimators are always smaller than that of unrestricted one. This conclusion is not the same as we draw in Section 3.3, but it is strongly supported by empirical evidences,

seeing Table 7-5. The crucial reason is that if we put non-stochastic measurement error in equation error term ($e = u\alpha + \varepsilon$), we have $\sigma_e^2 = \sigma_\varepsilon^2$, whereas $\sigma_e^2 = \sigma_\varepsilon^2 + \alpha^2\sigma_u^2$ if u is a stochastic variable.

Chapter IV
Empirical Estimation Issues on Measurement Errors

We have theoretically derived some formulas for evaluating bias, inconsistency, Change in variance and statistical inference due to measurement error. This Chapter will show how to impose restriction: $\beta = \alpha$ and how to hold measurement error constant at empirical level. It will also demonstrate how the gross and direct effects of the observed illicit drug use on birth weight are affected by this restriction, and explain why the coefficient on the observed by excluding measurement error and the coefficient on the actual are not comparable in assessing the bias of OLS estimator. In the following discussions, y is a dependent variable, z is an independent variable measured without error, x and X are the observed and the actual variables respectively, u is the measurement error that is $X - x$. We assume that $\text{cov}(x, z) = 0$, $\text{cov}(X, z) = 0$, $\text{cov}(u, \varepsilon) = 0$, and ε is independent of all regressors.

4.1. Regress y on z and x (without Restriction: $\beta_1 = \alpha$ & Excluding Measurement Error)

$$y = z\gamma + x\beta_1 + u\alpha + \varepsilon$$

$$= z\gamma + x\beta_{1ols} + e \qquad (4\text{-}1)$$

$$\text{plim } \hat{\beta}_{1ols} = \beta_1 (1 + \frac{\alpha}{\beta_1} \delta_{ux})$$

Where $\hat{\beta}_{1ols}$ is gross effect of x on y, β_1, α are direct effects of x, u on y.

If we run Model (4-1) by using SAS or other statistical software, we get gross effect, not direct effect because we do not hold u constant. As we discuss in Chapter III, $\hat{\beta}_{1ols}$ is

inconsistent and biased downwards due to measurement error.

4.2. Regress y on z, x and u (without Restriction: $\beta_1 = \alpha$ & Including Measurement Error)

$$y = z\gamma + x\beta_1 + u\alpha + \varepsilon \qquad (4\text{-}2)$$

$$u = z\delta_{uz} + x\delta_{ux} + v \qquad (4\text{-}3)$$

$$\text{plim } \hat{\beta}_1 = \beta_1$$

$$\text{plim } \hat{\alpha} = \alpha$$

Where $\hat{\beta}_1, \hat{\alpha}$ are consistent and unbiased estimates of β_1, α respectively. They are the direct effects of x, u on y. It should be noted that we do not impose restriction: $\beta_1 = \alpha$ here. It means that we allow x, u have different effects on y. In the study of the effect of illicit drug use on birth outcome, the reported and underreported use may not have the same effect on birth weight. If we regress u on z, x, we can obtain partial correlation between u and x, that is $\hat{\delta}_{ux}$ in Model (4-3). Model (4-2) is a population regression model without specification error in it. Now we check the relationship among $\hat{\beta}_{1ols}$ from Model (4-1), $\hat{\beta}_1, \hat{\alpha}, \hat{\delta}_{ux}$ from Model (4-2) and (4-3). As empirical results show, $\hat{\beta}_{1ols} = \hat{\beta}_1 + \hat{\alpha}\hat{\delta}_{ux}$.

4.3. Regress y on z and X (with Restriction: $\beta_2 = \alpha$ & Including Measurement Error Implicitly)

$$y = z\gamma + X\beta_2 + \varepsilon \qquad (4\text{-}4)$$

$$\text{plim } \hat{\beta}_2 = \beta_2$$

$\hat{\beta}_2$ is a consistent and unbiased estimator of β_2. If we run Model (4-4), we get direct effect of x or true effect of X on y. It is important to recognize that there is an implicit restriction: $\beta_2 = \alpha$ behind the Model (4-4) as we show in next section. Here $\hat{\beta}_2$ is the same as $\hat{\beta}_2$ in Model (4-5).

4.4. Regress y on z, x and u (with Restriction: $\beta_2 = \alpha$ & Including Measurement Error Explicitly)

$$y = z\gamma + x\beta_2 + u\beta_2 + \varepsilon \qquad (4\text{-}5)$$

$$= z\gamma + X\beta_2 + \varepsilon$$

$$\text{plim } \hat{\beta}_2 = \beta_2$$

Model (4-5) is equivalent with Model (4-4) if we assume that x and u have the same effect on y. In order to impose this restriction at empirical level, We need to run restricted OLS. $\hat{\beta}_2$ is direct effect of x, u on y in Model (4-5).

4.5. Regress y on z, and x (with Restriction: $\beta_2 = \alpha$ & Excluding Measurement Error)

$$y = z\gamma + x\beta_2 + u\beta_2 + \varepsilon$$

$$= z\gamma + x\beta_{2ols} + \omega \qquad (4\text{-}6)$$

$$\text{plim } \hat{\beta}_{2ols} = \beta_2(1 + \delta_{ux})$$

Here $\hat{\beta}_{2ols}$ is biased downwards and inconsistent due to measurement error. Can we get $\hat{\beta}_{2ols}$ by running Model (4-6)? The answer is no. If we run Model (4-6), We obtain $\hat{\beta}_{1ols}$ in Model (4-1), not $\hat{\beta}_{2ols}$ in Model (4-6). The reason is that if we include

measurement error in equation error term in Model (4-6), we can't impose restriction that x, u have the same effect on y. $\hat{\beta}_{2ols}$ in Model (4-6) and $\hat{\beta}_2$ in Model (4-5) are the gross effect and direct effect of x on y under the restriction. By running Model (4-3), $\hat{\delta}_{ux}$ can be obtained. The relationship that is $\hat{\beta}_{2ols} = \hat{\beta}_2(1+\hat{\delta}_{ux})$ should be held at empirical level also. In order to get $\hat{\beta}_{2ols}$, we have to impose restriction: $\beta_{2ols} = \hat{\beta}_2(1+\hat{\delta}_{ux})$ in Model (4-6) by running restricted OLS. The coefficient γ will be changed in contrast to that in Model (4-1).

4.6. Bias Indexes at Empirical Level

1. Bias index from Model (4-1) & (4-2) without the restriction: $\beta_1 = \alpha$

$$BI = \frac{\hat{\beta}_{1ols}}{\hat{\beta}_1} - 1 \qquad (4\text{-}7)$$

2. Bias index from Model (4-4), (4-5) &(4-6) with the restriction: $\beta_2 = \alpha$

$$BI = \frac{\hat{\beta}_{2ols}}{\hat{\beta}_2} - 1 \qquad (4\text{-}8)$$

Alternatively, one can use $\hat{\alpha}, \hat{\beta}_1, \hat{\delta}_{ux}$ to calculate Bias Indexes.

4.7 Summary

From the above discussions, we have raised some very important empirical estimation issues on measurement error. (1). It is easy to set $\beta = \alpha$ at theoretical level, but not that simple at empirical level, especially when one tries to impose inequality constraints or constraints across system equations. Whether or not to impose the restriction will have great impact on the absolute magnitude of effect of x on y. (2). How

to hold the measurement error constant in model at empirical level is to include it in model and to let it have some explanatory power, which is different from the concept at theoretical level. (3). The OLS estimator on the observed without restriction and that on the actual with restriction are not comparable in assessing bias for the coefficient of the observed given measurement error.

Chapter V
Measurement Errors in Health Production Function

5.1. Theoretical Models

Following household production theory created by G. S. Becker (1965), M. Grossman (1972) first constructs and estimates a model of demand for health. Health is a commodity that may directly or indirectly affect utility function. Medical care, own time, other healthy and unhealthy activities are inputs of health production. Health production function also depends on certain environmental variables such as schooling levels, age etc. In the content of infant health production, infants are commodities. Their parents can derive utility from quality of children. Infant birth outcomes are mainly a function of prenatal care, illicit drug use, alcohol and cigarette consumption, maternal physical conditions, nutrition, exercises, abortion and other infant and maternal characteristics. Pregnant women may get positive utility by using illicit drug, and disutility from fetus drug exposure.

The parents' utility function (U) and infant health production function (y) can be expressed as follows:

$$U = U(C, X, y) \qquad (5\text{-}1)$$

$$y = g(M, X, E) \qquad (5\text{-}2)$$

Where X is illicit drug use, M is prenatal care, C is other commodity, E is technological efficiency factor. We assume $\partial U / \partial C > 0$, $\partial U / \partial X > 0$, $\partial U / \partial y > 0$, $\partial y / \partial M > 0$ and $\partial y / \partial X < 0$ in this model. The full model is presented in Appendix I.

If we know the true infant health production function in that β is a coefficient on X, and y, M, X are measured with errors, we have

$$\beta = f(y, M, X, E) \qquad (5\text{-}3)$$

$$\overline{\beta} = f(\overline{y}, \overline{M}, \overline{X}, \overline{E}) \qquad (5\text{-}4)$$

$$\sigma_\beta = \sqrt{f_y^2\sigma_y^2 + f_M^2\sigma_M^2 + f_X^2\sigma_X^2}\,. \qquad (5\text{-}5)$$

Where f_y^2, f_M^2, f_X^2 should be evaluated at $(\overline{y}, \overline{M}, \overline{X}, \overline{E})$. Formula (5-5) shows how measurement errors are transmitted, and how they affect preciseness of the parameter in the system given that the true relationship among variables is fixed.

5.2. Empirical Models

As we present in the following models, y is birth weight, X_b is the actual binary illicit drug use ($X_b = x_b + u_b$), x_b is the observed binary illicit drug use, u_b is binary measurement error. z is a set of other regressors [7] measured without errors. We assume that $\text{cov}(u_b, \varepsilon) = 0$, $\text{cov}(u_b, x_b) \neq 0$, and ε is independent of all regressors.

1. Model 1 with the Binary Measure of the Actual Illicit Drug Use

$$y = z\gamma + X_b\beta + \varepsilon \qquad (5\text{-}6)$$

This is a population regression model. The estimates of γ, β are consistent and unbiased. $\hat{\beta}$ measures the treatment effect on the average users (or all users). Put differently, $\hat{\beta}$ indicates the mean group difference in birth weight between the users and

[7] Z includes msex, age1018, age3554, hsgrads, college, prenvis, plural, parity, alcohol, tobacco, hypert and induc. Variable definitions and descriptive statistics are in Table 6-2.

nonusers. If the drug program aims at all users, we can use $\hat{\beta}$ to do cost-benefit analysis for the program.

2. Model 2 with the Binary Measure of the Reported Illicit Drug Use

$$y = z\gamma + x_b\beta_{1ols} + e \qquad (5\text{-}7)$$

This is a sample regression model with specification error. $\hat{\beta}_{1ols}$ is biased and inconsistent due to omitted measurement error. $\hat{\beta}_{1ols}$ underestimates the treatment effect on the reported. If we use $\hat{\beta}_{1ols}$ to evaluate the drug program aimed at all users, the treatment effect will be overstated. The reason is that the impact of illicit drug use on infant health is less among women who deny prenatal use of illicit drugs but whose urine reflects exposure than among women admit illicit drug use during their pregnancies.

