BLS WORKING PAPERS

U.S. DEPARTMENT OF LABOR
Bureau of Labor Statistics

OFFICE OF EMPLOYMENT AND UNEMPLOYMENT STATISTICS

Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables

Harley Frazis, U.S. Bureau of Labor Statistics
Mark A. Loewenstein, U.S. Bureau of Labor Statistics

Working Paper 355
March 2002, Revised September 2002

Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables

Harley Frazis
Bureau of Labor Statistics
2 Massachusetts Ave. NE
Suite 4945
Washington D.C. 20212

Phone: 202-691-7394
Frazis_H@bls.gov

Mark A. Loewenstein
Bureau of Labor Statistics
2 Massachusetts Ave. NE
Suite 4130
Washington D.C. 20212

Phone: 202-691-7385
Loewenstein_M@bls.gov

February 2002, revised September 2002

We thank Dan Black and Tim Erickson for helpful comments, and Chris Foote for use of his job reallocation data. Anonymous referees' comments led to a substantial redirection and improvement of this paper. Opinions expressed are those of the authors and do not reflect the views of the Bureau of Labor Statistics or those of its staff.

Abstract

This paper is concerned with mismeasured binary explanatory variables in a linear regression. Modification of a technique in Hausman et al. (1998) allows simple computation of bounds under relatively weak assumptions. When one has instruments, we show how to obtain consistent parameter estimates using GMM. We show how to incorporate the estimated measurement error bounds into the GMM estimates, and we develop a specification test based on the compatibility of the GMM estimates with the measurement error bounds. When the mismeasured variable is endogenous, the IV estimate and the measurement error bounds can be used to bound the mismeasured variable's coefficient.

I. Introduction

This paper is concerned with mismeasured binary explanatory variables in a linear regression. We obtain results that allow us to improve on existing estimators under different assumptions about the extent of prior information. We examine three different cases: 1) the mismeasured variable is assumed exogenous, and no instruments are available; 2) the mismeasured variable is assumed exogenous, and one or more instruments are available; and 3) the mismeasured variable is not assumed exogenous, and instruments are available. In the first two cases, we derive bounds or point estimates under weaker assumptions on prior information than in the previous literature, and in the third case—which has not to our knowledge been analyzed—we also derive bounds.

In case 1, the traditional approach to the measurement error problem is to use auxiliary information on the measurement error process. More generally, one may not have good point estimates of the measurement error parameters, but may nevertheless be able to bound them—using a validation study, for example. Bollinger (1996) shows how these bounds can in turn be used to bound the regression coefficient.

Recently Hausman, Abrevaya, and Scott-Morton (1998) (hereafter HAS) have developed a technique that allows the analyst to bound the measurement error process by estimating the parameters from the data, without information from validation studies. We modify their technique to allow simple computation of bounds without functional form assumptions. Combining the estimated measurement error bounds with the OLS coefficient yields bounds on the true effect of the mismeasured explanatory variable.

When instruments are available, as in case 2, instrumental variable (IV) estimation is another common method of dealing with measurement error. However, recent research has shown that IV estimation is upwardly biased when the mismeasured variable is binary (Loewenstein and Spletzer 1997; Berger, Black and Scott 2000; Kane, Rouse and Staiger 1999) because measurement error in this case must be negatively correlated with the true value. Berger, Black, and Scott (2000) (BBS hereafter) and Kane, Rouse, and Staiger (1999) (KRS hereafter) show that when one has two erroneous measures, one can obtain a consistent estimate using a generalized method of moments (GMM) technique.

Two distinct measures of the same variable are not commonly available. We thus extend the analysis in BBS and KRS to the case where the second measure is replaced by one or more instruments. We provide a closed-form solution for the GMM parameter estimates. We also show how to incorporate the estimated measurement error bounds into the GMM estimates, and we develop a specification test of the measurement error model based on the compatibility of the GMM estimates with the measurement error bounds.

Lastly, we show that the GMM technique is not easily extended to the case where the mismeasured variable is endogenous. However, the IV estimate and estimates of the measurement error bounds can be used to bound the effect of the mismeasured variable, analogous to the OLS case without endogeneity.

The paper is organized as follows. Section II outlines the model, summarizes the HAS technique, and introduces our extension of HAS. Section III then turns to the case where one or more instruments exist for the mismeasured variable. Section IV considers the case where there are available instruments, but the mismeasured binary variable is endogenous. Section V presents an empirical example looking at the returns to on-the-job training and section VI concludes.

II. Bounding the Effect of a Mismeasured Binary Explanatory Variable When No Instrument is Available

Our model is:

(1) $Y_i = c + X_i\gamma + \beta T_i * + e_i$

for observation i in a random sample of n observations, where Y_i, T_i* and e_i are scalars, c is a constant, and X_i is a 1 x k row vector. We assume without loss of generality that $\beta >$ 0. Dropping the subscript for convenience, the variable T* is a binary variable that takes on the values of zero and one; T* and the elements of X are not linearly dependent. The error term e is mean zero and uncorrelated with T* and X. The variable T* is measured with error. Instead of T*, we observe the binary variable $T = T^* + U$, where U denotes measurement error (which can take on the values of 1, -1, or 0).

Define the measurement error probabilities $\alpha_0 \equiv \Pr(T = 1 | T^* = 0) = \Pr(U = 1 | T^* = 0)$ and $\alpha_1 \equiv \Pr(T = 0 | T^* = 1) = \Pr(U = -1 | T^* = 1)$. These probabilities are assumed to be independent of X and e. We should note that this may be a strong assumption in many applications (see Black, Sanders, and Taylor 2000); in a regression of earnings on education and (binary) training status, for example, more educated respondents may better understand the survey questions on training. Modifying this assumption would require modification of the standard results on the effect of measurement error on regression coefficients (Aigner 1973, BBS, KRS).

Let $p \equiv \Pr(T = 1)$ and $p^* \equiv \Pr(T^* = 1)$. It is straightforward to show that

(2) $\text{cov}(T,T^*) = p^*(1-p^*)(1-\alpha_0-\alpha_1)$.

Following Bollinger (1996) and others, we impose the restriction that cov(T,T*) > 0 (if this is not the case, then measurement error is so severe that (1-T) is a better measure of

T^* than is T), which implies $\alpha_0 + \alpha_1 < 1$.

Note that

$$(3) \qquad p\lim \hat{\beta}_{ols} = \beta(1 - \frac{Cov(T,U \mid X)}{Var(T \mid X)}).$$

Since $Cov(T,U|X) > 0$, $p\lim \hat{\beta}_{ols} < \beta$. After some algebra, one can show that

$$(4) \qquad \beta = p\lim \hat{\beta}_{ols} \chi(p, R, \alpha_0, \alpha_1) \ ,$$

where $\chi(\alpha_0, \alpha_1, p, R) \equiv \dfrac{p(1-p)(1-R)(1-\alpha_0-\alpha_1)}{(p-\alpha_0)(1-\alpha_1-p)-Rp(1-p)}$ and where

$R \equiv \dfrac{Cov(X,T)Var(X)^{-1}Cov(X,T)'}{p(1-p)}$ is the theoretical R-squared from a regression of T on X. It is straightforward to show that χ is an increasing function of both α_0 and α_1. Thus, if we have available upper bounds $\alpha_0^{\max}$ and $\alpha_1^{\max}$ on the measurement error parameters α_0 and α_1, we can bound β:

$$(5) \qquad p\lim \hat{\beta}_{ols} < \beta < p\lim \hat{\beta}_{ols} \chi(\alpha_0^{\max}, \alpha_1^{\max}, p, R).$$

<u>A Percentile Method for Bounding Measurement Error</u>

We now develop a method to bound α_0 and α_1. The independence of X and the measurement error process implies that

$$(6) \qquad Pr(T=1|X) = (1 - \alpha_1)\, Pr(T^*=1|X) + \alpha_0\, (1 - Pr(T^*=1|X))$$

$$= \alpha_0 + (1 - \alpha_0 - \alpha_1)\, Pr(T^*=1|X).$$

If one knows the functional form of $Pr(T^* = 1|X)$, one can estimate α_0 and α_1 from (6) by maximum likelihood, as in Hausman, Abrevaya, and Scott-Morton (1998) (HAS hereafter). But as we now show, simple $\sqrt{n}$-convergent bounds for α_0 and α_1 are available without knowledge of the true functional form.[1]

Equation (6) has the important implication that $\alpha_0 \leq \Pr(T=1|X) \leq 1-\alpha_1$ for all X. Thus, for any subset S of the range of X, an estimate of $Pr(T=1|X \in S)$ can be used to bound α_0 and α_1. Imagine a procedure where one takes a subset S comprising q percent of the sample in order to use the sample average T over S as a bound. If one has prior knowledge of the ranking of sample observations according to $Pr(T=1|X)$, one would obviously obtain the tightest possible bound for α_0 (α_1) by estimating $E(T | X \in S)$ over that subset of sample observations having the lowest (highest) expected value of T - that is, the set S should consist of observations with percentile rank of $Pr(T=1|X)$ less than q (greater than 1-q). Without prior knowledge of the ranking of observations by $Pr(T=1|X)$, one must estimate the ranking from the sample.

If the functional form chosen to estimate the ranking of $Pr(T=1|X)$ is incorrect, the estimated bounds will not be tight for a given q because some observations will be misclassified in the limit as having percentile rank above q when their true rank is below q, and vice versa (and correspondingly for rank 1-q). Since $\alpha_0 \leq \Pr(T=1|X) \leq 1-\alpha_1$ for all X, an incorrect functional form will affect only the tightness of the bounds, not their validity. In most cases, the percentile ranks of $Pr(T=1|X)$ are unlikely to be drastically affected by functional form.

---

[1] HAS present another procedure for estimating misclassification probabilities (and $Pr(T^*=1|X)$) that does not rely on prior knowledge of the functional form of $Pr(T^*=1|X)$, but this procedure is computationally quite complex and converges at a rate slower than $\sqrt{n}$.

Because we require asymptotic results where the functional form of *Pr*(*T=1|X)* may be misspecified, it is convenient to work with quasi-maximum-likelihood (QML) estimation (see White 1982) in estimating the ranking of *Pr*(*T=1|X)*. Specifically, consider estimation of the model $E(T|X) = G(X\,;\,\delta)$, where G is a cdf but not necessarily the true one and δ is a vector of parameters. Let $\hat{\delta}$ denote the quasi-maximum-likelihood estimator: $\hat{\delta} \equiv \arg\max L(\delta) \equiv$

$\arg\max (1/n)(\sum_{i:T_i=0} \ln(1 - G(x_i;\delta)) + \sum_{i:T_i=1} \ln G(x_i;\delta))$, so that the predicted value of T is simply $\hat{T} = G(X;\hat{\delta})$. We assume sufficient regularity such that the quasi-maximum likelihood estimator $\hat{\delta}$ exists and converges to a limit δ^* and such that $\sqrt{n}(\hat{\delta} - \delta^*)$ converges to a normal distribution.[2]

We now introduce some necessary notation. Let J_δ denote the cumulative distribution function of $G(X;\,\delta)$ and let $\kappa_q^* \equiv J_{\delta^*}^{-1}\ (q)$ denote the q-quantile for J_{δ^*}. Letting $\theta \equiv [\delta' \quad \kappa_q]'$, one can define the function, $A^0(\theta_q) \equiv E(T \mid G(X;\delta) \le \kappa_q)$. In addition, let $\hat{\kappa}_q \equiv \min(c \mid \hat{J}_{\hat{\delta}}(c) \ge q)$ denote the sample q-quantile for $\hat{T}$, where $\hat{J}_{\hat{\delta}}$ is the empirical cdf of $\hat{T}$. Let $I^0(\theta_q)$ denote the set of all sample observations such that $G(X;\delta) \le \kappa_q$ and similarly, define $I^1(\theta_{1-q})$ as the set of all sample observations such that $G(X;\delta) \ge \kappa_{1-q}$.