3. Model 3 with the Binary Measures of the Reported and the Unreported

$$y = z\gamma + x_b\beta_1 + u_b\alpha + \varepsilon \qquad (5\text{-}8)$$

This is a population regression model. $\hat{\beta}_1, \hat{\alpha}$ are unbiased and consistent. They measure the treatment effects on the reported and the unreported respectively. If we treat x_b, u_b as two kinds of illicit drug users in health production function, the coefficients for x_b, u_b can be different. If we have a drug program aimed at all users, we may use the weighted average of $\hat{\beta}_1$ and $\hat{\alpha}$ to assess it because the treatment effect on all users is bounded by $\hat{\beta}_1$ and $\hat{\alpha}$. The weighted average of $\hat{\beta}_1$ and $\hat{\alpha}$ is a better measure, and smaller than the coefficient $\hat{\beta}$ in Model 1 by taking account of the fact that the unreported have larger sampling weight than the reported do.

4. Model 4 with Treatment Effect on Heavy Users

Most of treatment programs are aimed at all users rather heavy users. M. Grossman (1996) [8] suggests applying the following models to get treatment effect on heavy users:

$$y = z\gamma + X_b\alpha + H\beta + \varepsilon \qquad (5\text{-}9)$$

$$y = z\gamma + L\alpha + H(\alpha + \beta) + \varepsilon \qquad (5\text{-}10)$$

Here $X_b = L + H$, H is a dummy variable for heavy illicit drug use, L is a dummy variable for light illicit drug use. Both Equation (5-9) and (5-10) are population regression models. The coefficient $(\alpha + \beta)$ in Equation (5-10) is most relevant in a cost–benefit analysis of drug treatment programs aimed at heavy users. If heavy users are not observable, One can obtain $\hat{\alpha}_{ols} = \alpha + \beta\delta_{HX}$ by regressing y on z, X_b with H omitted in Equation (5-9), where δ_{HX} is partial correlation between H and X_b, which is the ratio of the number of heavy users to the number of the total users ($\delta_{HX} < 1$). To evaluate these programs, one needs to compare $\hat{\alpha}_{ols}$ with $\hat{\alpha}$ and $(\hat{\alpha} + \hat{\beta})$. Generally Speaking, $\hat{\alpha}_{ols} > \hat{\alpha}$ and $\hat{\alpha}_{ols} < (\hat{\alpha} + \hat{\beta})$ in absolute values if $\hat{\alpha}$ and $\hat{\beta}$ have the same sign, which means that $\hat{\alpha}_{ols}$ underestimates the treatment effect on heavy users, and overestimates the treatment effect on all users or light users. Kaestner at el. (1996) argue that if one regresses y on z, x_b, u_b, the coefficients on x_b, u_b reflect the treatment effects on heavy users and light users respectively by assuming that light users are more likely to underreport their use, whereas heavy users are more likely to report their use. Comparing Equation (5-10) with Equation (5-8), one may use $\hat{\beta}$ and $\hat{\alpha}$ in Equation (5-8) to evaluate the drug programs aimed at heavy users and all users (or light users) respectively. It should be emphasized

[8] M. Grossman, Comments on Kaestner at el. (1996), 1996, File No: 2935.

that Model 3 and Model 4 are not equivalent in population because Model 4 assumes that all users and light users have the same treatment effects. However, Model 4 has its theoretical importance and policy relevance.

5.3. The Key Feature of Binary Measurement Error in Health production function

Suppose we have the following model:

$$y = z\gamma + x\beta + \varepsilon \qquad (5\text{-}11)$$

Where y is birth weight, x is a binary indicator of the reported or its measurement error. z is a set of other regressors which are assumed to be uncorrelatd with x. n is the number of observations in the sample. The OLS estimator of β is

$$\hat{\beta} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} \qquad (5\text{-}12)$$

$$= \frac{\bar{y}_1 - \bar{y}}{1 - \bar{x}}$$

Where $\bar{x} = n_1 / n$, $\bar{y}_1$ is the group mean of birth weight for the reported or the underreported. We assume that $\bar{y}_1 < \bar{y}$, so $\hat{\beta} < 0$. Intuitively speaking, as the number of reporters or underreporters (n_1) increases, $\bar{x}$ will increases, and $|\hat{\beta}|$ will increases, too. At the same time, as n_1 increases, $\bar{y}_1$ will increases, $|\hat{\beta}|$ will decreases at a great speed. The net effect of change in n_1 will pull down the coefficient in absolute value. Of cause, the coefficient in some cases can be pulled up or unchanged as n_1 goes up. This finding implies that the magnitude of the coefficient on binary indicator is not obviously associated with the prevalence of reporting or underreporting, especially in the context of

multiple regression models. For example, the coefficient ($|\hat{\beta}|$) for the group with high underreporting rate may be smaller than that for the group with low underreporting rate.

Chapter VI
Data Sets, Key Variables
And Some issues in Creating Measurement Errors

6.1. CA Data Set

CA urine test sample is a large, random and weighted sample. It comes from an epidemiologic cross-sectional study conducted under Department of Alcohol and Drug Programs in the State of California. The main purpose of the study is to obtain a 1992 estimate of prevalence of prenatal drug exposure by specific drug type for both the State of California as a whole and for the major regions within the state. Urine specimens and demographic information were collected anonymously from 29,494 pregnant women between July, 1991 to June, 1992 in 202 maternity hospitals throughout the state. It was hoped that the urine test data would provide an accurate population-based estimate of the number of substance-exposed infants born in California, and thereby provide a basis for program evaluations and clinical and community-based interventions. This data set has a total 46 variables which include the presence of licit and illicit substances and the demographic characteristics of women who were most at-risk for such deliveries. The urine toxicology screening was restricted to maternity patients being admitted for delivery. Therefor, estimates derived from testing at time of hospitalization could have varied from the true prevalence of substance exposure throughout the course of pregnancy and may underestimate the actual prevalence of drug use earlier in the pregnancy. The urine test may also have technical screening errors (ex: false negative or positive). But we make an assumption that the urine test reflects the true prevalence of illicit drug use among pregnant women. There are no key variables such as birth weight,

self-reported illicit drug use, prenatal visits, schooling levels and etc. in this data set. In order to use information of true illicit drug use in this sample to predict actual illicit drug use in NYC sample, we select 8406 pregnant women from 15 large cities with total population around 8 millions in California. The definition of some key variables and their descriptive statistics are shown in Table 6-1.

6.2. NYC Data Set

NYC birth certificate sample was documented by Department of Health of New York City in 1992. It is a large, random and unweighted sample with 126191 observations and about 150 variables. It does not include foreigners who do not live in NYC in this sample. Cocaine, Heroin, Marijuana, Methadone and other drugs were either reported by pregnant women or diagnosed by physicians when pregnant women were hospitalized for delivery. There is no full information about actual illicit drug use in NYC sample. In addition, birth weight, infant mortality, APGAR score, prenatal visits, early prenatal care, spontaneous and induced abortion, alcohol and cigarette consumption, physical conditions, parity, plurality, marital, employment and financial statuses, race, age, education, mother birth place, and participation in welfare programs etc. are reported in the sample. Self-reports regarding illicit drug use were believed to seriously underestimate the true prevalence of illicit drug use among pregnant women. The definitions of key variables and their descriptive statistics are listed in Table 6-2.

6.3. Key variables

1. illicit drug use

A drug is any chemical substance that produces a physical, behavioral, emotional or mental change in the user. Any drug can be harmful. Even both OTC and prescriptive drug have potential for becoming addictive. Illicit drug refers to Cocaine, Marijuana and Heroin etc. that are illegal in use. Some licit drugs such as Methadone, Tranquilizers, Sedatives, Amphelamines and Painkillers can be abused or misused, which may have negative effect on fetus. Both CA urine test sample and NYC birth certificate sample don't have quantity of illicit drug use, in stead of binary indicators showing whether pregnant women use illicit drug or not. Illicit drug is an input of infant health production. We define several versions of illicit drug use:

(a). X_b and x_b are binary actual and observed illicit drug use respectively.

(b). X_p and x_p are continuous predicted probability of the actual and the observed illicit drug use respectively.

(c). X and x are the actual and the observed illicit drug use being either binary or continuous .

(d). X^* and x^* are the actual and the observed quantity of illicit drug use which is not known by the researcher.

2. Measurement Errors

We define binary and continuous measurement errors in NYC sample as follows:

$$u_b = X_b - x_b \qquad (6\text{-}1)$$

$$u_p = X_p - x_p \qquad (6\text{-}2)$$

$$u = X - x \qquad (6\text{-}3)$$

$$u^* = X^* - x^* \qquad (6\text{-}4)$$

Where u_b and u_p are binary and continuous variables respectively. u is a general form of measurement error. u^* is continuous and uncensored latent measurement error. If $u_b < 0$, we set $u_b = 0$, and $X_b = 1$, which means that there is no over-reporting. Similarly, if $u_p < 0$, then we add absolute u_p to X_p and set $u_p = 0$.

3. Birth Weight

Birth weight, IMR, APGAR score can be defined as infant birth outcome, but birth weight is a more powerful predictor of birth outcome. We use birth weight as the dependent variable in infant health production function.

4. Prenatal Visit

Prenatal visit is treated as a proxy for prenatal care because there is no good measure for quality of prenatal care. It is a main input in infant health production function. We assume that prenatal visit is an exogenous variable in infant health production function.

5. Maternal Race

Maternal race is the key factor in constructing sampling weight. It is classified as Hispanics, Blacks (No-Hispanics), Whites (No-Hispanics), Asians (No-Hispanics) and Other (No-Hispanic). We split both CA sample and NYC sample into race-specific subsamples.

6.4.Validation for Using Out-of-Sample Information

If there are two large random samples, which may come from the same population or different populations, we need to make parametric or non-parametric statistical inferences before using out-of-sample information. Especially, one prefers to test equal means, variances or distributions by using F, t and χ^2 statistics. If two samples are from the

same population, information extracted from one sample has very strong prediction power to the other. We can treat those two samples as split samples with same population mean, variance, and even distribution. In reality, there are measurement errors and sampling errors in economic data. It is difficult to find two samples with the same multivariate distribution. Some confounding variables are unobservable. One should keep this caveat in mind. If two samples are not from the same population, it will be better to apply the bootstrap method for reshaping the target sample based on information form the source sample.

When Probit coefficients of actual illicit drug use in CA are used to predict probability of actual illicit drug use in NYC, we face validation issues for using out-of-sample information. Urine test sample in CA and birth certificate sample in NYC are assumed to be from the same population. The reasons are (1). All people in both samples are pregnant women; (2) They live in large cities which cover 8 million people each in CA and NYC; (3). Some key variables have same variances in both samples. Table 6-3 indicates that key variables have different distributions in CA and NYC samples, and their means are not same except for White dummy in both samples. Blacks, Hispanics and Whites are 12.2%, 50.19% and 23.2% respectively in CA sample, 31.5%, 33.7% and 25.8% respectively in NYC sample. One may see that compared to CA sample, Blacks are over-sampled, whereas Hispanics are under-sampled in NYC data set. We know races are an important factor in building sample weight, which corrects for sampling errors. Race-specific Probit coefficients in CA are used to predict race-specific probability of illicit drug use in NYC. Table 6-4 shows that for all races sample, one can reshape it with replacement for given proportion of Blacks, Hispanics and Whites in CA sample. For

example, 50.1% in CA sample are Hispanics, in order to keep same percentage of Hispanics in NYC sample, we need to randomly draw 63222 observations from an original pool of 42490 observations with replacement. We can do the same thing for other race groups. The newly shaped NYC data set has the same size as before, but proportion of races has been changed. This is the way to reduce sampling error in NYC data set. Table 6-5 demonstrates how to simulate the coefficients in health production function. Lets take Blacks group as an example. We know the original sample size is 40074 and the reshaped size is 15395 in Table 6-4. The Bootstrap method can be used to randomly draw 15395 observations from 40074 observations with replacement. For the drawn sample, we can estimate health production function for Blacks, and get the coefficient on the reported. After doing the same looped process as many as we like, we obtain a sample of coefficients, then the sample mean and standard deviation of the coefficient on the reported can be computed. Compared the coefficient (-324.65) without reshaping to that (-322.68) with reshaping for Blacks, one can conclude that as the number of loop increases, the difference between them will vanish gradually.

6.5. Converting Method of Equivalent Mean (EQM)

How to convert a continuous predicted probability into binary indicator did not cause econometricans' attention. At empirical level, we need to choose reasonable threshold value (X_{pt}), so that

$$X_b = 1 \quad \text{if } X_p \geq X_{pt} \qquad (6\text{-}5)$$

$$X_b = 0 \quad \text{otherwise.}$$

Where $X_{pt} = cm_p$, m_p is mean probability, c is any positive number.

Greene sets cm_p=50% in his LIMDEP software. According to his setting, true prevalence of illicit drug use will be underestimated heavily. Some researchers suggest using m_p as the threshold value. The result shows that true illicit drug use will be overestimated greatly. The mean of X_b (m_b) indicates percentage of people who use illicit drug. As sample size increases, m_b should converge to m_p. We select the value of c by trial and errors so that m_b is equal to m_p, which is so called equivalent mean method (EQM). If we know predicted probability of underreporting, we can convert it into a binary indicator in the same way.

Table 6-6 presents converted binary actual illicit drug use based on different threshold values. For example, according to EQM, 2.52% of Hispanic pregnant women use illicit drug. If we use other two criterions, one predicts that no one uses illicit drug, whereas the other predicts that 32.9% of Hispanic pregnant women use illicit drug. Both methods do not reflect the true prevalence of illicit drug use among Hispanic pregnant women .

6.6. Logical or Algebraic Mapping for Measurement Errors

Suppose that u_j, j=1 to k, is a binary measurement error of Cocaine or Heroin or Marijuana, we can use logical mapping method to create a summary variable called underreporting errors in illicit drug use. For example,

$$u_b = u_1 \cup u_2 \cup ..... \cup u_k \qquad (6\text{-}6)$$

$$u_b^* = u_1 \cap u_2 \cap ..... \cap u_k \qquad (6\text{-}7)$$

Where u_b=1 indicates one with at least one underreporting error, u_b^* =1 indicates one with multiple underreporting errors.

If u_j, j=1 to k, is a continuous measurement error of each illicit drug, and all u_j are correlated, we may apply principal component method to summarize those error information. There are k principal components, which are orthogonal among them. Each principal component is a linear combination of the original variables with coefficients equal to the eigenvectors (α) of the covariance matrix (Σ_u). The eigenvectors are customarily taken with unit length. The principal components are sorted by descending order of the eigenvalues (λ), which are equal to the variances of the components. We can choose one component, which summarizes the original variables best.

$$\Sigma_u \alpha_i = \lambda_i \alpha_i \qquad i = 1.....k \qquad (6\text{-}8)$$

$$u_{ip} = \sum_{j=1}^{k} \alpha_{ij} u_j \qquad i = 1.....k \qquad (6\text{-}9)$$

Where u_{ip} is summarizing variable – principal component, u_j is the original variable, α_i is an eigenvecoter.

The question here is how original variables affect those summarizing variables through logical or algebraic mapping. At empirical level, one can code continuous variables into binary variables, and summarize continuous and binary variables respectively. Which way will be better depends on research purpose under study. It is obvious that coding will lose information in the data. Generally speaking, logical mapping is much powerful in the world with large qualitative information. Algebraic mapping is a valuable tool for transformation from one structure to another.

6.7. Main Steps in Creating Measurement errors

1. Use F,t and χ^2 statistics to check out whether CA and NYC samples are from the same population or two populations have the same variances, means and distributions.

2. Get the coefficients (φ) from Probit model based on CA data set. The dependent variable is X_b. The independent variables (z) in race-specific Probit model include mothers' age, marital status, birthplace, financial status and early prenatal care. Table 6-7 shows that mother age, marital status, Medicaid and early prenatal care have very strong explanatory power in predicting probability of illicit drug use.

3. Use φ from CA Probit model and z from NYC data set to get index: $I = z\varphi$; z is a set of individual characteristics which is the same as that in Step 2.

4. Compute predicted probability of actual illicit drug use: $X_p = \Phi(I)$ in NYC sample;

5. Convert predicted probability of actual illicit drug use into a binary indicator (X_b) basing on certain threshold values;

6. Predict the probability of self-reported illicit drug use (x_p); We use x_b as dependent variable, and the same z as in Step 2 or 3 as independent variables.

7. Calculate measurement errors (u_b, u_p) by applying Formula (6-1) and (6-2); We also make adjustments for X,u in considering overreporting.

Table 6-1
Definitions and Descriptive Statistics of Key Variables in CA Sample

Variable	Definition	Mean	Std Dev
Xb	1 if actually using illicit drug during preganac	0.040	0.196
anydrug	1 if using licit or illicit drug during pregnancy	0.062	0.242
alcohol	1 if drinking alcohol during pregancy	0.072	0.258
tobacco	1 if smoking cigarettes during pregnancy	0.082	0.274
agemoth	mother age when delivering	26.67	6.159
age1018	1 if 10 <= mother age <= 18	0.079	0.270
age3554	1 if 35 <= mother age <= 54	0.116	0.320
married	1 if one is married	0.572	0.495
bornus	1 if one is born in U.S.	0.456	0.498
black	1 if one is Black, no Hispanic	0.122	0.328
hisp	1 if one is Hispanic	0.501	0.500
whitenh	1 if one is White, no Hispanic	0.232	0.422
asian	1 if one is Asian, no Hispanic	0.078	0.268
selffin	1 if financed by oneself	0.025	0.158
medicaid	1 if one is in Medicaid Program	0.537	0.499
trimest1	1 if prenatal care begins at first trimester	0.519	0.499
N	8406		

Note:
1.Mean,Std Dev are calculated by using sample weight.
2.The sample includes pregnant women in 15 large cities in CA.
3.This is an urine test sample.

Table 6-2
Variable Definitions and Descriptive Statistics in NYC Sample

Variable	Definition	Mean	Std Dev
xb	1 if reporting illicit drug use during pregnancy	0.018	0.132
ub	1 if underreporting illicit drug use	0.055	0.229
xp	predicted probability of the observed illicit drug use	0.018	0.032
up	predicted probability of underreporting illicit drug use	0.044	0.051
Xb	1 if actually using illicit drug	0.063	0.244
Xp	predicted probability of actual illicit drug use	0.062	0.077
y	infant birth weight in grams	3244.1	615.9
agemoth	mothers' age	27.367	6.137
age1018	1 if 10 <= mother age <= 18	0.073	0.260
age3554	1 if 35 <= mother age <=54	0.135	0.341
married	1 if getting married	0.526	0.499
bornus	1 if being born in U.S.	0.505	0.500
black	1 if mother's race is Black, and no Hispanic	0.315	0.465
hisp	1 if mother's race is Hispanic	0.337	0.473
whitenh	1 if mother's race is White and no Hispanic	0.258	0.437
selffin	1 if giving birth is self-financed	0.083	0.225
medicaid	1 if one is in Medicaid program	0.518	0.499
trimest1	1 if one starts prenatal care in first trimester	0.444	0.497
hsgrads	1 if one is a high school graduate	0.384	0.486
college	1 if one is a college graduate	0.350	0.477
prenvis	the number of prenatal visits during pregnancy	8.940	4.159
plural	1 if one produces more than one infant	0.026	0.159
parity	the number of living chidren so far	1.150	1.460
alcohol	1 if one drinks alcohol	0.009	0.096
ghypert	1 if one has gestational hypertension	0.013	0.113
induc	1 if one has induced abortions	0.235	0.424
msex	1 if the infant is male	0.511	0.499
AFDC	1 if one is in AFDC programe	0.315	0.465
WIC	1 if one is in WIC programe	0.022	0.147
emplyd	1 if one is employed during pregnancy	0.300	0.458

Note:
1. Variable definition is for both CA and NYC samples.
2. Descriptive statistics are only from NYC sample.
3. The number of observations is 126191.

Table 6-3
F, t &Chisq Tests for Equal Variances, Means & Distributions Between CA and NYC samples

Variable	Nca	Nny	Prob>F Ho:Equal Variance	Prob>t Ho: Equal Mean	Prob>Chisq Ho: Equal Distribution
agemoth	8406	126122	0.8378	0.0000	0.001
bornus	8406	126191	0.7790	0.0000	0.001
married	8406	126191	0.0873	0.0000	0.001
selffin	8406	126191	0.0000	0.0000	0.001
medicaid	8406	126191	0.8797	0.0185	0.001
trimest1	8406	126191	0.8638	0.0000	0.001
black	8406	126191	0.0000	0.0000	0.001
hisp	8406	126191	0.0000	0.0000	0.001
whitenh	8406	126191	0.9786	0.9845	0.001

1. Definitions of variables in Table 6-1.
2. Nca and Nny are the number of observations.

Table 6-4
Bootstrap Method for Resampling NYC Data Set

Race	N in CA Sample	Percentage	N in NYC Sample	Reshaping Size
Blacks	1336	12.2	40074	15395
Hispanics	3438	50.1	42490	63222
Whites	2168	23.2	32524	29276
Other	1464	14.5	11103	18298
All	8406	100	126191	126191

Note:

1. N is the number of observations.
2. NYC data set is a large random sample.
3. CA sample includes pregnant women in 15 large cities in which the total population is 8 millions.
4. Percentage is obtained by using CA sample weight.

Table 6-5
Simulating the Gross Effect
of Illicit Drug Use on Birth Weight in NYC
(The number of Loop=20)

Variable	Blacks	Hispanics	Whites
xb	-322.68 (27.66)	-347.43 (24.20)	-399.42 (46.75)
xp	-1274.15 (110.90)	-1978.55 (118.89)	-2097.52 (199.20)
N1	15395	63222	29276
xb*	-324.65 #	-349.99 #	-396.132 #
xp*	-1284.998 #	-2009.545 #	-2097.385 #
N2	40074	42490	32524

Note:

1. N1, N2 are the subsample size;

2. * indicates no simulation;

3.# indicates significance at 5% confidence level;

4. Standard deviations in brackets.

Table 6-6

Means for Binary or Continuous Measures of Illicit Drug Use in NYC

Variables	All	Blacks	Hispanics	Whites
xb	1.78 (13.2)	3.61 (18.6)	1.34 (11.5)	0.62 (7.80)
xp	1.78 (3.20)	3.60 (4.00)	1.34 (1.90)	0.62 (1.50)
Xb	6.30 (24.3)	15.0 (35.6)	2.52 (15.6)	4.70 (21.1)
Xp	6.20 (7.70)	15.0 (13.5)	2.56 (2.90)	4.69 (3.70)
Xm	31.6 (46.5)	36.7 (48.1)	32.9 (47.0)	36.3 (48.1)
X50	0.02 (1.30)	2.84 (16.6)	0.00 (0.00)	0.00 (0.00)
N	126191	40074	42490	32534

Note:

1. All numbers are in percentage.
2. Standard errors in brackets.
3. Xm -- A binary indicator when threshold = mean probability.
4. X50 -- A binary indicator when threshold = 50%.
5. N is sample size.