---

[2] See White (1982) for details. The results in the text and the appendix go through, with appropriate substitutions, for any estimator $\hat{\delta}$ such that $\sqrt{n}(\hat{\delta} - \delta^*)$ has the same limiting distribution as $K\sqrt{n}(\sum u(T,X;\delta)/n)$, where u is a mean zero function of the data fulfilling the conditions for the central limit theorem and K is a constant vector. Least

We now present our bounds estimates. Let $\hat{A}^0(\hat{\theta}_q) \equiv \frac{\sum_{i \in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))}$ be average observed T in $I^0(\hat{\theta}_q)$, where $\hat{\theta} \equiv [\hat{\delta}' \quad \hat{\kappa}_q]'$ and let $\hat{A}^1(\hat{\theta}_{1-q}) \equiv \frac{\sum_{i \in I^1(\hat{\theta}_{1-q})} (1 - T_i)}{\#(I^1(\hat{\theta}_{1-q}))}$. The statistics $\hat{A}^0(\hat{\theta}_q)$ and $\hat{A}^1(\hat{\theta}_{1-q})$ are obvious upper bound estimates for α_0 and α_1 and can be used to bound β. The following bounding results are straightforward:

Proposition 1:

(a) $\alpha_0 \leq \; plim\, \hat{A}^0(\hat{\theta}_q) = A^0(\theta_q^*)$

(b) $\alpha_1 \leq \; plim\, \hat{A}^1(\hat{\theta}_{1-q}) = A^1(\theta_{1-q}^*)$

(c) $p\lim \hat{\beta}_{ols} < \beta < p\lim \hat{\beta}_{ols}\chi(\hat{A}^0(\hat{\theta}_q), \hat{A}^1(\hat{\theta}_{1-q}), p, R) = p\lim \hat{\beta}_{ols}\chi(A^0(\theta_q^*), A^1(\theta_{1-q}^*), p, R)$,

where $\theta_q^* \equiv [\delta^{*\prime} \quad \kappa_q^*]'$.

We now discuss the asymptotic distributions of $\hat{A}^0(\hat{\theta}_q)$ and $\hat{A}^1(\hat{\theta}_{1-q})$ and how to use the estimates in constructing asymptotic confidence intervals for α_0, α_1, and β. We discuss $\hat{A}^0(\hat{\theta}_q)$; the properties of $\hat{A}^1(\hat{\theta}_{1-q})$ are symmetrical.

Decompose $(\hat{A}^0(\hat{\theta}_q) - A^0(\theta_q^*))$ into $(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q)) + (A^0(\hat{\theta}_q) - A^0(\theta_q^*))$. The asymptotic variance of the second term depends on the gradient vector of A^0 at θ_q^*, and is not straightforward to estimate. However, note that $A^0(\hat{\theta}_q)$ (the population mean T for the set defined by $\hat{T} \leq \hat{\kappa}_q$) is a valid bound for α_0. We can therefore treat $A^0(\hat{\theta}_q)$ as an

squares regression is one example of such an estimator.

object of estimation in its own right, and deal only with the first term in the decomposition. We show in the appendix, under weak additional regularity conditions, that the fact that $\hat{\theta}_q$ is estimated has no effect on the limiting distribution of $\sqrt{n}(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q))$, which takes the simple form $(0, A^0(\hat{\theta}_q)(1-A^0(\hat{\theta}_q))/q)$. In effect, we can treat $\hat{\theta}_q$ as fixed.

Constructing a confidence interval for $A^0(\hat{\theta}_q)$ is immediate. It is also straightforward to construct a conservative confidence interval for the parameter of interest, α_0. We summarize our results in the following proposition.

Proposition 2:

$$(7a) \qquad \lim_{n\to\infty} \Pr(\hat{A}^0(\hat{\theta}_q) - s(n,\hat{\theta})z_{1-r/2} \le A^0(\hat{\theta}_q) \le \hat{A}^0(\hat{\theta}_q) + s(n,\hat{\theta})z_{1-r/2}) = 1-r$$

$$(7b) \qquad \lim_{n\to\infty} \Pr(0 \le \alpha_0 \le \hat{A}^0(\hat{\theta}_q) + s(n,\hat{\theta})z_{1-r}) \ge 1-r\,.$$

where z_r denotes the rth percentile of the standard normal cdf and

$$s(n,\hat{\theta}) = \sqrt{\frac{\hat{A}^0(\hat{\theta}_q)(1-\hat{A}^0(\hat{\theta}_q))}{nq}}\,.$$

Let $\beta_{Ubound} \equiv p\lim \hat{\beta}_{ols}\chi(\alpha_0^{\max}, \alpha_1^{\max}, p, R)$ denote the theoretical upward bound on β, as determined by the measurement error bounds $\alpha_0^{\max}$ and $\alpha_1^{\max}$. Using the delta method, one can derive a confidence interval for β_{Ubound} similar to (7a) from the joint distribution of $\hat{\beta}_{ols}$, p, R, $\hat{A}^0(\hat{\theta}_q)$, and $\hat{A}^1(\hat{\theta}_{1-q})$. The joint distribution can be estimated using a stacked regression. For this purpose, it is convenient to let $\sigma^2_{T|X}$ denote the average variance of T conditional on X and to replace R in (4) with $\sigma^2_{T|X}$ using the relation $R = 1 - \frac{\sigma_T^2}{p(1-p)}$. Regressing the squared residual, e_T^2, from a regression of T on X on a constant yields the coefficient $\sigma_{T|X}^2$. In addition, let I_0 and I_1 be indicators of membership in the sets $I^0(\hat{\theta}_q)$ and $I^1(\hat{\theta}_{1-q})$, respectively, and let $V \equiv (1 - I_0 - I_1)/(1-2q)$. It can be verified that regressing T on V, $I_0 - qV$, and $I_1 - qV$ yields the coefficients p, $A^0(\theta_q)$, and $A^1(\theta_{1-q})$. This gives us the stacked regression:

$$\begin{bmatrix} Y \\ T \\ e_T^2 \end{bmatrix} = \begin{bmatrix} 1 & X & T & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & V & I_0 - qV & I_1 - qV & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} c \\ \gamma \\ \beta \\ p \\ A^0 \\ 1 - A^1 \\ \sigma_{T|X}^2 \end{bmatrix} + \begin{bmatrix} e \\ \varepsilon_T \\ \varepsilon_{e_T^2} \end{bmatrix}.$$

To construct a w-percent confidence interval for β analogous to (7b), note that $(\hat{\beta}_{ols}, \hat{\beta}_{Ubound})$ are distributed bivariate normal. Let $\Phi(c,d,r)$ denote the probability that $X_1 \leq c$ and $X_2 \leq d$, where X_1 and X_2 are standard normal random variables with correlation r. Let c_q and d_q satisfy $\Phi(c_w, d_w, -\rho) = w$, where ρ is the correlation between $\hat{\beta}_{Ubound}$ and $\hat{\beta}_{ols}$. We then have

$$(8) \quad w = \lim_{n\to\infty} \Pr(\hat{\beta}_{ols} - c_w\sigma_{ols} \leq p\lim \hat{\beta}_{ols}, p\lim \hat{\beta}_{Ubound} \leq \hat{\beta}_{Ubound} + d_w\sigma_{Ubound})$$

$$\leq \lim_{n\to\infty} \Pr(\hat{\beta}_{ols} - c_w\sigma_{ols} \leq \beta \leq \hat{\beta}_{Ubound} + d_w\sigma_{Ubound}),$$

where σ denotes the standard error of the subscripted estimator. A minimum length w-percent confidence interval can be constructed by choosing c_w and d_w to solve the following minimization problem:

$$\min_{c_w, d_w} \quad (d_w \sigma_{Ubound} + c_w \sigma_{ols}) \text{ subject to } \Phi(c_w, d_w, -\rho) = w.$$

The preceding analysis treats q as fixed. $Var(\hat{A}^0(\hat{\theta}_q))$ decreases with q since the subsample used to compute $\hat{A}^0(\hat{\theta}_q)$ increases with q. However, $Pr(T^*=1)$, and consequently, $\hat{A}^0(\hat{\theta}_q)$ will in general be increasing in q. The choice of q thus involves a trade-off between the tightness and the variance of the bounds. We leave the optimal choice of q as a topic for research.

Our upper bound on β does not take into account the constraint that the variance of e is bounded below by zero. Whether this constraint binds can be checked by examining the "reverse regression" coefficient generated by regressing T on Y and X. Let B denote the coefficient on Y and $\hat{\beta}_{rev} \equiv 1/B$ the slope implied by the reverse regression. Bollinger (1996) shows that the constraint binds only if $\hat{\beta}_{rev} < (1 - A^0(\hat{\theta}_q) - A^1(\hat{\theta}_q))\hat{\beta}_{Ubound}$. We refer the reader to the bounds in Bollinger (1996) for this case (which is unlikely to occur for low and moderate values of R^2 in the main regression).

Stronger prior information will of course yield tighter bounds. We noted above that knowledge of the functional form of $Pr(T^* = 1|X)$ can be used to estimate α_0 and α_1 from (6) by maximum likelihood, as in HAS.[3] We compare the percentile method with HAS at several points below. Without knowledge of $Pr(T^* = 1|X)$, but with knowledge

[3] More precisely, one can estimate upper bounds for α_0 and α_1. As HAS themselves note, their model is indistinguishable from one where $T^* = 1$ for a fraction of the population α_0, and $T^* = 0$ for a fraction α_1, independent of X. More generally, one can envision a mixture of these two extreme cases, where the proportions α_0 and α_1 are composed partly of those whose response is independent of X and partly of those who misreport, so true measurement error is bounded below by zero and above by the estimated values of α_0 and α_1.

of a functional form $h(X;\ \delta)$ such that $Pr(T=1|X) = G(h(X;\ \delta))$ for an unknown cdf G--that is, a single-index condition--one can consistently estimate δ up to a multiplicative constant as in Ichimura (1986) and Han (1987), and thus estimate bounds that are tight for a given q. Finally, if there is a set S such that it is known a priori that $Pr(T^*=1|\ X \in S) = 0$ (for example, knowledge that the training program whose impact is of interest is not offered at a given location), then $Pr(T=1|\ X \in S) = \alpha_0$, and correspondingly for α_1.

III. A GMM Estimator for Estimating the Effect of a Mismeasured Binary Explanatory Variable When Instruments are Available

Where instruments are available, instrumental variable estimation is a commonly prescribed fix to measurement error in a regressor. We begin this section by showing that IV estimation by itself is inconsistent for the coefficient of the mismeasured variable but consistent for other variables (assumed to be correctly measured). We next derive a consistent closed-form GMM estimator and then show how to incorporate the estimated measurement error bounds developed in the previous section into the GMM estimates. We also develop a specification test for the measurement error model.

<u>IV Estimation</u>

When $\text{cov}(T^*,U) = 0$, any variable Z which is correlated with T^* and independent of e and the measurement error process can be used as an instrument. However, if T^* is binary, T^* and U will be negatively correlated for any (non-degenerate) distribution of T^* and U. Because the classical assumption that T^* and U are uncorrelated cannot hold when T^* is binary, Z will not be a valid instrument. To be more precise, let Z be a vector of variables such that $\text{Cov}(Z,T^*) \neq 0$, $\text{Cov}(Z, e)=0$, and $\text{Cov}(Z,U|T^*)=0$. The last equality captures the idea that Z is independent of the measurement error process conditional on T^*, but since Cov(T^*,U) ≠ 0, this does not imply that the unconditional covariance Cov(Z,U)=0.

As others have demonstrated, the fact that Cov(Z,U)≠0 means that the use of Z as instruments will result in an inconsistent estimate of β.[4] Interestingly, as we now show, the IV estimate of γ is consistent. In anticipation of the discussion to follow, it is convenient to frame the analysis in terms of GMM estimation.