Table 6-7
Coefficinets from Probit Model Using the CA Sample
(Dependent Variable Xb=0,1)

Variable	Black	Hisp	White	All Races
intercept	-3.43020 #	-3.40773 #	-1.69351 #	-1.69351 #
agemoth	0.07705 #	0.03754 #	0.00911	0.04002 #
married	-0.54209 #	-0.41686 #	-0.34764 #	-0.45613 #
bornus	0.45141 ##	0.75857 #	0.12962	0.61126 #
selffin	0.62146 ###	0.25446	0.12389	0.30471 ##
medicaid	0.34901 #	0.37069 #	0.39017 #	0.38659 #
trimest1	-0.53999 #	-0.24208 #	-0.38512 #	-0.41871 #
black				0.85264 #
hisp				0.17321
whitenh				0.61089 #
N	1336	3438	2168	8406

Note:

1. # , ##, ### indicate significance at 5%, 10%, 15%.
2. Variable definitions in Table 6-1.
3. N is the number of observations in the samples.

Chapter VII
Empirical Results for Binary Measurement Errors in Health Production Function

7.1. Discussing Findings about Bias of OLS Estimator

In Table 7-1, the gross effect is the treatment effect on the reported with omitted measurement error, the direct effects are the treatment effects on the reported and the unreported respectively, and the actual effect is the treatment effect on all users. We do not impose the restriction that the effects of the observed and its error on birth weight are the same in Table 7-1. If one omits binary measurement error, OLS estimators of illicit drug use are biased downwards by 8.8% for all races group, 10.1% for Blacks, 3.7% for Hispanics, 2.5% for Whites respectively. This finding implies that the costs and consequences of prenatal illicit drug use based on hospital discharge data and vital records probably have been underestimated if one use OLS estimators to evaluate drug program aimed at the reported.

Compared with Table 7-1, Table 7-2 indicates all race-specific gross and direct effects from restricted OLS estimation. One needs to impose restriction: $\hat{\beta}_{ols} = \beta(1 + \delta_{ux})$ in Equation (5-7) for getting gross effect, and to impose restriction: $\beta = \alpha$ in Equation (5-8) for getting direct effect which is the same as the actual effect. Under these restrictions, bias index is the same as partial correlation between the observed illicit drug use and its error. The restricted OLS estimators are biased downwards by 19.76% for all races group, 32.25% for Blacks, 7.62 for Hispanics, 12.66 for Whites respectively. The finding suggests that if one uses the restricted gross effect to evaluate drug program aimed at all users, the treatment effect on all users will be underestimated.

The bias indexes under restriction $\beta = \alpha$ are much bigger in absolute values than that without restriction. The finding about direction of bias for illicit drug use due to binary measurement error is consistent with Aigner (1973). One should recognize that our favorable specification is $y = z\gamma + x\beta + u\alpha + \varepsilon$. If we omit u in the model, $\hat{\beta}$ is gross effect of the observed illicit drug use on birth weight without imposing $\beta = \alpha$. Put differently, it is gross treatment effect on the reported. When imposing $\beta = \alpha$ and including u in the model, $\hat{\beta}$ is the direct effect of illicit drug use on birth weight (alternatively, treatment effect on all users) which is the same as the coefficient on the actual by running OLS. So the gross effect without restriction and direct effect with restriction are not comparable in assessing bias direction of OLS estimator due to binary measurement error. Lets take some numbers for all races group to explain this point a little more. The gross effect of the reported, the direct effect of the reported, the direct effect of the unreported and the partial correlation between the unreported and the reported in Table 7-1 are -387.700, -424.523, -190.229 and -0.1976 respectively. It is easy to verify that -387.700= -424.523+(-190.229)×(-0.1976). The gross effect of the reported, the direct effect of the reported and the partial correlation between the unreported and the reported in Table 7-2 are -190.210, -236.880 and -0.1976 respectively. It will be held that -190.210=-236.880(1-0.1976). But one should note that $-387.700 \neq -236.880(1-0.1976)$, which means that one can't draw the conclusion that the OLS estimator is biased upwards due to binary measurement error by comparing 387.700 with 236.880. Basing on Equation (3-2) or Model 1 in Chapter V, Aigner (1973) theoretically claims that if one has knowledge about the covariance of the observed and its measurement error, the consistent OLS estimator can be obtained. Actually at the

empirical level, we can not use $\hat{\delta}_{uz}$ from out-of-sample to correct inconsistency of the OLS estimator according to Aigner (1973)'s formula when measurement error is unknown. As we have discussed in Chapter IV, it is not correct to think that the biased OLS estimator without restriction ($\beta = \alpha$) and that with restriction ($\beta = \alpha$) are the same if the given measurement error is ignored. Put differently, one can not get the biased OLS coefficient on the observed illicit drug use with its error omitted by assuming that the observed and its error have the same effect on birth outcomes.

One may argue why for the pure additive measurement error, the coefficients for the observed and its error should be equal. As we have seen, this restriction play a crucial role in computing bias indexes and gross effect of illicit drug use on birth outcomes at the empirical level. People may have certain knowledge about relationships among parameters in population. For instance, marginal propensity of consumption (MPC) should be in between 0 and 1. If one uses correct prior information in specifying the linear regression model, the OLS estimator will be BLUE. In most of cases, we need to build the model based on theory and evidences from the data. If we impose wrong restrictions in the model, of cause, we obtain biased estimators. There are two competing models in this study which are Equation (3-2) & (3-3) in Section 3.1. If one believes that x and u does not have the same effect on y in population, Equation (3-3) should be selected. Otherwise, Equation (3-2) is assumed to be the correct specification in population. Whether the restriction is reasonable or not depends on the theory and the results from statistical tests in reality. We test three null hypotheses as seen in Table 7-3. F statistics in Table 7-3 indicate that (1). The restriction ($\alpha = 0$) is accepted only by Whites group; (2). The restriction ($\beta = \alpha$) is rejected by all races group, Blacks,

Hispanics and Whites; (3) The restrictions $((\alpha = 0, \hat{\beta}_{ols} = \hat{\beta}(1 + \hat{\delta}_{ur}))$ are rejected by all races group and race-specific groups. The knowledge about $\beta = \alpha$ in population is not a favorable assumption or restriction in OLS estimation for all races, Blacks, Hispanics and Whites. The reason is that binary measurement error in this study is non-random and indicates a group of actual users being different from the reported according to their consumption levels of illicit drugs. From this view, the coefficients on the reported and unreported can be different!

7.2. Evaluating Treatment Effect on Heavy Users

Table 7-4 gives us a vivid picture of connection between consumption levels and reporting possibility for illicit drug use among pregnant women. The information in Table 7-4 reveals that illicit drug use has a negative effect on birth weight because mean birth weight for users is smaller than that for nonusers, and mean birth weight for the underreported is larger than that for the reported. Furthermore, The evidence states that light users are more likely to underreport their use, whereas heavy users are more likely to report their use.

Table 7-1 shows that (1). The coefficient on the reported is bigger than that of the underreported in absolute value; (2). The absolute magnitude of coefficient on the actual is smaller than that on the reported holding its error constant or with its error omitted, and bigger than that on the unreported. These two findings have very important policy implications.

The first finding suggests that drug treatment program should aim at heavy users. Medical scientists have found the harmful effects of illicit drug on birth outcomes.

Pregnant cocaine users are more likely to give the birth with a smaller head circumference. Marijuana smoking may impair oxygenation, with a consequent impairment of fetal growth. Heavy illicit drug use results in low birth weight and high infant mortality rate in short run. It also has long term effects on children's cognitive development, physical conditions and well beings in their whole life. If quantity of illicit drug use has negative marginal effect on birth weight, this finding also convinces that light users are more likely to underreport their use.

The second finding implies that to evaluate drug program aimed at heavy users, one needs to use the coefficient on the reported which measures the mean difference in birth weight between heavy users and nonusers, rather than the coefficient on the actual use with heavy use omitted because it only measures the gross mean difference in birth weight between all users and nonusers. If one uses the coefficient on the actual with heavy illicit drug use omitted in the model to evaluate drug program aimed at heavy users, the treatment effect on heavy users will be understated. Conversely, if one use gross effect or direct effect of the reported in Table 7-1 to evaluate drug program aimed at all users, the treatment effect is overestimated. According to our wisdom, the number of heavy users is plausibly less than that of light users in population. But this wisdom may not be true in some subsamples. Whether heavy users or light users are more likely to underreport their use depends on the data under study. If reduction in birth weight is proportionally related to quantity of illicit drug use, and heavy users are more likely to underreport their use, the coefficient on the underreported may be bigger than that on the reported in absolute value.

7.3. Examining Variance Correction Indexes

Table 7-5 explains what happens to the variance of the OLS estimator if one omits binary measurement error in the model. $f_1(\rho_{ux})$ is a variance correction index at theoretical level. It is derived from large sample theory in Chapter III. It fairly predicts the change direction of variance due to omitted measurement error in the model if one has information about measurement error. $f_2(\rho_{ux})$ is a variance correction index at empirical level. All variances of restricted OLS estimators are smaller than that of unrestricted one. This finding convinces the conclusion we draw in Section 3.5. For the binary measure of illicit drug use, if we omit the measurement error, the variance of OLS estimator for the variable given error is downwards by 6.1‰ for all races group, 26.6‰ for Blacks, 1‰ for Hispanics and 1.4‰ for Whites respectively. Both BI and $f(\rho_{ux})$ can be used in other studies where measurement error is unknown.

7.4. Comparing the results in the Study with Kaestner at el. (1996)

Kaestner at el. (1996) find that the coefficient on the underreported is half of that on the reported. They argue that the coefficient on the reported measures the treatment effect on heavy users because light users are more likely to underreport their use. The coefficient on the actual is smaller than that on the reported in absolute value. Table 7-6 shows that we obtain similar findings as compared to Kaestner at el. (1996). They use log(birth weight) as the dependent variable in infant health production function, whereas we use birth weight as the dependent variable. We divide our coefficients by mean birth weight so as to make these coefficients comparable with their results. It is really interesting that our results are bounded by their two corresponding results. For example, the treatment effect on all users in this study is 0.064 in between 0.057 and 0.089.

Table 7-1
Estimates of the Effect of Illicit Drug Use on Birth Weight in NYC
(Without Restriction: Alfa=Beta)

Race	Variable	Actual Effect	Gross Effect	Direct Effect	Partial Correlation	Bias Index	N
All	Xb	-236.880 #					126191
	xb		-387.010 #	-424.523 #			
	ub			-190.229 #			
	Delta				-0.1976 # (0.005)		
	BI					-0.088	
Blacks	Xb	-156.666 #					40074
	xb		-324.650 #	-360.870 #			
	ub			-112.420 #			
	Delta				-0.3225 # (0.010)		
	BI					-0.101	
Hispanics	Xb	-238.573 #					42490
	xb		-349.990 #	-363.307 #			
	ub			-174.707 #			
	Delta				-0.0763 # (0.007)		
	BI					-0.037	
Whites	Xb	-116.800 #					32534
	xb		-396.132 #	-406.354 #			
	ub			-80.775			
	Delta				-0.1266 # (0.016)		
	BI					-0.025	

Note:
1. # indicates significance at 5% confidence level.
2. Standard errors for partial correlation are in brackets.
3. Unit for actual, gross and direct effects is gram.
4. Partical correlation is the coefficient on x from regressing u on x and other regressors.
5. Bias index=(gross effect / direct effect)-1.

Table 7-2
Estimates of the Effect of Illicit Drug Use on Birth Weight in NYC
(With Restriction: Beta=Alfa)

Race	Variable	Actual Effect	Gross Effect	Direct Effect	Partial Correlation or	N
All	Xb	-236.880 #				126191
	xb		-190.210 @	-236.880 #		
	ub			-236.880 #		
	Delta				-0.1976 # (0.005)	
Blacks	Xb	-156.666 #				40074
	xb		-106.063 @	-156.666 #		
	ub			-156.666 #		
	Delta				-0.3225 # (0.010)	
Hispanic	Xb	-238.573 #				42490
	xb		-220.441 @	-238.573 #		
	ub			-238.573 #		
	Delta				-0.0762 # (0.007)	
Whites	Xb	-116.800 #				32534
	xb		-101.966 @	-116.800 #		
	ub			-116.800 #		
	Delta				-0.1266 # (0.018)	

Note:
1. # indicates significance at 5% confidence level;
2. @ shows zero standard error;
3. Standard errors in brackets;
4. Unit is gram for actual, gross or direct effects in the table.

Table 7-3
F Statistics for Testing the Restrictions In Infant Health Production Functions (Based on NYC sample)

Population Model 1: $y = z\gamma + x_b\beta + u\alpha + \varepsilon$

Population Model 2: $y = z\gamma + X_b\beta^* + \varepsilon$

Null hypothesis H_0	All Races	Blacks	Hispanics	Whites
$\alpha = 0$ (Model 1)	Rej	Rej	Rej	Ace
$\beta = \alpha$ (Model 1)	Rej	Rej	Rej	Rej
$\alpha = 0, \beta = \beta^*(1+\delta_{ux})$ (Model 1)	Rej	Rej	Rej	Rej

Note:

1. All, Black, Hispanic and White indicate race-specific infant health production functions.
2. Rej—Reject H_0, Ace—Accept H_0.
3. Statistical inference at 5% confidence level.
4. β, α are coefficients for illicit drug use and its error in health production functions.
5. δ_{ux} is the partial correlation between u and x.
6. NYC sample has 126191 observations.

Table 7-4
Mean Birth Weights & Percentages among Different race Groups

Race	All Users		Reportors		Underreporter		Nonusers		N
	MBW	Percentage	MBW	Percentage	MBW	Percentage	MBW	Percentage	
All Races	2916.51	7.3	2630.53	1.8	3008.34	5.5	3269.93	92.7	126191
Blacks	2908.94	16.6	2556.57	2.4	3006.20	13.3	3175.34	83.4	40074
Hispanics	2955.64	3.6	2727.47	1.34	3091.40	2.26	3277.69	96.4	42490
Whites	3180.75	5.1	2859.68	0.7	3225.86	4.4	3370.62	94.9	32524

Note:
1. N is the number of observations in NYC sample.
2. MBW is mean birth weight.

Table 7-5
The Variance Correction Index-- $f(\rho_{ux})$

Item	All Races	Blacks	Hispanics	Whites
σ_u^2	0.0523	0.1132	0.0221	0.0425
α^2	36187	12639	30523	6524
σ_ε^2	323308	382609	304406	282017
a	0.0059	0.0037	0.0022	0.0001
$\sqrt{\frac{a}{1+a}}$	0.0763	0.0610	0.0470	0.0099
ρ_{ux}	-0.0327	-0.0740	-0.0180	-0.0171
$f_1(\rho_{ux})$	0.0048	-0.0018	0.0019	-0.0002
$\text{var}_1(\hat{\beta})$	212.43	424.77	712.04	1880
$\text{var}_2(\hat{\beta})$	213.74	436.39	712.78	1882.7
$f_2(\rho_{ux})$	-0.0061	-0.0266	-0.0010	-0.0014

Note:

1. Black, Hispanic and White indicate race-specific health production functions.
2. Illicit drug use (x) and its measurement error (u) are binary variables in health production functions.
3. The Model 1 with specification error is: $y = z\gamma + x\beta + e$, While the Model 2 without specification error is : $y = z\gamma + x\beta + u\alpha + \varepsilon$. y is birth weight. z is a set of other regressors without measurement errors. $\alpha^2, \sigma_\varepsilon^2, \text{var}_2(\hat{\beta})$ are from the Model 2. $\text{var}_1(\hat{\beta})$ is from the Model 1.
4. σ_u^2 is the variance of measurement error.
5. $a = \frac{\alpha^2\sigma_u^2}{\sigma_\varepsilon^2}$, $\rho_{ux} = \frac{\sigma_{ux}}{\sigma_u\sigma_x}$, $f_1(\rho_{ux}) = a - (1+a)\rho_{ux}^2$, where ρ_{ux} is correlation coefficient between u and x.
6. $f_2(\rho_{ux}) = [\text{var}_1(\hat{\beta}) / \text{var}_2(\hat{\beta})] - 1$.
7. * indicates that α is not significantly different from zero at 5% confidence level.

Table 7-6
Comparing Estimates of the Effect of Illicit Drug Use on Birth Weight in This Study with Kaestner at el. (1996)

Variable	Actual Effect (1)	Gross Effect (2)	Partial Correlation (4)	Direct Effect (5)
Any Drug Use @				
Actual Use	-0.057 *			
Reported Use		-0.080 *		-0.089 *
Unreported Use				-0.042 **
Delta			-0.233 *	
Cocaine Use @				
Actual Use	-0.089 *			
Reported Use		-0.134 *		-0.148 *
Unreported Use				-0.066 **
Delta			-0.212 *	
Illicit Drug Use #				
Actual Use	-0.064 *			
Reported Use		-0.119 *		-0.131 *
Unreported Use				-0.059 *
Delta			-0.197 *	

Note:

1. Illicit drug use in this study includes Cocaine, Heroin, and Marijuana.
2. The sample size in this study is 126191.
3. Any drug use in Kaestner at el. (1996) includes Cocaine, Heroin, Marijuana and Methadone.
4. The sample size in Kaestner at el. (1996) is 1279.
5. @ and # indicate the results from Kaestner at el. (1996) and this study respectively.
6. Kaestner at el. (1996) use Log(birth weight) as the dependent variable, so we divide Our coefficients by mean birth weight to make things comparable.
7. * and ** indicate statistical significance at 5% and 10% respectively.

Chapter VIII
Endogeneity, Multicollinearity, Heteroscdaticity and Selectivity Due to Measurement Errors in Health Production Function

8.1. Two Step (IV) Method

If one has no information about measurement error in illicit drug use, the observed binary illicit drug use (x_b) is an endogenous variable in the model. The two step (IV) method can be used.

$$y = z\gamma + x_b\beta + \varepsilon \qquad (8\text{-}1)$$

$$x^* = w\eta + v \qquad (8\text{-}2)$$

$x_b = 1$ if $x^* > 0$

$x_b = 0$ otherwise.

Where x^* is reported quantity of illicit drug use that is not observed by the researcher. w is a vector of instruments which are agemoth, married, bornus, selffin, medicaid, and trimest1. It is assumed to be correlated with x_b, and not correlated to ε, v. z is other regressors such as msex, age1018, age3554, hsgrads, college, prenvis, plural, parity, alcohol, tobacco, hypert, induc. y is birth weight. This is a recursive simultaneous system with endogenous binary and latent variables. The model is identified even if ε and v are not independent, and z includes all variables in w. The two-stage estimation of this model proceeds as follows. We first get an estimate $\hat{\eta}$ of η by using Probit ML method for Equation (8-2), As for Equation (8-1), we can write it as

$$y = z\gamma + F(w\eta)\beta + \omega \qquad (8\text{-}3)$$

Where $\omega = \varepsilon + (x_b - F(w\eta))\beta$, $F(*)$ is the distribution function of ν. Because ω has zero mean and is uncorrelated with regressors, OLS can be used after substituting $\hat{\eta}$ for η. We get consistent estimators of β, γ due to consistent η. If ν and ε are independent, we can estimate equation (8-1) by OLS simply. Generally, the coefficient of x_b in Model (8-1) is inconsistent. But the coefficient of $F(w\eta)$ in Model (8-3) is consistent, and it is biased in the same direction as the OLS estimator. Which one will be less biased or more asymptotically efficient is not clear. Both the coefficient of x_b in Model (8-1) and that of $F(w\eta)$ in Model (8-3) show the mean difference in birth weight between illicit drug users and nonusers even though the specifications are not the same. In estimating Model (8-3), we correct standard errors according to Murphy and Topel (1985)'s formulation. We also compute conditional indexes for diagnosing multicollinearity. All conditional indexes are less 20. After correcting heteroscdaticity, the coefficients do not change too much, but standard errors become much smaller.

Table 8-1 shows that the coefficients from T-S (IV) are much bigger in absolute value than that from simple OLS. Can we interpret both of them in the same way? The answer is yes only if simultaneous equations system is recursive. The coefficients both from simple OLS and from T-S (IV) indicate the mean difference of birth weight between users and nonusers. In our example, mean birth weights for users are 387.01 grams lower than that of nonusers in all races sample. The coefficient from simultaneous equations system tells the mean difference in birth weight between users and nonusers is 2118 grams for all races group. Actually, simple OLS is a special case of simultaneous equations when error terms in both equations are not correlated.

8.2. Continuous Measurement Error in Omitted Variable Model

When one has good proxies for continuous measurement errors, the following model can be used:

$$y = z\gamma + X_p\beta + \varepsilon \qquad (8\text{-}4)$$

$$y = z\gamma + x_p\beta + u_p\alpha + \varepsilon \qquad (8\text{-}5)$$

Where y is birth weight, x_p is the predicted probability of illicit drug use. u_p is continuous measurement error. z is the same as the set of other variables in Equation (8-1). We assume that $\text{cov}(u_p, \varepsilon) = 0$, $\text{cov}(u_p, x_p) \neq 0$, and ε is independent of all regressors. As we discuss in the Chapter III, OLS estimator is inconsistent and biased by excluding u_p in the model.

Table 8-2 indicates that for excluding continuous measurement error, if one does not impose restriction $\alpha = \beta$, OLS estimators are biased upwards by 13.4% for all races group, 61.2% for Hispanics and 0.3% for Whites respectively, and OLS estimator is biased downwards by 14.3% for Blacks. We also find that the coefficient on the actual predicted probability (X_p) is smaller than that on the observed (x_p) with measurement error (u_p) omitted for all and race-specific groups. One additional finding is that the coefficient on X_p or u_p may be bigger or smaller in the absolute value than that on the x_p by holding u_p constant.

The classical error in variable model predicts that if one omits continuous measurement error, the OLS estimator will be biased downwards. The findings in this study do not totally support the conclusion from the classical error in variable model because continuous measurement error is not random.