As is well known, the IV estimator $\begin{bmatrix}\hat{\beta}_{iv}\\ \hat{\gamma}_{iv}\end{bmatrix}$ is equivalent to the GMM estimator

$\begin{bmatrix}\hat{\beta}_1\\ \hat{\gamma}\end{bmatrix}$ whose probability limit is given by

$$(9)\qquad \begin{bmatrix} p\lim\hat{\beta}_1 \\ p\lim\hat{\gamma}\end{bmatrix} = \arg\min_{\tilde{\beta}_1,\tilde{\gamma}}\left[m_1(\tilde{\beta}_1,\tilde{\gamma}) \quad m_2(\tilde{\beta}_1,\tilde{\gamma})\right]' A\left[m_1(\tilde{\beta}_1,\tilde{\gamma}) \quad m_2(\tilde{\beta}_1,\tilde{\gamma})\right],$$

where

$$(10a)\qquad m_1(\tilde{\beta}_1,\tilde{\gamma}) = Cov(Z, Y-\tilde{\beta}_1 T - X\tilde{\gamma})$$

$$(10b)\qquad m_2(\tilde{\beta}_1,\tilde{\gamma}) = Cov(X, Y-\tilde{\beta}_1 T - X\tilde{\gamma}),$$

and A is a weighting matrix (in the case of TSLS, $A = Var(Z\ X)^{-1}$).[5] (The subscript on the estimate for β will be convenient in the discussion below where we obtain other estimates using additional moment conditions.) We now prove

---

[4] We follow previous papers (Loewenstein and Spletzer 1997, Berger, Black and Scott 1997, Kane, Rouse and Staiger 1999) in referring to estimators using the elements of Z as instruments as IV estimators, in spite of the fact that technically the Z variables are not instruments.

[5] Naturally, in carrying out the GMM estimation, the population moments are replaced by the sample moments.

Proposition 3: $p\lim\hat{\beta}_1 = k_1\beta$ and $p\lim\hat{\gamma}_{iv} = \gamma$, where $k_1 = 1/(1-\alpha_0-\alpha_1)$.

Proof:

It follows immediately from equation (1) that

(11a) $Cov(Z,Y) = \beta Cov(Z,T^*) + Cov(Z,X)\gamma$

(11b) $Cov(X,Y) = \beta Cov(X,T^*) + Var(X)\gamma$.

Letting r_{AW} denote the coefficient of correlation between two random variables *A* and *W*, and $r_{AW\cdot V}$ the partial correlation between *A* and *W* conditional on *V*, the coefficient of correlation between *X* and *U* conditional on *T** can be expressed as (Gujarati 1978, p. 112)

$$r_{XU\cdot T^*} = \frac{r_{XU} - r_{XT^*}r_{UT^*}}{\sqrt{(1-r_{XT^*}^2)(1-r_{UT^*}^2)}}.$$

Using the fact that $r_{XU\cdot T^*} = 0$, one finds that Var(*T**)Cov(*X*,*U*) = Cov(T*,U)Cov(*X*,*T**). This result together with (2) and the fact that Cov(*T**,*U*) = $-p^*(1-p^*)(\alpha_0+\alpha_1)$ gives us

(12) $Cov(X,U) = -(\alpha_0+\alpha_1)k_1 Cov(X,T)$.

From (12) and its analogue for *Z*, we have $Cov(Z,T^*)/Cov(Z,T) = Cov(X,T^*)/Cov(X,T) = k_1$. Substituting into (11) yields

(13a) $Cov(Z,Y) = k_1\beta Cov(Z,T) + Cov(Z,X)\gamma$

(13b) $Cov(X,Y) = k_1\beta Cov(X,T) + Var(X)\gamma$.

Substituting (13) into (10), we see that

(12*a*') $m_1(\tilde{\beta}_1,\tilde{\gamma}) = (k_1\beta - \tilde{\beta}_1)Cov(Z,T) + Cov(Z,X)(\gamma-\tilde{\gamma})$

(12*b*') $m_2(\tilde{\beta}_1,\tilde{\gamma}) = (k_1\beta - \tilde{\beta}_1)Cov(X,T) - Var(X)(\gamma-\tilde{\gamma})$, which implies that $p\lim \hat{\beta}_1 = k_1\beta$ and $p\lim \hat{\gamma}_2 = \gamma$. Q.E.D.

Although IV yields an inconsistent estimate of β, the proof of proposition 3 suggests a way that one might be able to obtain a consistent estimate. The GMM estimator using the moment conditions (11) yields an estimate of $\hat{\beta}_1$ that is a simple function of β and the measurement error parameters α_0 and α_1. If one can use additional moments that allow determination of α_0 and α_1, then one can solve for β.

GMM Estimation

Consider for the moment the case without covariates. KRS and BBS analyze a model with two mismeasured indicators T_1 and T_2 of T^*. They note that seven moments are observable: $E(Y \mid T_1 = i, T_2 = j)$, $\Pr(T_1 = i, T_2 = j)$, i,j={0,1} (one of the cell probabilities is redundant). This allows the identification of the seven parameters c, β, p^*, α_{0k} , α_{1k} , k={1,2}. Note that knowledge of $E(Y \mid T_1 = i, T_2 = j)$ and $\Pr(T_1 = i, T_2 = j)$ is equivalent to knowledge of the following sets of moments: E(Y), Cov(T_k,Y), Cov(T_1T_2,Y), E(T_k), and Cov(T_1,T_2). In our analysis, an instrument takes the place of one of the alternate measures, so that the moments are E(Y), Cov(T,Y), Cov(Z,Y), Cov(ZT,Y), E(T), E(Z), and Cov(Z,T). Thus, our model is identified and can be estimated using GMM.

Turning our attention back to the full model (1), note that we have already used the covariances between Z and Y in the moment conditions (13a) and the covariances between X and Y in the moment conditions (13b). To use Cov(T,Y) and Cov(ZT,Y), we note from (1) that

$$(13c) \quad Cov(T,Y) = k_2\beta Var(T) + Cov(T,X)\gamma$$
$$(13d) \quad Cov(W,Y) = k_3\beta Cov(W,T) + Cov(W,X)\gamma$$

where $W \equiv (Z - \bar{Z}) \cdot T$, $k_2 \equiv Cov(T,T^*) / Var(T)$, and $k_3 \equiv Cov(W,T^*) / Cov(W,T)$.[6] Algebra establishes that:

$$(14) \quad k_2 = \frac{(p-\alpha_0)(1-p-\alpha_1)}{p(1-p)(1-\alpha_0-\alpha_1)};$$

$$k_3 = \frac{(1-p-\alpha_1)+\alpha_0}{(1-p)(1-\alpha_0-\alpha_1)}.$$

To estimate $\beta_2 \equiv k_2\beta$ and $\beta_3 \equiv k_3\beta$, we expand the IV moment conditions (10) to include:

$$(10c) \quad m_3(\tilde{\beta}_2, \tilde{\gamma}) = Cov(T, Y - \tilde{\beta}_2 T - X\tilde{\gamma})$$

$$(10d) \quad m_4(\tilde{\beta}_3, \tilde{\gamma}) = Cov(W, Y - \tilde{\beta}_3 T - X\tilde{\gamma}).$$

Since the factors k_2 and k_3 are both functions of p as well as α_0 and α_1, we need an estimate of p to close the model, so we add a final moment condition:

$$(10e) \quad m_5 = E(T) - \tilde{p}.$$

Substituting (13c) and (13d) into (10) yields

$$(12a') \quad m_1(\tilde{\beta}_1, \tilde{\gamma}) = (k_1\beta - \tilde{\beta}_1)Cov(Z,T) + Cov(Z,X)(\gamma - \tilde{\gamma})$$
$$(12b') \quad m_2(\tilde{\beta}_1, \tilde{\gamma}) = (k_1\beta - \tilde{\beta}_1)Cov(X,T) - Var(X)(\gamma - \tilde{\gamma})$$
$$(12c') \quad m_3(\tilde{\beta}_2, \tilde{\gamma}) = (k_2\beta - \tilde{\beta}_2)Var(T) + Cov(T,X)(\gamma - \tilde{\gamma})$$
$$(12d') \quad m_4(\tilde{\beta}_3, \tilde{\gamma}) = (k_3\beta - \tilde{\beta}_3)Cov(W,T) + Cov(W,X)(\gamma - \tilde{\gamma})$$
$$(12e') \quad m_5(\tilde{p}) = E(T) - \tilde{p}.$$

This leads to

[6] Note the definition of W. The formula for k_3 below is correct only if Z is normalized to zero before multiplying by T.

Proposition 4: Let $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{p}$ be GMM estimates using moments (12). Then

(15a) $$\hat{\beta} = \sqrt{4\hat{p}(1-\hat{p})\hat{\beta}_1\hat{\beta}_2 + ((1-\hat{p})\hat{\beta}_3 - \hat{p}\hat{\beta}_1)^2}$$

(15b) $$\hat{\alpha}_0 = \frac{\hat{p}\hat{\beta}_1 + (1-\hat{p})\hat{\beta}_3 - \hat{\beta}}{2\hat{\beta}_1}$$

and

(15c) $$\hat{\alpha}_1 = \frac{(2-\hat{p})\hat{\beta}_1 - (1-\hat{p})\hat{\beta}_3 - \hat{\beta}}{2\hat{\beta}_1}$$

are consistent estimators for β, α_0 and α_1.[7]

Proof:

From (12a') - (12e'), it is clear that any GMM estimator using moment conditions (10a) – (10e) will have:

(16*a*) $p\lim\hat{\beta}_1 = p\lim\hat{\beta}_{iv} = k_1\beta$

(16*b*) $p\lim\hat{\gamma} = p\lim\hat{\gamma}_{iv} = \gamma$

(16*c*) $p\lim\hat{\beta}_2 = k_2\beta$

(16*d*) $p\lim\hat{\beta}_3 = k_3\beta$

(16*e*) $p\lim\hat{p} = p.$

Substituting (14) and Proposition 3 into (16) and solving for β, α_0 and α_1 yields the consistency result. Q.E.D.

An optimal GMM estimator can be derived as follows. Let

---

[7] Note that $\text{sgn}(p\lim(\hat{\beta}_1)) = \text{sgn}(p\lim(\hat{\beta}_2)) = \text{sgn}(p\lim(\hat{\beta}_3)) = \text{sgn}(\beta)$ since k_1, k_2, and k_3 are all greater than zero. Allowing $\beta < 0$, (15a) becomes
$\hat{\beta} = \text{sgn}(\hat{\beta}_1)\sqrt{4\hat{p}(1-\hat{p})\hat{\beta}_1\hat{\beta}_2 + ((1-\hat{p})\hat{\beta}_3 - \hat{p}\hat{\beta}_1)^2}$.

$$H \equiv \begin{bmatrix} z & x & 0 & 0 & 0 \\ 0 & 0 & t & 0 & 0 \\ 0 & 0 & 0 & w & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},\ \Omega \equiv \begin{bmatrix} y \\ y \\ y \\ y \end{bmatrix},\ \Gamma \equiv \begin{bmatrix} t & 0 & 0 & x & 0 \\ 0 & t & 0 & x & 0 \\ 0 & 0 & t & x & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix},\ \Pi^0 \equiv \begin{bmatrix} \beta_1^0 \\ \beta_2^0 \\ \beta_3^0 \\ \gamma^0 \\ p^0 \end{bmatrix}, \text{ and } \hat{\varepsilon} \equiv H'[\Omega - \Gamma\Pi^0],$$

where Π^0 is some initial consistent estimate of Π and where *z, x, y,* and *t* devote the deviations of *Z, X, Y,* and *T* from their sample means. Write the sample moments (10) in stacked form:

$$(17) \qquad \overline{m} \equiv (1/n)H'[\Omega - \Gamma\Pi].$$

The optimal GMM estimator minimizes $\overline{m}V^{-1}\overline{m}$, where *V* is the asymptotic variance matrix of $\overline{m}$ (Hansen 1982). Letting $[\Omega - \Gamma\Pi^0]' \equiv \begin{bmatrix} \varepsilon^1 & \varepsilon^2 & \varepsilon^3 & \varepsilon^4 \end{bmatrix}'$, note that ε^i is not generally homoscedastic. For example, $\varepsilon_i^1 = y_i - \tilde{\beta}_1 t_i - x_i\tilde{\gamma} = \beta t_i^* - \tilde{\beta}_1 t_i - x_i\tilde{\gamma} + e_i^1$, so that $Var(\varepsilon^1 \mid t) = \beta^2 Var(T^* \mid T = t + \overline{T})$ $+ Var(e)$ will not in general be constant across *t*. The optimal GMM estimator given the moment conditions (10) and taking into account heteroscedasticity is $\hat{\pi} = (\Gamma' H S^{-1} H'\Gamma)^{-1}\Gamma' H S^{-1} H'\Omega$, where $S = \sum (H_i' \hat{\varepsilon}_i \hat{\varepsilon}_i' H_i)$; see Wooldridge (1996).