8.3. Treatment Effect on Illicit Drug Users with Self-selection

In case of knowing binary indicator of the observed illicit drug use, one can estimate "treatment effect" by taking account of self-selection.

$$y = z\gamma + x_b\beta + \varepsilon \qquad (8\text{-}6)$$

$$x^* = w\eta + v \qquad (8\text{-}7)$$

$x_b = 1$ if $x^* > 0$

$x_b = 0$ otherwise.

$$y = z\gamma + x_b\beta + \rho\sigma_\varepsilon \frac{\phi(w\eta)}{\Phi(w\eta)} + \omega \qquad (8\text{-}8)$$

Where x^* is the quantity of actual illicit drug use which is not observed by the researcher, $\rho = corr(\varepsilon, v)$, $\omega = \varepsilon - \rho\sigma_\varepsilon \frac{\phi(w\eta)}{\Phi(w\eta)}$, w includes key factors such as bornus, married, selffin, medicaid, trimest1, WIC, AFDC, emplyd. It will affect the participation of illicit drug use. The notations for z, x_b and y are the same as before. If $\rho = 0$, we can estimate equation (8-6) by OLS. When $\rho \neq 0$, people's "intention" will affect their "action". Equation (8-8) is a deviation from Heckman sample selection model. This model is logically consistent and identified. We can first get an estimate $\hat{\eta}$ of η by using Probit ML method for equation (8-7); then we estimate equation (8-8) by OLS. It should be noted that in the second step, we need to use all observations in the sample. The estimates of γ, β are consistent, but the standard errors in the second step need to be corrected.

Table 8-3 shows the "treatment effect" of illicit drug use by taking account of self-selection. The coefficient of illicit drug use by using simple OLS indicates the mean effect of treatment on randomly selected persons. When considering self-selection, the coefficient of illicit drug use measures the mean effect of treatment on the treated. λ tells us the relationship between "intention" and "action" of illicit drug use. In our example, all λ's in race-specific health production are positive and statistically significant, which means that probability of participation has positive effect on the action taken by the treated for illicit drug use. The mean effects of treatment on the treated are smaller than that on randomly selected persons for Blacks, Hispanics and Whites. The mean effect of treatment on the treated is almost the same as that on randomly selected persons for all races group. The finding from Table 8-3 implies that the mean effect of treatment on the treated is overestimated if one omits the inverse of the Mills ratio in Equation (8-8), which is so called "sample selection bias".

8.4. Comparison between Continuous and Binary Measurement Errors in Health Production Function

As we have explored in Section 5.3, the coefficient of binary measurement error only indicates group mean difference between underreporters and nonusers. The coefficients from race-specific health production function on binary measurement error are comparable, but we can not draw a picture about prevalence of underreporting. Compared with binary illicit drug use, continuous predicted probability is a good & comprehensive index of frequency and participation in illicit drug use. So does continuous measurement error of illicit drug use. Kaestner at el. (1996) point out that using predicted probability of underreporting can eliminate technical screen error. The coefficient divided by 100 on

continuous measurement error shows the marginal effect of underreporting on the birth outcome – which means increment in birth weight due to 1% increment in the underreporting probability. All race-specific regression coefficients on continuous measurement errors are absolutely comparable. The conclusions here are : (1). Using marginal effect is better than using group mean difference for evaluating the effect of illicit drug use on birth outcomes; (2). The coefficients for both binary and continuous indicators of illicit drug use or its error are not comparable if one views the predicted probability as a proxy.

Table 8-1
Two Step (IV) & OLS Estimators on illicit drug use
From Race-Specific Health Production Functions in NYC

Race	Variable	Coefficient on Dummy	Correction for Endogeneity	Correction for Heterscdaticity	Multicollinearity Index	N
Two Step:						
All	**xp**		**-2117.67 #**	**-2116.99 #**	**9.134**	**126191**
Blacks	**xp**		**-1284.95 #**	**-1284.21 #**	**9.096**	**40074**
Hispanics	**xp**		**-2009.54 #**	**-2006.52 #**	**8.468**	**42490**
Whites	**xp**		**-2097.38 #**	**-2100.84 #**	**11.971**	**32524**
OLS:						
All	**xb**	**-387.01 #**				**126191**
Blacks	**xb**	**-324.64 #**				**40074**
Hispanics	**xb**	**-349.99 #**				**42490**
Whites	**xb**	**-396.13 #**				**32424**

Note:
1. # indicates significance at 5% confidence level;
2. Unit in Table 8-1 is gram.
3. Standard errors in second step are corrected;
4. Critical value for multicollinearity index is 20;
5. N is the number of observations.

Table 8-2
Estimates of the Effects of Illicit Drug Use on Birth Weight in NYC
(For Continuous Measurement Error)

Race	Variable	Actual Effect	Gross Effect	Direct Effect	Partial Correlation	Bias Index	N
All	Xp	-817.454 #					126191
	xp		-2117.665 #	-1866.753 #			
	up			-226.997 #			
	Delta				1.1050 # (0.004)		
	BI					0.134	
Blacks	Xp	-405.700 #					40074
	xp		-1284.948 #	-1498.730 #			
	up			186.014 #			
	Delta				1.1490 # (0.007)		
	BI					-0.143	
Hispanics	Xp	-1439.660 #					42490
	xp		-2009.545 #	-1246.306 #			
	up			-1765.727 #			
	Delta				0.4321 # (0.002)		
	BI					0.612	
Whites	Xp	-694.710 #					32534
	xp		-2097.385 #	-2092.085 #			
	up			-5.052			
	Delta				1.0480 # (0.008)		
	BI					0.003	

Note:
1. # indicates significance at 5% confidence level;
2.Standard errors for partial correlation are in brackets;
3.Unit for gross and direct effects is gram.
4. Partical correlation is the coefficient on x from regressing u on x and other regressors.
5. Bias index=(gross effect/direct effect)-1.

Table 8-3
Estimates of the Effect of Illicit Drug Use on Birth Weight in NYC
(With Self-selection)
(Dependent Variable:Birth Weight)

Race	Variable	Coefficient (No Self-Slection)	Coefficient (With Self-Slection)	N
All	xb	-378.010 #	-378.272 #	126191
	Lamda		52.909 #	
Blacks	xb	-324.650 #	-300.970 #	40074
	Lamda		67.585 #	
Hispanics	xb	-349.990 #	-336.424 #	42490
	Lamda		57.852 #	
Whites	xb	-396.132 #	-357.151 #	32534
	Lamda		28.829 #	

Note:
1. # indicates significance at 5% confidence level;
2. Birth weight in grams.
3. Bornus, married, trimest1, WIC, AFDC, emplyd, selffin and medicaid are independent variables in first step Probit equation.
4. Lamda is the inverse of the Mills ratio.

Chapter IX
Reutilization of the Information About Measurement Errors form or to Out-of-Samples

It is common that economic data involve measurement errors. Excluding measurement errors will bias OLS estimators. One can extract some information about measurement errors from historical data or other studies, and use it in out-of-sample in which measurement errors are not available. This is not an expensive way to get consistent estimator for the variable measured with errors.

9.1. Bias Indexes Used for Correcting Bias of OLS estimators

In Table 7-1 and 7-5, there are some race-specific bias indexes and variance correction indexes due to omitted binary measurement errors. One can use those indexes as out-of-sample information for correcting inconsistency and variances of OLS estimators in other studies by applying following formulas.

$$\hat{\beta}_{ols} = \beta(1 + \frac{\alpha}{\beta}\delta_{ux}) \qquad (9\text{-}1)$$

$$\text{var}(\hat{\beta}_{ols}) = \text{var}(\beta)(1 + f(\rho_{ux})) \qquad (9\text{-}2)$$

It should be noted that we can not use the formula $\hat{\beta}_{ols} = \beta(1 + \delta_{ux})$ to correct bias or inconsistency of OLS estimator by utilizing out-of-sample information (δ_{ux}) if we can not obtain the biased $\hat{\beta}_{ols}$. One may also use the corrected coefficient and variance to make statistical inference at empirical level.

9.2. Probit Coefficients Used for Predicting Binary Measurement Error

The following formulas are used to get binary measurement error:

$$I = z\varphi \qquad (9\text{-}3)$$

$$U_p = \Phi(I) \qquad (9\text{-}4)$$

$$u_b = 1 \quad \text{if } U_p \geq c \qquad (9\text{-}5)$$

$$u_b = 0 \quad \text{otherwise.}$$

Where φ is Probit coefficients from other studies, z is the set of individual characteristics in current sample including agemoth, hsgrads, college, married, bornus, black, whitenh, WIC, AFDC, emplyd. U_p is predicted probability of underreporting illicit drug use. u_b is converted binary measurement error, c is the threshold value.

Table 9-1 shows that schooling level has no effect on underreporting. Blacks and Whites are more likely to underreport illicit drug use. Someone who was born in U.S. and with poor social living status is more likely to be underreporting.

9.3. Tobit & OLS Coefficients Used for Predicting Continuous Measurement Error

We use the following models:

$$\text{Tobit:} \quad u_p^* = z\varphi + \varepsilon \qquad (9\text{-}6)$$

$$u_p = 0 \quad \text{if } u_p^* \leq 0$$

$$u_p = u_p^* \quad \text{if } u_p^* > 0$$

$$\text{OLS:} \quad \lg(u_p) = z_1\varphi_1 + z_2\varphi_2 + \varepsilon \qquad (9\text{-}7)$$

Where u_p^* is continuous measurement error without censoring, u_p is continuous censored measurement error, z, z_2 are set of maternal and infant characteristics. z is the same as that in Equation (9-4). z_2 is a subset of z which excludes agemoth. z_1 is expressed as lg(*agemoth*). Table 9-2 indicates that the coefficients from Tobit and OLS are all statistically significant. Schooling levels have effect on underreporting. One can uses those coefficients in out-of-sample.

9.4. Simulating Measurement Errors

This study obtains three kinds of density functions for continuous measurement error in illicit drug use by applying the curve fitting technique.

If $0 \le u_p \le 0.3000$ then,

$$f_1(u_p) = N(0.04, 0.025^2)$$
$$f_2(u_p) = \frac{2.38}{(1.0589 + u_p)^{35}} \qquad (9\text{-}8)$$
$$f_3(u_p) = \frac{0.00015}{u_p^2} e^{\frac{0.015}{u_p}}$$

otherwise, $f_i(u_p) = 0 \qquad i = 1,2,3 \qquad (9\text{-}9)$

In order to get these three density functions, we divide u_p into certain categories (or percentiles), and obtain mean values and cumulative frequencies for all categories. After plotting this distribution function from real data, we can choose relevant function forms to fit it by selecting proper values of parameters in those theoretical functions. One can take first derivatives of those theoretical distribution functions to get their corresponding density functions for different ranges of u_p.

Actually, the distribution function ($F(u_p)$) of u_p is truncated as $G(u_p)$:

$$G(u_p) = \frac{F(u_p)}{1 - F(u_p < 0) - F(u_p > 0.3000)} \qquad (9\text{-}11)$$

Where $0 \le u_p \le 0.3000$.

For binary measurement error, its binomial density function is

$$f(u_b) = 0.055^{u_b}(1-0.055)^{1-u_b}, \quad \text{for } u_b = 0,1. \qquad (9\text{-}12)$$

One can use random standard normal density to create the random sample for continuous measurement error, or use random binomial density function to create the random sample for binary measurement error. It is important to note that simulated measurement error is more likely to be random. As we know, measurement errors in economic data tend to be more systematic. If correlation among measurement error and the observed as well as other regressors in the model is low or close to zero, we may simulate measurement errors. Otherwise, it will be better to use the methods introduced in Section 9-1, 9-2, 9-3.