Note that the asymptotic distribution for $\sqrt{n}\begin{bmatrix} \hat{\beta} - \beta \\ \hat{\gamma} - \gamma \end{bmatrix}$ is

$N(0, \begin{bmatrix} d & 0 \\ 0 & 1 \end{bmatrix}'(\Gamma' H S^{-1} H'\Gamma)^{-1}\begin{bmatrix} d & 0 \\ 0 & 1 \end{bmatrix})$, where $d' \equiv \begin{bmatrix} \partial\hat{\beta}/\partial\hat{\beta}_1 & \partial\hat{\beta}/\partial\hat{\beta}_2 & \partial\hat{\beta}/\partial\hat{\beta}_3 & \partial\hat{\beta}/\partial\hat{p} \end{bmatrix}$ can be obtained from (15a). If desired, $\hat{c}$ may be calculated as $\overline{Y} - \hat{\beta}\hat{p}^* - \overline{X}\hat{\gamma}$.

Finally, we might point out that there is one substantive difference between the situation in which there is more than one imperfect measure of a binary explanatory variable and that where one or more instruments take the place of one of the alternative measures. In the former case, Cov(X,T_1)Cov(T_2,T^*) = Cov(X,T_2)Cov(T_1,T^*) since Cov(X,T_i) = Cov(X,T^*)Cov(T_i,T^*) for i = 1,2. KRS use this restriction to obtain additional identifying information. Since Cov(X,Z) is unrestricted, there is no analogous restriction when instruments take the place of one of the measures.[8]

Incorporating Restrictions on the Measurement Error Parameters Into the GMM Estimation

Estimates of α_0 and α_1 based on (15b-c) are not guaranteed to be between 0 and 1. Additionally, they are not guaranteed to be less than bounds derived from the HAS or percentile methods in Section II. Bounds on the measurement error parameters can be incorporated into the GMM estimation procedure.[9] We first discuss imposing bounds using the percentile method, than briefly discuss using the HAS bounds.

Let $A_i \equiv p\lim \hat{A}^i(\hat{\theta}_q)$, $D_i \equiv A_i - \alpha_i$, and $\hat{D}_i \equiv \hat{A}^i(\hat{\theta}_q) - \hat{\alpha}_i$ The restriction that the measurement error parameters be between zero and the bounds A_i can be expressed as $L \geq 0$, where $L \equiv \begin{bmatrix} D_0 & D_1 & \alpha_0 & \alpha_1 \end{bmatrix}$. Note that these restrictions imply that that $\alpha_i \leq 1$.

Let Π denote the vector of parameters [π A_0 A_1]'. One can directly incorporate the percentile bounds into the estimation procedure by extending the moment conditions (10) to include a regression of T on a constant over the subset of the data with $X\hat{\delta} \leq \hat{\kappa}_q$ to estimate A_0, and a regression of $1 - T$ on a constant over the subset of the data with

[8] If one makes the stronger assumption that e is not just uncorrelated with, but is independent of X and T^*, additional moment conditions analogous to (13d) can be imposed using Cov(xT ,Y). More broadly, if any interactions between elements j of X and T^* can be excluded from (1), Cov(x_jT, Y) can be used as in (13d).

[9] BBS do not take into account the bounds on α_0 and α_1 in their GMM procedure. KRS account for it by parameterizing α_0 and α_1 to be between zero and one, resulting in a non-linear GMM procedure.

$X\hat{\delta} \geq \hat{\kappa}_{1-q}$ to estimate A_1, similar to the stacked regressions above. Denote this extended moment set m_a.

The classical approach to point estimation is to estimate Π by minimizing the weighted sum of squares $m_a' V_a^{-1} m_a$ subject to $\hat{L} \equiv \left[\hat{D}_0 \quad \hat{D}_1 \quad \hat{\alpha}_0 \quad \hat{\alpha}_1\right] \geq 0$. Note that the constraints are non-linear since the $\hat{\alpha}_i$ estimates are non-linear functions of the regression parameters. Incorporating inequality constraints into classical inference presents challenges. The asymptotic distribution depends on whether the true parameter is on the boundary of the feasible set. If the true parameter is in the interior of the feasible set, then the asymptotic distribution is equal to that of the equivalent unconstrained estimator, but this may be a poor guide to finite-sample behavior. The bootstrap, a common method of improving finite-sample performance of variance estimates, is inconsistent in inequality-constrained problems (Andrews 2000).

We believe that Geweke's (1986) Bayesian method presents the simplest satisfactory approach to this problem. Geweke shows that if the prior for the parameter vector is diffuse over the feasible set, the posterior distribution is the portion of the estimated sampling distribution of the unrestricted parameter estimate in the feasible region. The posterior mean can be evaluated by taking random draws from the sampling distribution and averaging over draws in the feasible region. Applied to our problem, one first uses GMM to obtain an unconstrained estimate $\hat{\Pi}$. One then takes draws from the distribution $N(\hat{\Pi}, V(\hat{\Pi}))$ and averages over those draws where $\hat{L} \geq 0$.

A Specification Test for the Measurement Error Model.

The foregoing assumes that the data were generated by (1) (i.e., that e is uncorrelated with T^* and X), that the measurement error is uncorrelated with X and Z

conditional on T^*, and that the instruments Z are valid. If these assumptions are violated, it is not necessarily the case that $p\lim \hat{L} \geq 0$. Thus we can use a test of the null hypothesis that $L \geq 0$ as a specification test of the measurement error model.

Let δ denote the 4 x 1 vector $(\hat{L}^R - \hat{L})$', where $\hat{L}^R$ denotes the inequality-restricted estimator of L. Letting δ_A denote the subvector of δ corresponding to all the binding constraints and letting $V_A(\hat{L})$ denote the corresponding elements of the variance matrix of $\hat{L}$, construct the test statistic

(18) $\tau = \delta_A' V_A(\hat{L})^{-1} \delta_A$.

We now show that asymptotically the $1-q$ percentile of the chi-square distribution with 2 degrees of freedom is a valid critical value for a test with significance level q.

Proposition 5:

$$\lim_{n\to\infty} \Pr(\tau \leq c \mid L \geq 0) \geq \chi_2^2(c) \text{ and } \lim_{n\to\infty} \Pr(\tau \leq c \mid L = 0) = \chi_2^2(c)$$

Proof:

The $\hat{A}^i(\hat{\theta}_q)$ parameters must be non-negative since they are sample means of non-negative variables. Thus, no more than two out of the four elements of $\hat{L}$ can violate the inequality constraints and δ_A will never have more than two elements.

All four constraints are satisfied with equality when $\alpha_0 = A_0 = \alpha_1 = A_1 = 0$. Asymptotically, $Var(\hat{A}^i(\hat{\theta}_q)) = A_i(1 - A_i) = 0$ at this point, so that the distribution of $\hat{D}_i$ converges to the distribution of $-\hat{\alpha}_i$ and the restrictions collapse to $[-\hat{\alpha}_0 \quad -\hat{\alpha}_1 \quad \hat{\alpha}_0 \quad \hat{\alpha}_1] \geq 0$, or equivalently, $\hat{\alpha}_0 = \hat{\alpha}_1 = 0$. Whatever combinations of constraints are violated in the sample, the statistic τ thus equals

$[\hat{\alpha}_0 \quad \hat{\alpha}_1] Var(\hat{\alpha})^{-1} [\hat{\alpha}_0 \quad \hat{\alpha}_1]'$, which is asymptotically distributed $\chi^2(2)$.[10]

Now consider the case where two constraints are satisfied with equality, one corresponding to α_0 and one to α_1. If $\alpha_i = 0$, let $\lambda_i = \alpha_i$ and $\mu_i = D_i$. Conversely, if $D_i = 0$, let $\lambda_i = D_i$ and $\mu_i = \alpha_i$. As the sample becomes large, the probability that the inequality constraints involving λ_i are violated approaches zero. We therefore need only consider cases where the constraints involving α_i are violated in the sample. If no constraints are violated in the sample, τ = 0. If one constraint *i* is violated, $\tau = \hat{\lambda}_i^2 / Var(\hat{\lambda}_i)$, which is distributed $\chi^2(1)$. If both constraints are violated, $\tau = \left[\hat{\lambda}_0 - \hat{\lambda}_0^R \;\; \hat{\lambda}_1 - \hat{\lambda}_1^R\right]$ $Var(\hat{\lambda})^{-1}\left[\hat{\lambda}_0 - \hat{\lambda}_0^R \;\; \hat{\lambda}_1 - \hat{\lambda}_1^R\right]'$, which is precisely Wolak's (1987) test statistic *W*. Wolak (1987) shows that *W* is equal to the distance between the inequality-restricted and unrestricted estimates, evaluated in the norm of the covariance matrix of the unrestricted estimates. This distance must be less than or equal to the distance Δ between the unrestricted estimates and the estimates obtained imposing the equality restrictions $\lambda_0 = \lambda_1 = 0$. Δ is distributed $\chi^2(2)$. Thus τ is a mixture of random variables all of which are less than or equal to $\chi^2(2)$.

Applying a similar argument for cases with other combinations of constraints holding with equality, we can conclude that the point where all four constraints are

---

[10] Note that the more traditional chi-square statistic $\delta' Var(\hat{L})^{-1} \delta$, as in Wolak (1991), does not exist at this point, as $Var(\hat{L})$ is singular.

satisfied with equality is the least favorable point for the purpose of computing a distribution of the test statistic under the null hypothesis.[11, 12] Q.E.D.

Applying the HAS bounds

The application of HAS bounds to the Geweke technique is immediate. However, the distribution of the test statistic τ under the null is not easily derived under the HAS method. Negative estimates of A_0 and A_1 are possible, so all four constraints may be violated at the same time. As explained in Wolak (1991), the distribution of τ at the point where all constraints bind is a weighted mixture of chi-squares from zero to four degrees of freedom, with the weights depending on $Var(\hat{\lambda})$. Since $Var(\hat{\lambda})$ will vary across points in the null due to the non-linearity of the constraints, one can no longer show that this point is the least favorable point in the null.

IV. Bounding the Effect of a Mismeasured Endogenous Binary Explanatory Variable

Now suppose the residual *e* in (1) is uncorrelated with *X* and *Z*, but correlated with *T**. Projecting *T* on *T** and *Z*, we have: $T = \alpha_0 + (1 - \alpha_0 - \alpha_1)T^* + \eta$, where η is orthogonal to *T** and *Z*. Thus, $Cov(T,e) = (1 - \alpha_0 - \alpha_1)Cov(T^*,e)$ and $Cov(W,e) =$

---

[11] Our constraints involve $\hat{\alpha}_0$ and $\hat{\alpha}_1$, which are nonlinear functions of $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$, and $\hat{p}$. While it is always true for linear constraints that the least favorable point in the null occurs where all constraints are satisfied with equality, we needed to demonstrate this here because it is not always true when the constraints are non-linear; see Wolak (1991).