Table 9-1
The Coefficients from Probit Model in NYC For Predicting Binary Measurement Error (Dependent Variable: Ub)

Independent Variable	Coefficient
intercept	-12.856 #
agemoth	0.17159 #
hsgrads	0.02755
college	0.00732
married	-2.1477 #
bornus	2.8772 #
black	3.3802 #
whitenh	2.0079 #
WIC	-0.1133 #
AFDC	-0.15306 #
emplyd	-0.01416
selffin	1.70408 #
medicaid	2.12398 #

Note:
1. # indicates significance at 5%;
2. The sample size is 126191 (All races);
3. Definitions of variables in Table 6-2.

Table 9-2
The Coefficients from Tobit & OLS in NYC
For Predicting Continuous Measurement Error
(Dependent Variable=Up for Tobit)
(Dependent Variable=Log(Up) for OLS)

Independent Variable	Tobit Coefficient	OLS coefficient
intercept	-0.07487 #	-7.4618 #
agemoth	0.00271 #	0.0731 # **
hsgrads	0.00067 #	-0.0971 #
college	-0.00236 #	-0.0696 #
married	-0.02715 #	-0.9341 #
bornus	0.03457 #	1.0513 #
black	0.06706 #	1.9271 #
whitenh	0.02993 #	1.5585 #
WIC	-0.00277 #	0.0233 #
AFDC	0.00491 #	-0.1240 #
emplyd	-0.00564 #	-0.0211 #
selffin	0.00961 #	0.3808 #
medicaid	0.02844 #	0.9680 #
scale	0.02496	

Note:
1. ** indicates that agemoth should be changed to log(agemoth);
2.# shows significance at 5% confidence level;
3.The number of observations is 126191.
4. Definitions of variables in Table 6-2.

Chapter X
Additional Findings
Related to Human Behaviors of Illicit Drug Use

10.1. Choices of Illicit Drug Use

Why some pregnant women use drug and the others don't? Random utility theory states that if pregnant women face k choices, they will choose one that gives them the highest utility with the greatest probability. Personal indexes and the distribution of the unobserved error term determine utility differentials among those choices. The discrete choice model can predict the behavior pattern of pregnant women in illicit drug use. Table 10-1 reveals that (1). Pregnant women whose ages are in between 10 and 18 are less likely to use multiple illicit drugs, whereas pregnant women whose ages are in 35 and 54 are more likely to use at least one illicit drug; (2). Married women are less likely to use illicit drugs; (3). Someone who is born in U.S. is more likely to use at least one illicit drug; (4). Blacks and Whites are more likely to use multiple illicit drugs; (5). Someone who is in Medicaid program is more likely to use at least one illicit drug; (6). Pregnant women who start prenatal care early are less likely to use illicit drugs.

10.2. Consumption Preference of Illicit Drug Use

According to traditional utility theory, price of illicit drugs, shadow price of infant health, prices of other commodities and income mainly determine demand for illicit drugs. Suppose that parents' utility function is only a function of illicit drug use and infant health. Holding real income constant, as price of illicit drug goes up, the shadow price of health goes down, so relative price of illicit drug to infant health increases. Opportunity set will be away from illicit drugs and towards infant health. The rational

consumers will demand less illicit drugs and more infant health so as to maximize their utility. The irrational consumers can make choices randomly within opportunity set. But on average, they will consume less illicit drug and more infant health. Statistically speaking, law of demand still holds. One should keep in mind that illicit drug market is underground ant not fully competitive. Based on our wisdom, heavy addictive illicit drug users have low price sensitivity, and the young and poor beginners of illicit drug use have high price sensitivity.

Table 10-2 shows that (1). Blacks are the first heavy users of alcohol, cocaine, marijuana and tobacco; (2). Whites are the second heavy users of alcohol, marijuana and tobacco; (3). Asians like to drink alcohol, whereas Hispanics tend to smoke cigarettes.

10.3. The Effects of Personal Characteristics on Birth Outcome

Infant health production function is a function of prenatal care, illicit drug use and other maternal and infant characteristics. Table 10-3 suggests that (1). Pregnant women with ages between 10 and 18, or 35 and 54 give lower weight birth compared to age group in between 19 and 34; (2). High school and college graduates give high weight birth compared to high school dropouts; (3). Alcohol, illicit drug and tobacco have significantly negative effects on birth weight; (4). Gestational hypertension has large negative effects on birth weight; (5). Early prenatal care and induced abortion almost have no effects on birth weight in race-specific health production functions.

Table 10-1
The Coefficients from Mlogit on CA Sample
(Dependent Variable Dchoice=0,1,2)

Variable	Dchoice=1	Dchoice=2
intercept	-7.201 #	-4.725 #
age1018	-13.008	-1.896 #
age3554	0.954 #	0.415 #
married	-0.797 ##	-0.917 #
bornus	2.053 #	1.275 #
black	-12.575	1.891 #
hisp	0.306	0.284
whitenh	1.257	1.193 #
selffin	1.298	0.477
medicaid	1.036 #	0.637 #
trimest1	-1.385 #	-0.803 #

Note:

1.# and ## indicate significance at 5% and 10% respectively.

2. Dcoice=0 for nousers, 1 for one illicit drug users, 2 for multiple illicit drug users.

3. The number of observations is 8406.

Table 10-2
Race-Specific Prevalence of Illicit Drug Use
(In 15 Large Cities of California)

Race	Alcohol	Cocaine	Marijuna	Heroin	Methadone	Tobacco	N
Asians	4.10	0.28	0.19	0.19	0.00	1.69	1061
Blacks	12.05	10.25	5.54	0.30	0.22	23.90	1336
Hispanic	0.81	0.81	0.73	0.20	0.20	3.91	3438
Whites	6.27	1.02	3.64	0.18	0.28	16.49	2168
All	7.13	2.33	2.18	0.20	0.20	10.28	8406

Note:

1. Unit is percentage.
2. Name list of 15 large cities in Appendix Table 1.
3. N is the number of observations.
4. The original 1993 urine test sample has 29494 observations.

Table 10-3
The Coefficients from Race-Specific Infant Health Production Functions in NYC (Dependent Variable: Birth Weight)

Variable	All Races	Blacks	Hispanics	Whites
msex	113.051 #	107.694 #	103.119 #	138.762 #
age1018	-90.828 #	-62.225 #	-78.506 #	-51.362 #
age3554	-12.143 #	23.995 #	-37.527 #	-33.458 #
hsgrads	42.905 #	44.893 #	42.087 #	40.787 #
college	40.816 #	42.229 #	21.021 #	49.601 #
prenvis	22.772 #	26.152 #	23.137 #	12.130 #
plural	-927.965 #	-929.876 #	-928.772 #	-928.825 #
parity	29.058 #	23.260 #	29.382 #	32.369 #
alcohol	-107.306 #	-174.135 #	-93.311 #	296.129 #
Xb	-236.883 #	-156.665 #	-238.573 #	-116.779 #
tobacco	-184.867 #	-223.552 #	-209.674 #	-132.756 #
hypert	-188.616 #	-140.226 #	-182.273 #	-219.153 #
induc	-14.585 #	0.735	5.869	3.347
N	126191	40074	42490	32534

Note:

1. # indicates significance at 5% confidence level.
2. Xb is a binary indicator of actual illicit drug use.
3. N is the number of observations in race-specific samples in NYC.

Chapter XI
Conclusions

In this paper, we have derived the theoretical formulas for correcting bias and variances of OLS estimators due to omitted binary or continuous measurement errors. Especially, we have examined the effect of illicit drug use and its measurement error on infant birth outcome. There are some important findings as follows.

1. Binary underreported error in illicit drug use results in an underestimate of the direct effect of illicit drug use on birth weight. This is consistent with Aigner (1973)'s conclusion.
2. Continuous underreported error in illicit drug use produces an overestimate of direct effect of illicit drug use on birth outcome except for Black pregnant women. This result contrasts with the conclusion in classical error in variable model. The reason is that NYC data does not support the assumptions such as $E(u_p) = 0$, $\text{cov}(X_p, u_p) = 0$ and $\text{cov}(x_p, z) = 0$ in this study.
3. The direction and magnitude of change in the variance of OLS estimator for the variable given error is determinate if one has information about measurement error. For omitted non-stochastic measurement error in the finite sample, the variance of restricted OLS estimator is less than that of unrestricted one.
4. The coefficient on the underreported is smaller than that on the reported in absolute value, which means that light users are more likely to underreport their use.
5. The coefficient on the actual is smaller in absolute value than that on the reported by holding measurement error constant. If one uses the coefficient on the actual to

evaluate drug program aimed at heavy users, the treatment effect on heavy users will be underestimated.

6. If one uses the mean probability as the threshold value, the converted binary indicator will overestimate the true prevalence of illicit drug use. The coefficients for continuous and binary measurement errors in health production are not comparable because of different interpretations. The magnitude of the coefficient on binary illicit drug use is not obviously associated with illicit drug prevalence.
7. Under recursive simultaneous equations system with endogenous binary and latent variables, two step estimation method is preferred. The coefficient on illicit drug use from this model and that from simple OLS by assuming that error terms among equations are independent can be interpreted in the same way. But if one views predicted probability of illicit drug use as a proxy in the model, the coefficients for continuous predicted probability and for binary indicator are not comparable.
8. The information about measurement error in one study can be applied to the others. The matter is that we need to verify whether target and source samples are from the same population.

This study mainly focuses on measurement error in structure equations system and simple OLS regression model. The further researches should be in following directions: (1). How to simulate the actual measurement error with a given density function; (2). How to filter out "noise" from the observed.

Appendix I

Utility Maximization & Health Production Function

Suppose that we have the following model:

$$Max \quad U = U(C, X, y) \qquad (2\text{-}1)$$

$$y = g(X, M, E) \qquad (2\text{-}2)$$

$$ST: \quad I = P_C C + P_X X + P_M M \qquad (2\text{-}3)$$

Where U is parents' utility function, C is the other good, X is the actual illicit drug use, y is infant birth weight, M is prenatal care, E is mothers' or infant characteristics, P_C, P_X, P_M are prices of C, X, M respectively, I is the income.

We assume $MP_X < 0, MP_M > 0, U_C > 0, U_X > 0, U_y > 0$.

We have Lagrange function:

$$L = U(C, X, y) + \lambda(I - P_C C - P_X X - P_M M) \qquad (2\text{-}4)$$

The first order conditions are:

$$U_C - \lambda P_C = 0 \qquad (2\text{-}5)$$

$$U_X + U_y MP_X - \lambda P_X = 0 \qquad (2\text{-}6)$$

$$U_y MP_M - \lambda P_M = 0 \qquad (2\text{-}7)$$

$$I - P_C C - P_X X - P_M M = 0 \qquad (2\text{-}8)$$

There are four equations for four unknowns (C, X, M, λ). We can get optimal solution for this model.

Now we discuss two cases:

1. If $X = x + u$, where x is the reported illicit drug use, u is the unreported illicit drug use, then $MP_x = MP_u = MP_X$, $U_x = U_u = U_X$. This is the case in which we set equal coefficients for x, u in health production function, especially for pure additive measurement error. So if we treat u as measurement error, we can define health production functions as follows:

$$y = z\gamma + X\beta + \varepsilon \qquad (2\text{-}9)$$

$$y = z\gamma + x\beta + u\beta + \varepsilon \qquad (2\text{-}10)$$

Where $z = (M, E)$, ε is equation error term.