[12] The statistic τ, while it involves estimating $\hat{L}^R$ subject to non-linear constraints, is in many cases easy to calculate. Cases with zero or one constraint binding were dealt with in the text. If two constraints are violated, and if each constraint is still violated after imposing the other constraint as an equality and calculating the constrained optimum, then $\hat{L}^R = 0$. One can also show that the constraint that $\alpha_0 \geq 0$ is equivalent to the linear constraint $\beta_2 \leq \beta_3$.

$(1-\alpha_0-\alpha_1)Cov(zT^*,e)$, which means that the population moment conditions (13c) and (13d) become:

$$(13c') \quad Cov(T,Y) = k_2\beta Var(T) + Cov(T,X)\gamma + (1-\alpha_0-\alpha_1)Cov(T^*,e)$$
$$(13d') \quad Cov(W,Y) = k_3\beta Cov(W,T) + Cov(W,X)\gamma + (1-\alpha_0-\alpha_1)Cov(zT^*,e).$$

By assumption, *Cov*(*T**,*e*) is not equal to zero. What about *Cov*(*zT**,*e*)? Projecting *T** onto *z, X,* and *e*, we have: $T^* = \delta_0 + \delta_1 X + \delta_2 z + \delta_3 e + \omega$, where ω is orthogonal to *X, Z,* and *e*. So

$$(19) \quad Cov(zT^*,e) = Cov(ze,T^*) = \delta_1 E(zXe) + \delta_2 E(z^2 e) + \delta_3 E(ze^2) + Cov(ze,\omega)$$

Independence of X and Z with e, not just orthogonality, is required to guarantee that the first three terms on the right hand side of (19) are zero. Cov(ze,ω) will not in general be zero except under the strong assumption that E(*T**| *X, Z, e*) is linear in *X, Z,* and *e*--that is, that the linear probability model applies to *T**.

We can conclude that endogeneity of *T** adds two more sets of unknown parameters to the moment conditions (13) -- *Cov*(*T**,*e*) and *Cov*(*zT**,*e*). Consequently, the GMM method described above is now underidentified. However, note that Proposition 3 still holds, so that $p\lim\hat{\beta}_{iv}$ still equals $\beta/(1-\alpha_0-\alpha_1)$ (and $p\lim\hat{\gamma}_{iv}=\gamma$). Under the maintained assumption that $\alpha_0+\alpha_1<1$, $\hat{\beta}_{iv}$ is asymptotically still an upper bound, and zero is a lower bound. This lower bound can be tightened by employing the HAS or the percentile method to obtain upper bounds $\alpha_0^{\max}$ and $\alpha_1^{\max}$ for the measurement error parameters α_0 and α_1.[13] Specifically, one has:

$$p\lim\hat{\beta}_{iv}(1-\alpha_0^{\max}-\alpha_1^{\max}) < \beta < p\lim\hat{\beta}_{iv}.$$

---

[13] Naturally, the instruments *Z* can be used in estimating *Pr*(*T*=1|*X, Z*) (for the percentile

Confidence intervals for the lower bound can be generated as in Section II.

V. Empirical Example

We now illustrate the use of our measurement error techniques with an analysis of the effect of training incidence on wage growth. Employee training is a particularly interesting application of these techniques because there is evidence that it is measured with a great deal of error. Using a survey of matched employer-employee responses to the same training questions, Barron, Berger, and Black (1997) find that the correlation between worker reported training and employer reported training is only .32 for the incidence of on-site formal training, and .38 for off-site formal training.

We use data from the National Longitudinal Survey of Youth 1979 Cohort (NLSY79). NLSY79 is a dataset of 12,686 individuals who were aged 14 to 21 in 1979. These youth have been interviewed annually since 1979, and the response rate has been 90 percent or greater in each year. We use data from the 1987 through 1994 surveys. Our dependent variable is the change in real log wages between interviews. We exclude job changers, so all wage growth is within-job. We also exclude the military subsample, observations with real wages below $1 or above $100 in 1982-84 dollars, and observations where the respondent is an active member of the armed forces, self-employed, in a farm occupation, or enrolled in school. Finally, we exclude observations where variables have missing values (except for the cases noted below where we use missing indicators).

---

method) or $Pr(T^*=1|X, Z)$ (for HAS) to help tighten the measurement error bounds.

Our training measure equals one if the respondent reports completing a training program on the current job since the last interview (the training may have started before the last interview) and zero otherwise. Our control variables are age, tenure, experience, the Armed Forces Qualifying Test measure of cognitive skills (AFQT),[14] and dummies for female, black, Hispanic, ever married, one-digit occupational categories, collective bargaining, part-time status, and calendar year. In addition, there are dummies for missing AFQT, collective bargaining status, and part-time status, with the variables set equal to zero if their missing indicators equal one. Our final sample has 20,300 observations from 8,031 jobs and 6,345 individuals. The observed incidence of training is 12.9 percent.

We use two instruments. The first is a measure of gross job destruction and creation by 2 digit industry created from Michigan unemployment insurance data.[15] The second is years of completed schooling. Our first instrument can be justified by the fact that the employer's return to a given training-induced increase in worker productivity is higher for a longer-lived job match (see Royalty 1996). Job creation and destruction rates are plausibly related to the magnitude of exogenous demand shocks and hence exogenously shift turnover rates. The second instrument is motivated by the consideration that years of school is an indicator of trainability. In this context, note that the inclusion of AFQT and one-digit occupation dummies in the wage growth equation arguably controls for the direct effect of schooling on productivity growth. We also include runs where the reallocation variable serves as the sole instrument (rows denoted "1 instrument" in the table).

[14] Specifically, the residual from a regression of AFQT on dummies for year of birth.

To illustrate the effects of our measurement error bounds on estimates of γ, we also show OLS bounds and IV estimates for the coefficient on AFQT. Our uncorrected OLS results indicate a statistically significant link between wage growth and AFQT. Given that we find (as is common in this literature) a strong effect of AFQT on the probability of receiving training, it is of interest to what extent the observed AFQT effect on wage growth may be due to measurement error in training.

Means and standard deviations for the variables of interest are shown in table 1. Point estimates and bounds for the wage return to training, β, for the measurement parameters, α_0, α_1, and for γ_{afqt} are shown in table 2. All standard errors are from a panel version of the White heteroscedasticity-consistent estimator (see Froot 1989 and Rogers 1993). Table 3 shows minimum length confidence intervals as in (8).

The OLS results indicate that training during a period raises next period wages by 1.9 percent. IV estimation raises the training coefficient dramatically, to between 13 and 14 percent. This increase is consistent with the hypothesis of substantial measurement error.

To generate percentile bounds, we first estimate a probit of training incidence on the independent variables and instruments and then observe the incidence of training below the 5$^{\text{th}}$ and above the 95$^{\text{th}}$ percentiles of the distribution of predicted training. The bound for α_0 is relatively tight at 2.4 percent. The bound for α_1, 70 percent, is much higher, indicating the potential for a great deal of measurement error. The relative magnitudes of these bounds are intuitively plausible, as it seems more likely that respondents would forget or neglect to report training spells rather than to report training

---

[15] We thank Chris Foote for supplying these data. See Foote (1998) for more details.

that did not occur. We generate HAS bounds by estimating (6) taking *F* to be the normal cdf. The HAS bounds are slightly tighter than the percentile bounds.

Applying the bounds on α_0 and α_1 to the OLS results using (5) yields upper bounds on β of about 4 percent for both the percentile and HAS methods, roughly double the OLS estimate. [16] The confidence intervals from (8) are similar to, but slightly to the left of what one would get by adding 1.96 σ_{Ubound} to the upper bound estimate and subtracting 1.96 σ_{ols} from the OLS estimate. Because $\sigma_{Ubound} > \sigma_{ols}$, taking a smaller multiple of σ_{Ubound} and a larger multiple of σ_{ols} allows us to shorten the confidence interval slightly.

GMM results based on (17) using both instruments give a point estimate of 4 percent, similar to the bounds.[17] However, the GMM estimates of α_0 and α_1 are infeasible, with α_0 negative and α_1 above the percentile (and hence HAS) bound. We apply the Geweke technique, generating 10,000 draws in the feasible region, to produce feasible estimates of α_0 and α_1. Relative to GMM, the Geweke method does not affect the point estimate of β very much, but reduces the standard error from .008 to .007.[18] The results are similar when only the reallocation variable is used as an instrument. However, the improvement in precision from using the Geweke technique is more

---

[16] The reverse-regression upper bound on β from Bollinger (1996), referred to in Section II, is 22.2, corresponding to values of (α_0, α_1) of (0.123, 0). The much tighter bound produced by our method stems from the fact that, as explained in Bollinger (1996, p. 396), information reducing misclassification from the larger to the smaller group – in this case, α_0 - is particularly powerful in reducing the upper bound.

[17] Adding an overidentifying instrument generates two overidentifying moments (15a) and (15d). The test statistic for the conventional GMM overidentification test (Greene 2000, p. 482) is 1.64, far from the $\chi^2(2)$ critical values.

[18] Of course, the GMM estimator and the Geweke estimator are not strictly comparable unless one considers GMM as a Bayesian estimator combining the data and a diffuse

dramatic, as the GMM standard error for β with only one instrument is .012 while the Geweke standard error is still .007 to three decimal places.

To test whether the measurement error model underlying the GMM and Geweke estimates is compatible with the data, we calculate the test statistic τ in (18) (using the percentile bounds). With two instruments, τ is 2.68, which is not significant at conventional levels using a $\chi^2(2)$ distribution (the 5% critical value is 5.99). Using one instrument, τ is zero, as the GMM values are feasible.

The results for γ_{afqt} show a substantial effect of measurement error. The OLS coefficient is statistically significant at the 1 percent level whether highest grade completed is included in the regression or not. However, IV estimation reduces the coefficient by at least two-thirds; the estimated coefficient is less than its standard error. Recall that γ is consistently estimated by IV. The OLS bounds using the estimated α parameters are consistent with the low IV estimates, with the percentile bounds slightly lower than the IV estimates and the HAS bounds quite close.

Finally, if we allow for endogeneity, the lower bounds for β from IV estimation range from .036 to .041 across the different specifications, though with relatively large standard errors. These lower bounds are similar to the upper bounds from OLS and to the point estimates from the GMM and Geweke methods. The length of the confidence intervals reflects the large standard errors, σ_{Lbound}, on the lower bounds estimates; as above, they are to the left of intervals generated by subtracting 1.96 σ_{Lbound} to the bound and adding 1.96 σ_{iv} to the IV estimate.

---

prior.

Our prior expectation was that training incidence may be positively associated with unobservable determinants of wage growth. However, the bounds estimates indicate that true training is either exogenous or negatively correlated with the wage growth residual, as the lower bound estimates allowing for endogeneity coincide with the point estimates when one takes training to be exogenous. We conclude that our evidence is consistent with measurement error cutting the OLS estimate of the return to training in half, but we cannot rule out an additional downward bias to OLS due to endogeneity. Much of the apparent effect of AFQT on wage growth appears to be due to measurement error in training.

VI. Conclusion

This paper has explored techniques for dealing with a mismeasured binary explanatory variable in a linear regression. If the binary variable is measured with error, is uncorrelated with the error term in the regression, and there is no instrument available, then its true coefficient, β, can be bounded by combining the least squares coefficient with the HAS or percentile method for bounding measurement error presented in Section II. If an instrument is available, IV is inconsistent, but β can be consistently estimated by the GMM estimator in Section III. The estimated measurement bounds can be incorporated into the GMM estimates, and the specification can be tested by comparing the GMM estimates with the measurement error bounds. Finally, if the mismeasured binary explanatory variable is correlated with the error term in the regression, the GMM estimator is inconsistent, but β can be bounded by applying the HAS or percentile measurement error bounds to the IV estimate.

References

Aigner, Dennis J. 1973. "Regression with a Binary Independent Variable Subject to Errors of Observation." *Journal of Econometrics* 1, pp. 49-60.

Andrews, Donald W. K. 2000. "Inconsistency of the Bootstrap When a Parameter Is On the Boundary of the Parameter Space." *Econometrica* 68:2, pp. 399-405.

Barron, John M., Mark C. Berger, and Dan A. Black. 1997. "How Well do We Measure Training?" *Journal of Labor Economics* 15:3 (part 1), pp. 507-528.

Berger, Mark, Dan Black, and Frank Scott. 2000. "Bounding Parameter Estimates with Non-Classical Measurement Error." *Journal of the American Statistical Association* 95:451, pp. 739-48.

Black, Dan, Seth Sanders, and Lowell Taylor. 2000. "Measurement of Higher Education in the Census and CPS". Paper presented at Southern Economic Association Meetings, Arlington VA.

Bollinger, Christopher R. 1996. "Bounding Mean Regressions When a Binary Regressor is Mismeasured." *Journal Of Econometrics* 73:2, pp. 387-399.

Foote, Christopher. 1998. "Trend Employment Growth and the Bunching of Job Creation and Destruction." *Quarterly Journal of Economics* 113:3, pp. 809-34.

Froot, Kenneth A. 1989. "Consistent Covariance Matrix Estimation with Cross-Sectional Dependence and Heteroskedasticity in Financial Data." *Journal of Financial and Quantitative Analysis* 24:3, pp. 333-55.

Geweke, John. 1986. "Exact Inference in the Inequality Constrained Normal Linear Regression Model." *Journal of Applied Econometrics* 1:2, pp. 127-141.

Greene, William H. 2000. *Econometric Analysis,* 4th edition. Upper Saddle River, New Jersey: Prentice Hall.

Gujarati, Damodar. 1978. *Basic Econometrics.* New York: McGraw-Hill.

Han, A. K. 1987. "Non-parametric analysis of a generalized regression model." *Journal of Econometrics* 35, pp. 303-16.

Hansen, L. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50, pp. 1029-54.

Hausman, J. A., Abrevaya, Jason, and F. M. Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-Response Setting." *Journal Of Econometrics* 87, pp. 239-69.

Ichimura, H. 1993. "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models." *Journal of Econometrics* 58, pp. 71-120.

Kane, Thomas J., Cecilia Elena Rouse and Douglas Staiger. 1999. "Estimating Returns to Schooling when Schooling is Misreported." Working Paper 7235, National Bureau of Economic Research.

Koenker, Roger, and Gilbert Bassett Jr. 1982. "Robust Tests for Heteroscedasticity Based on Regression Quantiles." *Econometrica* 50:1, pp. 43-61.

Loewenstein, Mark A. and James. R. Spletzer. 1997. "Belated Training: The Relationship Between Training, Tenure, and Wages." Unpublished paper, Bureau of Labor Statistics.

Rogers, W. H. 1993. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 13, pp. 19-23.

Royalty, Anne Beeson. 1996. "The Effects of Job Turnover on the Training of Men and Women." *Industrial and Labor Relations Review* 49:3, pp. 506-21.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1, pp. 1-25.

Wolak, Frank A. 1987. "An Exact Test for Multiple Inequality and Equality Constraints in the Linear Regression Model." *Journal of the American Statistical Association* 82:399, pp. 782-93.

Wolak, Frank A. 1991. "The Local Nature of Hypothesis Tests Involving Inequality Constraints in Nonlinear Models." *Econometrica* 59:4, pp. 981-95.

Wooldridge, Jeffrey M. 1996. "Estimating Systems of Equations with Different Instruments for Different Equations." *Journal of Econometrics* 74, pp. 387-40

Table 1

Descriptive Statistics, Selected Variables

Variable	Mean	Std. Deviation	Min.	Max.
Change in Ln Wage	0.025	0.225	-1.492	1.487
Training Incidence	0.129	0.335	0	1
AFQT	0.7	20.3	-65.5	45.9

Table 2

Estimates and Bounds for Selected Coefficients and Measurement Error Parameters

Parameter	β	α_0	α_1	γ_{afqt} (x 10^{-4})
Specification				
Point Estimates				
OLS, 2 instruments	.019 (.005)			3.24 (0.81)
OLS, 1 instrument	.019 (.005)			2.37 (0.87)
IV, 2 instruments	.140 (.045)			0.72 (1.22)
IV, 1 instrument	.133 (.064)			0.79 (1.27)
GMM(2)	.040 (.008)	-.047 (.046)	.757 (.029)	
Geweke(2) percentile[1]	.038 (.007)	.008 (.006)	.634 (.082)	
Geweke(2) HAS	.038 (.007)	.006 (.004)	.636 (.072)	
GMM(1)	.039 (.012)	.020 (.053)	.688 (.091)	
Geweke(1) percentile[1]	.037 (.007)	.014 (.008)	.599 (.124)	
Geweke(1) HAS	.036 (.006)	.009 (.006)	.587 (.124)	
Upper Bounds[2]				
Percentile[1]		.024 (.005)	.702 (.017)	
HAS		.015 (.003)	.695 (.008)	
OLS percentile bound[1] (2)	.043 (.011)			0.39 (1.67)
OLS percentile bound[1] (1)	.043 (.011)			0.44 (1.36)
OLS HAS bound (2)	.038 (.009)			0.88 (1.10)
OLS HAS bound (1)	.038 (.010)			0.79 (1.16)
Lower Bounds				
Endogenous T*, IV(2) percentile bound[1]	.038 (.012)			
Endogenous T*, IV(2) HAS bound	.041 (.013)			
Endogenous T*, IV(1) percentile bound[1]	.036 (.017)			
Endogenous T*, IV(1) HAS bound	.039 (.018)			

[1] Standard error conditional on $\hat{\theta}$ (see Section II).

[2] Lower bound for γ_{afqt} corresponding to upper bound for β.

Table 3

Minimum Length 95% Confidence Intervals for Training Coefficient, Selected Specifications

Specification	Interval
OLS (2), Percentile Bounds	[.009, .063]
OLS (1), Percentile Bounds	[.008, .063]
Endogenous *T**, IV (2), Percentile Bounds	[.014, .219]
Endogenous *T**, IV (1), Percentile Bounds	[.002, .244]
OLS (2), HAS Bounds	[.009, .055]
OLS (1), HAS Bounds	[.008, .056]
Endogenous *T**, IV (2), HAS Bounds	[.009, .217]
Endogenous *T**, IV (1), HAS Bounds	[-.004, .240]

Appendix

In the text, we showed that $\hat{A}^0(\hat{\theta}_q) \xrightarrow{p} A^0(\theta_q^*)$. In this appendix, we show that convergence occurs at rate $\sqrt{nq}$. As the analysis below makes clear, the argument is complicated by a small sample effect arising from the fact that $\hat{\theta}_q$ depends on the realized values of T_i over the entire sample. After demonstrating that $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\theta_q^*))$ converges to a normal distribution, we go on to establish the useful result that $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q)) \xrightarrow{d} N(0, A^0(\hat{\theta}_q)(1 - A^0(\hat{\theta}_q))$.

In the following, let $h(X)$ denote the density of X and let $p_T(X)$ denote the conditional probability $Pr(T=1|X)$. Letting $J_\delta(c) \equiv Pr(G(X;\delta) \leq c)$ denote the cumulative distribution function of $G(X;\delta)$, we assume that for all δ in some neighborhood of δ*, (i) $J_\delta(c)$ is continuous and differentiable in δ and (ii) $J_\delta^{'}(c) \equiv j_\delta(c)$ is continuous and positive everywhere that $0 < J_\delta(c) < 1$. Let $\kappa_q(\delta) \equiv J_\delta^{-1}(q)$ denote the q-quantile for $J_\delta(c)$ and let $\hat{\kappa}_q(\delta) \equiv \min(c \mid \hat{J}_\delta(c) \geq q))$, where $\hat{J}_\delta$ is the empirical cdf for $G(X;\delta)$ for a given δ.

As a preliminary, we establish the asymptotic normality of $\hat{\kappa}_q$ and thus $\hat{\theta}$:

Lemma 1: $\sqrt{n}(\hat{\kappa}_q - \kappa_q*) \xrightarrow{d} N(0, V_{\hat{\kappa}})$ *, where* $V_{\hat{\kappa}} = V_1 + V_2 + V_3$,

$$V_1 = \frac{\partial \kappa_q(\delta*)}{\partial \delta}'(E(\frac{\partial^2 L(\delta*)}{\partial\delta\partial\delta'})^{-1} E(\frac{\partial L(\delta*)}{\partial \delta}' \frac{\partial L(\delta*)}{\partial \delta}) E(\frac{\partial^2 L(\delta*)}{\partial\delta\partial\delta'})^{-1}) \frac{\partial \kappa_q(\delta*)}{\partial \delta},\ V_2 = \frac{q(1-q)}{j_{\delta*}(\kappa_q*)^2},$$

and $V_3 = 2Cov\,(\frac{\partial \kappa_q(\delta^*)}{\partial \delta}' \; E(\frac{\partial^2 L}{\partial\delta\partial\delta'})^{-1} \frac{\partial L}{\partial \delta}', \frac{\psi(X\delta^* - \kappa_q{}^*)}{j_{\delta^*}(\kappa_q{}^*)})$, *and where L is the log-likelihood and* $\psi(Y) = \begin{cases} q & if\ Y > 0 \\ -(1-q) & if\ Y \le 0 \end{cases}$.

$$\sqrt{n}(\hat{\theta}_q - \theta_q^*) \xrightarrow{d} N(0, \begin{bmatrix} V_{\hat{\delta}} & C_{\hat{\delta},\hat{\kappa}} \\ C_{\hat{\delta},\hat{\kappa}} & V_{\hat{\kappa}} \end{bmatrix}), \textit{ where}$$

$$V_{\hat{\delta}} = E(\frac{\partial^2 L(\delta^*)}{\partial\delta\partial\delta'})^{-1} E(\frac{\partial L(\delta^*)}{\partial \delta}' \frac{\partial L(\delta^*)}{\partial \delta}) E(\frac{\partial^2 L(\delta^*)}{\partial\delta\partial\delta'})^{-1} \textit{ and}$$

$$C_{\hat{\delta},\hat{\kappa}} = \frac{\partial \kappa_q(\delta^*)}{\partial \delta}' V_{\hat{\delta}} + Cov(E(\frac{\partial^2 L(\delta^*)}{\partial\delta\partial\delta'})^{-1}(\frac{\partial L(\delta^*)}{\partial \delta}'), \frac{\psi(X\delta^* - \kappa_q{}^*)}{j_{\delta^*}(\kappa_q{}^*)}).$$

Proof: We can decompose $\sqrt{n}(\hat{\kappa}_q - \kappa_q{}^*)$ as:

(A1) $\sqrt{n}(\hat{\kappa}_q - \kappa_q{}^*) = \sqrt{n}(\hat{\kappa}_q(\hat{\delta}) - \hat{\kappa}_q(\delta^*)) + \sqrt{n}(\hat{\kappa}_q(\delta^*) - \kappa_q{}^*).$

Koenker and Bassett (1982) show that under our continuity assumptions,

(A2) $\sqrt{n}(\hat{\kappa}_q(\delta^*) - \kappa_q(\delta^*)) = \frac{1}{j_{\delta^*}(\kappa_q(\delta^*))}\sqrt{n}\frac{\sum \psi(G(X;\delta^*) - \kappa_q(\delta^*))}{n} + o_p(1).$