2. If we simply have dichotomous measure of the actual illicit drug use, we can decompose the actual illicit drug users (X) into light users (u) and heavy users (x) by assuming that light users may be more likely to underreport their use, which means that x and u can have different coefficients in health production function. We define health production function as follows:

$$y = z\gamma + x\beta + u\alpha + \varepsilon \qquad (2\text{-}11)$$

It is worthy to note that (2-9) is started from the classical error in variable model which emphasizes that the coefficients for x, u should be equal. (2-11) is started from the omitted variable model which stresses u as a variable with a different coefficient. Generally speaking, (2-9) and (2-11) are two competing population models in specifying health production functions.

Appendix II

15 Large Cities in CA and Their Drug Prevalence

City	Population	N	Illicit Drug Use (%)	Anydrug Use (%)
Anaheim	272,327	392	2.55	4.34
Bakersfield	271,347	312	7.69	8.33
Fresno	347,905	497	3.22	5.63
Hunting Ton	255,681	170	2.35	3.53
Long Beach	299,651	148	4.05	5.41
Los Angeles	2,102,295	1553	5.09	7.21
Modesto	216,439	260	6.15	9.23
Oakland	314,487	730	10.14	15.48
Riverside	253,487	219	5.48	6.39
Sacramento	628,279	938	7.04	9.60
San Diego	1,049,298	864	2.84	4.26
San Francisco	723,993	869	4.60	8.40
Stockton	267,258	224	3.57	4.46
Santa anA	290,992	369	1.63	4.07
San Jose	816,653	879	3.75	5.96
Total	8,110,092	8406	4.97	7.40

Note:

N is the number of observations in subsamples.

Appendix III
Underreporting Rates of Illicit Drug Use in NYC

Race	N	Actual Use	Underreporting	Overrepoting	Underreporting Rate
Blacks	40074	15.0	12.4		82.7
		16.6	13.0	1.7	78.3
Hispanic	42490	2.5	1.2		48.0
		3.6	2.26	1.1	62.8
Wites	32524	4.7	4.1		87.2
		5.1	4.4	0.4	86.3
All	126191	6.4	4.6		71.9
		7.3	5.5	1.0	75.3

Note:

1. Unit in the Table is percentage.
2. Underrepoting=actual-selfreporting.
3. Undereporting Rate=Underreporting/Actual Use.
4. Assuming that there is no overreporting, add overreporting to Actual Use.

Appendix IV
Correlation among Some Key Variables in NYC Sample

Correlation	All	Blacks	Hispanics	Whites
Corr(ub, xb)	-0.0327 #	-0.074 #	-0.018 #	-0.017 #
Delta(ub, xb)	-0.197 #	-0.323 #	-0.074 #	-0.127 #
Corr(up, xp)	0.721 #	0.723 #	0.669 #	0.644 #
Delta(up, xp)	1.105 #	1.149 #	0.432 #	1.408 #
Corr(ub,Xb)	0.861 #	0.866 #	0.786 #	0.933 #
Corr(up, Xp)	0.985 #	0.967 #	0.894 #	0.954 #
Corr(xb, Xb)	0.48 #	0.433 #	0.600 #	0.343 #
Corr(xp, Xp)	0.889 #	0.875 #	0.928 #	0.843 #
|Corr(xb, z)|<=	0.25	0.32	0.25	0.20
|Corr(xp, z)|<=	0.30	0.35	0.21	0.20
N	126191	40074	42490	32524

Note:

1. # indicates significance at 5% confidence level.
2. Ub and up are binary and continuous measurement errors respectively.
3. Xb, Xp, xb, xp are the actual and the observed illicit drug use.
4. Delta is partial correlation between the observed and its error.
5. N is the number of observations in the samples.
6. z is the set of other regressors in health production function.

References

Aigner, D. "Regression with a Binary Independent Variable Subject to Error of Observation." Journal of Econometrics, March, 1973, 49-60.

Becker, G. S. "A Theory of the Allocation of Time." Economic Journal, Sept., 1965, 493-517.

Bound J. et al. "Problems with Instrumental Variables Estimation When the Correlation between Instruments and Endogenous Explanatory Variable Is Week." Journal of the American Statistical Association, June, 1995, 443-450.

Buse, A. "The bias of Instrumental Variable Estimators." Econometrica, Vol. 60, No. 1, Jan. 1992, 173-180.

Chasnoff, I. Et al. "The Prevalence of Illicit-Drug or Alcohol Use during Pregnancy and Discrepancies in Mandatory Reporting in Pinellas County, Florida." The New England Journal of Medicine, Apr. 1990, 1202-1206.

Dagenais, M. G. et al. "Higher Moment Estimators for Linear Regression Models with Errors in the variables." Journal of Econometrics, 76, 1997, 193-221.

Greene, W. H. Econometric Analysis. Macmillan Publishing Company, 1993.

Griffiths, W. E. et al. Learning and Practicing Econometrics, McGraw-Hill, Inc., 1993.

Grossman, M. The Demand for Health: A Theoretical and Empirical Investigation, New York, Columbia University Press, 1972.

Gujarati, D. Basic Econometrics, McGraw-Hill, Inc., 1988.

Heckman, J. "Instrumental Variables A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations." The Journal of Human Resources, XXXII, 3, Jan., 1997, 441-462.

Heckman, J. "Sample Selection Bias as a Specification Error." Econometrica, 47, 1979, 153-161.

Joyce, T. et al. "The Consequences and Costs of Maternal Substance abuse in New York City A Pooled Time-Series, Cross-Section Analysis."

Journal of Health Economics, 11, 1992, 297-314.

Kaestner, et al. "The Effect of Maternal Drug Use on Birth Weight: Measurement Error in Binary Variables." Economic Inquiry, Oct., 1996, 617-629.

Klepper, S. et al. "Consistent sets of Estimates for Regressions with Errors in All Variables." Econometrica, Vol. 52, No. 1, Jan., 1984, 163-183.

Klepper, S. "Bounding the Effects of Measurement Error in Regressions Involving Dichotomous Variables." Journal of Econometrics, 37, 1988, 343-359.

Leamer, E. E. "Destructive Diagnostics for the Error-in-Variable Model." Department of Economics, UCLA, 1983.

Maddala, G. S. Econometrics, New York, McGraw-Hill, 1977.

Maddala, G. S. Limited Dependent and Qualitative Variables in Econometrics. Cambridge University Press, 1983.

Maddala, G. S. Introduction to Econometrics. Macmillan Publishing Company,1992.

Murphy, K. M. et al. "Estimation and Inference in Two-Step Econometric Models." Journal of Business & Economic Statistics, Oct., 1985, 370-379.

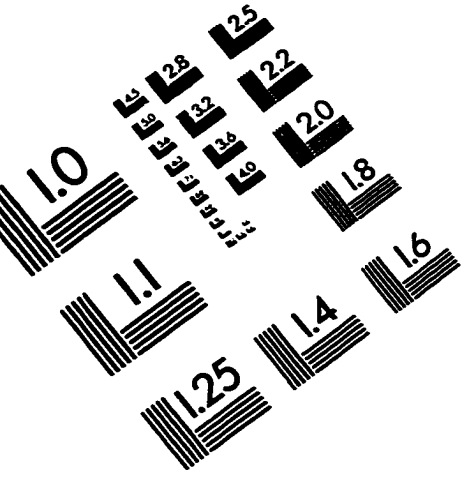
Nelson, C. R. et al. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." Econometrica, Vol. 58, No. 4, July, 1990, 967-976.

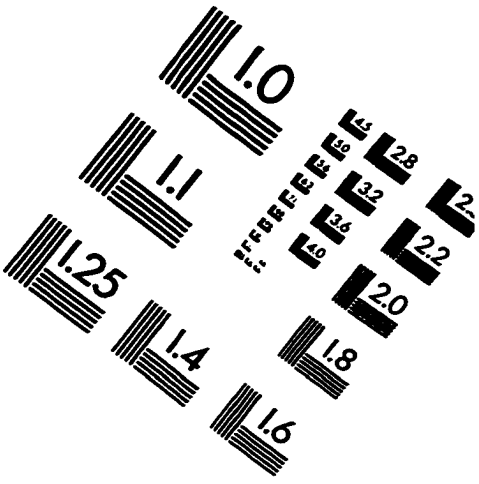
Vega, W. et al. "Prevalence and Magnitude of Prenatal Substance Exposures in California." The New England Journal of Medicine, Sept. 16, 1993, 850-854.

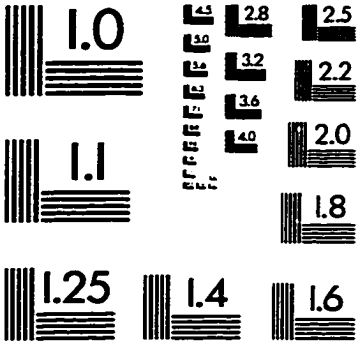
Welch, F. "Human Capital Theory: Education, Discrimination and Life Cycle." American Economic Review, May, 1975, 67.

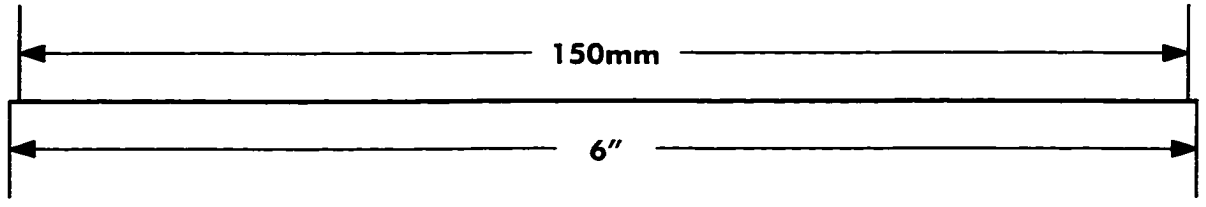
Zukerman, B. et al. "Effects of Maternal Marijuana and Cocaine Use on Fetal Growth." The New England Journal of Medicine, Oct.,1989, 762-768.

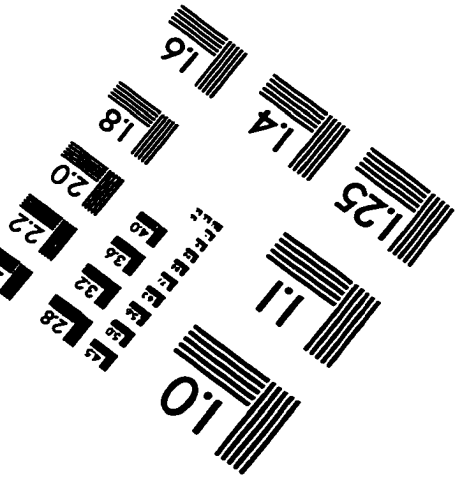
IMAGE EVALUATION TEST TARGET (QA-3)

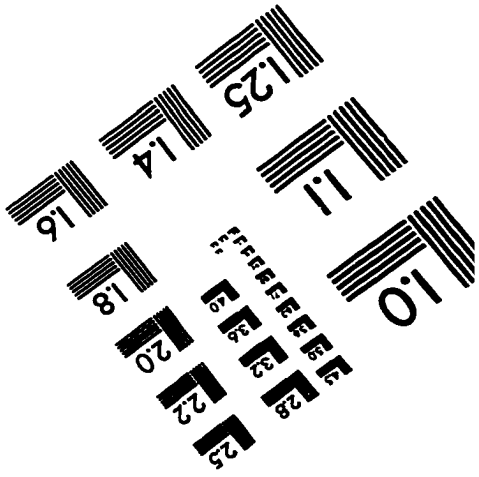












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989