To determine the distribution of the first term of (A1), we now show that $\sqrt{n}(\hat{\kappa}_q(\hat{\delta}) - \hat{\kappa}_q(\delta^*))$ has the same limiting distribution as $\sqrt{n}(\kappa_q(\hat{\delta}) - \kappa_q(\delta^*))$. From (A2) and its analogue for $\hat{\delta}$, we have:

(A3) $$\sqrt{n}(\hat{\kappa}_q(\hat{\delta}) - \kappa_q(\hat{\delta})) - \sqrt{n}(\hat{\kappa}_q(\delta^*) - \kappa_q(\delta^*)) = \frac{1}{j_{\hat{\delta}}(\kappa_q(\hat{\delta}))}\sqrt{n}\frac{\sum \psi(G(X;\hat{\delta}) - \kappa_q(\hat{\delta}))}{n} - \frac{1}{j_{\delta^*}(\kappa_q(\delta^*))}\sqrt{n}\frac{\sum \psi(G(X;\delta^*) - \kappa_q(\delta^*))}{n} + o_p(1).$$

Since $j_{\hat{\delta}}(\kappa_q(\hat{\delta})) \xrightarrow{p} j_{\delta*}(\kappa_q(\delta*))$, it follows from(A3) that

$\sqrt{n}(\hat{\kappa}_q(\hat{\delta}) - \kappa_q(\hat{\delta})) - \sqrt{n}(\hat{\kappa}_q(\delta*) - \kappa_q(\delta*))$ has the same limiting distribution as

$$\frac{1}{j_{\delta*}(\kappa_q*)}\sqrt{n}\left(\frac{\sum \psi(G(X;\hat{\delta}) - \kappa_q(\hat{\delta}))}{n} - \frac{\sum \psi(G(X;\delta*) - \kappa_q*)}{n}\right) .$$

Let $s_1(\hat{\delta}) \equiv I^0(\hat{\theta}_q) \cap ((\theta_q^*))^c$, $s_2(\hat{\delta}) \equiv (\hat{\theta}_q)^c \cap I_q^0(\theta*)$, $p_1(\hat{\delta}) \equiv \Pr(X \in s_1(\hat{\delta}))$, and

$p_2(\hat{\delta}) \equiv \Pr(X \in s_2(\hat{\delta}))$. Note that

$$\sqrt{n}\left(\frac{\sum \psi(G(X;\hat{\delta}) - \kappa_q(\hat{\delta}))}{n} - \frac{\sum \psi(G(X;\delta*) - \kappa_q*)}{n}\right) = \frac{\#(s_2(\hat{\delta})) - \#(s_1(\hat{\delta}))}{\sqrt{n}}$$
$$= \sqrt{n}\,\frac{\sum_i b_i(\hat{\delta})}{n},$$

where $b_i(\hat{\delta}) = \begin{cases} 1 & \text{if } X_i \in s_1(\hat{\delta}), \\ -1 & \text{if } X_i \in s_2(\hat{\delta}), \\ 0 & \text{otherwise.} \end{cases}$

By the central limit theorem, $\sqrt{n}\left(\frac{\sum_i b_i(\hat{\delta})}{n} - E(b_i)\right)$ converges in distribution to

$N(0, Var(b_i))$. Note that $E(b_i) = p_1(\hat{\delta}) - p_2(\hat{\delta})$ and

$Var(b_i) = p_1(\hat{\delta}) + p_2(\hat{\delta}) - (p_1(\hat{\delta}) - p_2(\hat{\delta}))^2$.

Since $p_1(\hat{\delta}) \xrightarrow{p} 0$, $p_2(\hat{\delta}) \xrightarrow{p} 0$, we can conclude that $\sqrt{n}\,\frac{\sum_i b_i(\hat{\delta})}{n} \xrightarrow{p} 0$. It follows that $\sqrt{n}(\hat{\kappa}_q(\hat{\delta}) - \kappa_q(\hat{\delta})) - \sqrt{n}(\hat{\kappa}_q(\delta*) - \kappa_q(\delta*)) \xrightarrow{p} 0$, which implies that

$\sqrt{n}(\hat{\kappa}_q(\hat{\delta}) - \hat{\kappa}_q(\delta*))$ has the same limiting distribution as $\sqrt{n}(\kappa_q(\hat{\delta}) - \kappa_q(\delta*))$.

A standard delta-method argument establishes that $\sqrt{n}(\kappa_q(\hat{\delta})-\kappa_q*)$ has the same limiting distribution as $\frac{\partial\kappa_q(\delta*)}{\partial\delta}\sqrt{n}(\hat{\delta}-\delta*)$, which in turn, as shown by White (1982), has the same limiting distribution as $\frac{\partial\kappa_q(\delta*)}{\partial\delta}(-E(\frac{\partial^2 L(\delta*)}{\partial\delta\partial\delta'})^{-1})\sqrt{n}(\sum\frac{\partial L(\delta*)}{\partial\delta'}\Big/n)$, where L is the log-likelihood. A multivariate central limit theorem applies to $\sqrt{n}\begin{bmatrix}\hat{\kappa}_q(\hat{\delta})-\hat{\kappa}_q(\delta*)\\ \hat{\kappa}_q(\delta*)-\kappa_q*\end{bmatrix}$, implying that the sum $\sqrt{n}(\hat{\kappa}_q-\kappa_q*)=$ $\sqrt{n}(\hat{\kappa}_q(\hat{\delta})-\hat{\kappa}_q(\delta*))+\sqrt{n}(\hat{\kappa}_q(\delta*)-\kappa_q*)$ is asymptotically normal with the specified variance.

Similarly, $\sqrt{n}(\hat{\theta}-\theta*)$ is asymptotically normal, with variance $V_{\hat{\theta}}=\begin{bmatrix}V_{\hat{\delta}} & C_{\hat{\delta},\hat{\kappa}}\\ C_{\hat{\delta},\hat{\kappa}} & V_{\hat{\kappa}}\end{bmatrix}$.

Q.E.D.

Lemma 2 below is key to the argument demonstrating the convergence of $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q)-A^0(\theta_q^*))$:

Lemma 2: *Let* $\hat{A}^0(\theta_q^*)\equiv\frac{\sum_{i\in I^0(\theta_q^*)}T_i}{\#(I^0(\theta_q^*))}$. *Then* $\sqrt{n}(\hat{A}^0(\hat{\theta}_q)-\hat{A}^0(\theta_q^*))\xrightarrow{d}N(0,\Delta' V_{\hat{\theta}_q}\Delta)$,

where $\Delta\equiv\partial A^0(\theta_q^*)/\partial\theta_q$ denotes the gradient vector of A^0 at θ_q^* .

Proof: Let $S_1(\hat{\theta}_q)\equiv I^0(\hat{\theta}_q)\cap I^0(\theta_q^*)^c$, $S_2(\hat{\theta}_q)\equiv I^0(\theta_q^*)\cap I^0(\hat{\theta}_q)^c$, $P_1(\hat{\theta}_q)\equiv\Pr(X\in S_1(\hat{\theta}_q))$, and $P_2(\hat{\theta}_q)=\Pr(X\in S_2(\hat{\theta}_q))$. One can write

$$\text{(A4)}\qquad \sqrt{n}(\hat{A}^0(\hat{\theta}_q)-\hat{A}^0(\theta_q^*))=\sqrt{n}(\frac{\sum_{i\in I^0(\hat{\theta}_q)}T_i}{\#(I^0(\hat{\theta}_q))}-\frac{\sum_{i\in I^0(\theta_q^*)}T_i}{\#(I^0(\theta_q^*))})$$

$$= \sqrt{n}\left(\frac{\sum_{i \in I^0(\hat{\theta}_q)} T_i - \sum_{i \in I^0(\theta_q^*)} T_i}{\#(I^0(\theta_q^*))} + \frac{\sum_{i \in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\theta_q^*))} \frac{\#(I^0(\theta_q^*)) - \#(I^0(\hat{\theta}_q))}{\#(I^0(\hat{\theta}_q))}\right)$$

$$= \sqrt{n}\left(\frac{\sum_{i \in S_1} T_i - \sum_{i \in S_2} T_i}{\#(I^0(\theta_q^*))} + \frac{\sum_{i \in I_q^0(\hat{\theta}_q)} T_i}{\#(I^0(\theta_q^*))} \frac{\#(S_2(\hat{\theta}_q)) - \#(S_1(\hat{\theta}_q))}{\#(I^0(\hat{\theta}_q))}\right)$$

$$= \sqrt{n} \frac{n}{\#(I^0(\theta_q^*))}\left(\frac{\sum_i T_i B_i(\hat{\theta}_q)}{n} - \frac{\sum_{i \in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))} \frac{\sum_i B_i(\hat{\theta}_q)}{n}\right)$$

where

$$B_i(\hat{\theta}_q) = \begin{cases} 1 & if \;\; X_i \in S_1(\hat{\theta}_q) \\ -1 & if \;\; X_i \in S_2(\hat{\theta}_q) \\ 0 & otherwise \end{cases} \quad .$$

In determining the asymptotic distribution of

$\sqrt{n} \frac{n}{\#(I^0(\theta_q^*))}\left(\frac{\sum_i T_i B_i(\hat{\theta}_q)}{n} - \frac{\sum_{i \in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))} \frac{\sum_i B_i(\hat{\theta}_q)}{n}\right)$, one must take into account the small sample effect arising from the fact that $\hat{\theta}_q$ depends on the realized values of T_{i} over the entire sample. Thus, let $p_{T1}(\hat{\theta}_q) \equiv \Pr(T = 1 | X \in S_1(\hat{\theta}_q))$, and let $\tilde{p}_{T_1}(\hat{\theta}_q) \equiv \Pr(T = 1 | X \in S_1(\hat{\theta}_q), \hat{\theta}_q) \equiv \int_{X \in S_1(\hat{\theta}_q)} p_T(X) h(X) dX / P_1(\hat{\theta}_q)$ denote the probability that T = 1 when $X \in S_1(\hat{\theta}_q)$ and $\hat{\theta}_q$ is taken to be exogenous. For convenience, let $a_1(\hat{\theta}_q) \equiv p_{T_1}(\hat{\theta}_q) - \tilde{p}_{T_1}(\hat{\theta}_q)$. Similarly, let $p_{T_2}(\hat{\theta}_q) \equiv \Pr(T = 1 | X \in S_2(\hat{\theta}_q))$ and write $p_{T_2}(\hat{\theta}_q) \equiv \tilde{p}_{T_2}(\hat{\theta}_q) + a_2(\hat{\theta}_q)$, where $\tilde{p}_{T_2}(\hat{\theta}_q) \equiv \Pr(T=1 \mid X \in S_2(\hat{\theta}_q), \hat{\theta}_q) \equiv \int_{X \in S_2(\hat{\theta}_q)} p_T(X) h(X) dX / P_2(\hat{\theta}_q)$ and $a_2(\hat{\theta}_q)$ is a small sample effect. Since $a_1(\hat{\theta}_q)$ and $a_2(\hat{\theta}_q)$ approach 0 as $\hat{\theta}_q$ approaches θ_q^* and $\hat{\theta}_q \xrightarrow{p} \theta_q^*$, $a_1(\hat{\theta}_q) \xrightarrow{p} 0$ and $a_2(\hat{\theta}_q) \xrightarrow{p} 0$.

Note that

$$(A5) \quad \sqrt{n}\frac{n}{\#(I^0(\theta_q^*))}\frac{\sum_i T_i B_i(\hat{\theta}_q)}{n} = \frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}\Big(\frac{\sum_i T_i B_i(\hat{\theta}_q)}{n} - (p_{T1}(\hat{\theta}_q)P_1(\hat{\theta}_q) - p_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q))\Big)$$

$$+\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}(a_1(\hat{\theta}_q)P_1(\hat{\theta}_q) - a_2(\hat{\theta}_q)P_2(\hat{\theta}_q))$$
$$+\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}(\tilde{p}_{T_1}(\hat{\theta}_q)P_1(\hat{\theta}_q) - \tilde{p}_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q)).$$

Substituting (A5) into (A4) and adding and subtracting

$\sqrt{n}\frac{n}{\#(I^0(\theta_q^*))}\frac{\sum_{i \in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))}(P_2(\hat{\theta}_q) - P_1(\hat{\theta}_q))$, one obtains

$$(A6) \quad \sqrt{n}(\hat{A}^0(\hat{\theta}_q) - \hat{A}^0(\theta_q^*)) = \frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}\Big(\frac{\sum_i T_i B_i(\hat{\theta}_q)}{n} - (p_{T1}(\hat{\theta}_q)P_1(\hat{\theta}_q) - p_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q))\Big))$$

$$+\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}(a_1(\hat{\theta}_q)P_1(\hat{\theta}_q) - a_2(\hat{\theta}_q)P_2(\hat{\theta}_q))$$
$$+\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}(\tilde{p}_{T_1}(\hat{\theta}_q)P_1(\hat{\theta}_q) - \tilde{p}_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q)).$$

$$-\sqrt{n}\frac{n}{\#(I^0(\theta_q^*))}\frac{\sum_{i \in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))}\Big(\frac{\sum_i B_i(\hat{\theta}_q)}{n} - (P_1(\hat{\theta}_q) - P_2(\hat{\theta}_q))\Big) + (P_1(\hat{\theta}_q) - P_2(\hat{\theta}_q)).$$

Letting $\hat{q} \equiv \Pr(G(X;\hat{\delta}) \le \hat{\kappa}_q) \equiv \int_{X:G(X;\hat{\delta}) \le \hat{\kappa}_q} h(X) dX$, one can write

$$\text{(A7)} \qquad A^0(\hat{\theta}_q) - A^0(\theta_q^*) = \frac{\int_{X:G(X;\hat{\delta})\le\hat{\kappa}_q} p_T(X)h(X)dX}{\hat{q}} - \frac{\int_{X:G(X;\delta^*)\le\kappa_q^*} p_T(X)h(X)dX}{q}$$

$$= \frac{q\left(\int_{X\in S_1(\hat{\theta}_q)} p_T(X)h(X)dX - \int_{X\in S_2(\hat{\theta}_q)} p_T(X)h(X)dX\right)}{q\hat{q}}$$

$$+ \frac{(q-\hat{q})\int_{X:G(X;\delta)^*\le\kappa_q^*} p_T(X)h(X)dX}{\hat{q}q}$$

$$= \frac{\tilde{p}_{T_1}(\hat{\theta}_q)P_1(\hat{\theta}_q) - \tilde{p}_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q) + (q-\hat{q})A^0(\theta_q^*)}{\hat{q}},$$

which implies that

$$(A8) \qquad \tilde{p}_{T_1}(\hat{\theta}_q)P_1(\hat{\theta}_q) - \tilde{p}_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q) = \hat{q}(A^0(\hat{\theta}_q) - A^0(\theta_q^*) + (P_2(\hat{\theta}_q) - P_1(\hat{\theta}_q))A^0(\theta_q^*)$$

.

Substituting (A8) into (A6) and rearranging yields:

$$\text{(A9)} \qquad \sqrt{n}(\hat{A}^0(\hat{\theta}_q) - \hat{A}^0(\theta_q^*)) = \frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}(\hat{q}(A^0(\hat{\theta}_q) - A^0(\theta_q^*))) + X(n,\hat{\theta}_q)$$

where

$$\text{(A10)} \qquad X(n,\hat{\theta}_q) = \frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}\left(\frac{\sum_i T_i B_i(\hat{\theta}_q)}{n} - (p_{T1}(\hat{\theta}_q)P_1(\hat{\theta}_q) - p_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q))\right)$$

$$+ \frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}(a_1(\hat{\theta}_q)P_1(\hat{\theta}_q) - a_2(\hat{\theta}_q)P_2(\hat{\theta}_q))$$

$$-\sqrt{n}\frac{n}{\#(I^0(\theta_q^*))}\frac{\sum_{i\in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))}\left(\frac{\sum_i B_i(\hat{\theta}_q)}{n} - (P_1(\hat{\theta}_q) - P_2(\hat{\theta}_q))\right)$$

$$-\sqrt{n}\frac{n}{\#(I^0(\theta_q^*))}(\frac{\sum_{i\in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))}-A_q^0(\theta_q^*))(P_1(\hat{\theta}_q)-P_2(\hat{\theta}_q)).$$

As n becomes large, $X(n,\hat{\theta})$ converges in probability to zero. To see this, note that by the central limit theorem, $\sqrt{n}(\frac{\sum_i T_i B_i(\hat{\theta}_q)}{n}-(p_{T1}(\hat{\theta}_q)P_1(\hat{\theta}_q)-p_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q)))$ converges in distribution to $N(0,\chi(P_1(\hat{\theta}_q),P_2(\hat{\theta}_q)))$, where

$$\chi(P_1(\hat{\theta}_q),P_2(\hat{\theta}_q))=P_1(\hat{\theta}_q)(1-p_{T1}(\hat{\theta}_q)P_1(\hat{\theta}_q)+p_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q))^2+P_2(\hat{\theta}_q)(-1-p_{T1}(\hat{\theta}_q)P_1(\hat{\theta}_q)+p_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q))^2.$$

Since $P_1(\hat{\theta}_q)\xrightarrow{p}0$, $P_2(\hat{\theta}_q)\xrightarrow{p}0$, and $\frac{\#(I^0(\theta_q^*))}{n}\xrightarrow{p}q$,

$\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}(\frac{\sum_i T_i B_i}{n}-(p_{T1}(\hat{\theta}_q)P_1(\hat{\theta}_q)-p_{T2}(\hat{\theta}_q)P_2(\hat{\theta}_q)))\xrightarrow{p}0$. Similarly, the third term on the right hand side of (A7), $\sqrt{n}\frac{n}{\#(I^0(\theta_q^*))}\frac{\sum_{i\in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))}(\frac{\sum_i B_i(\hat{\theta}_q)}{n}-(P_2(\hat{\theta}_q)-P_1(\hat{\theta}_q)))$, converges in probability to zero. Also, note that $\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}P_1(\hat{\theta}_q)$ and $\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}P_2(\hat{\theta}_q)$ converge asymptotically to stable (normal) distributions. Since $a_1(\hat{\theta}_q)$ and $a_2(\hat{\theta}_q)$ both converge in probability to 0,

$$\frac{n}{\#(I^0(\theta_q^*))}\frac{\sum_{i\in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))}\sqrt{n}(a_1(\hat{\theta}_q)P_1(\hat{\theta}_q)-a_2(\hat{\theta}_q)P_2(\hat{\theta}_q))\xrightarrow{p}0.$$

Finally, consider the fourth term on the right hand side of (A10). Since

$P_1(\theta_q^*) = P_2(\theta_q^*) = 0$ and $\sqrt{n}(\hat{\theta}_q - \theta_q^*)$ is asymptotically normal,

$$\sqrt{n}(P_1(\hat{\theta}_q) - P_2(\hat{\theta}_q)) \xrightarrow{d} N(0, (\partial P_1(\theta_q^*)/\partial\theta) - (\partial P_2(\theta_q^*)/\partial\theta))V_{\hat{\theta}_q}(\partial P_1(\theta_q^*)/\partial\theta) - (\partial P_2(\theta_q^*)/\partial\theta)).$$

Since $\dfrac{\sum_{i\in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))} \xrightarrow{p} A^0(\theta_q^*)$ and $\dfrac{\#(I^0(\theta_q^*))}{n} \xrightarrow{p} q$, it follows that

$$\frac{n}{\#(I^0(\theta_q^*))}\sqrt{n}\left(\frac{\sum_{i\in I^0(\hat{\theta}_q)} T_i}{\#(I^0(\hat{\theta}_q))} - A_q^0(\theta_q^*)\right)(P_1(\hat{\theta}_q) - P_2(\hat{\theta}_q))) \xrightarrow{p} 0 .$$

We can conclude that

(A11) $\sqrt{n}(\hat{A}^0(\hat{\theta}_q) - \hat{A}^0(\theta_q^*))) \xrightarrow{d} \dfrac{n}{\#(I^0(\theta_q^*))} q\sqrt{n}((A^0(\hat{\theta}_q) - A^0(\theta_q^*))$, which

implies that

$$\sqrt{n}(\hat{A}^0(\hat{\theta}_q) - \hat{A}^0(\theta_q^*)) \xrightarrow{d} N(0, \Delta' V_{\hat{\theta}_q}\Delta) \qquad \text{Q.E.D.}$$

Note that in order to estimate the variance of $\hat{A}^0(\hat{\theta}_q) - \hat{A}^0(\theta_q^*)$ from the formula in the lemma, one would need to estimate the derivative vectors Δ, $\dfrac{\partial \kappa_q(\delta^*)}{\partial \delta}$, and the density $j_{\delta^*}(\kappa_q{}^*)$.

Using the above lemma, it is easy to demonstrate that $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\theta_q^*))$ converges to a normal distribution with mean zero.

Theorem 1: $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\theta_q^*))$ *converges asymptotically to a normal distribution with zero mean.*

Proof:

Decompose $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\theta_q^*))$ as follows:

(A12) $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\theta_q^*)) = \sqrt{nq}(\hat{A}^0(\theta_q^*) - A^0(\theta_q^*)) + \sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - \hat{A}^0(\theta_q^*))$.

Note that $\sqrt{nq}(\hat{A}^0(\theta_q^*) - A^0(\theta_q^*)) \overset{LD}{=} \sqrt{nq}(\frac{\sum_{i\in I^0(\theta_q^*)} T_i}{nq} - A^0(\theta_q^*))$. From the preceding lemmas, we know that $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - \hat{A}^0(\theta_q^*))$ has the same asymptotic distribution as a linear combination of $\sqrt{nq}(\sum \frac{\partial L}{\partial \delta^{:}} / n)$ and $\sqrt{nq}\frac{\sum \psi(G(X;\delta^*) - \kappa_q(\delta^*))}{n}$, both of which are mean zero. By the multivariate central limit theorem, the components of (A12) are multivariate normal. Thus, the sum $\sqrt{nq}(\hat{A}^0(\theta_q^*) - A^0(\theta_q^*)) + \sqrt{nq}(\hat{A}^0(\hat{\theta}) - \hat{A}^0(\theta_q^*))$ is asymptotically normal with mean zero. Q.E.D.

Corollary 1: $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q)) \xrightarrow{d} N(0, A^0(\theta_q^*)(1 - A^0(\theta_q^*))$.

Proof:

Instead of (A12), one can decompose $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q))$ as

(A13) $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\theta_q^*)) = \sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q)) + \sqrt{nq}(A^0(\hat{\theta}_q) - A^0(\theta_q^*))$

Substituting (A9) and (A12) into (A13) and rearranging terms yields

(A14) $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q)) = \sqrt{nq}(\hat{A}^0(\theta_q^*) - A^0(\theta_q^*)) + \sqrt{q}X(n, \hat{\theta}_q)$

We have already shown than $X(n,\hat{\theta}) \xrightarrow{p} 0$ and $\sqrt{nq}(\hat{A}^0(\theta_q^*) - A^0(\theta_q^*)) \xrightarrow{d} N(0, A^0(\theta_q^*)(1 - A^0(\theta_q^*)))$. Q.E.D.

Corollary 1 and the fact that $A^0(\hat{\theta}_q) \xrightarrow{p} A^0(\theta_q^*)$ immediately give us

Corollary 2: $\sqrt{nq}(\hat{A}^0(\hat{\theta}_q) - A^0(\hat{\theta}_q)) \xrightarrow{d} N(0, A^0(\hat{\theta}_q)(1 - A^0(\hat{\theta}_q))$.