

(preliminary and incomplete)

Diagnosis Measurement Error and Corrected Instrumental Variables

Donna Chen, Brent Kreider, Elizabeth Merwin, and Steven Stern*

University of Virginia

August 2000

Abstract

Health diagnosis indicators used as explanatory variables in econometric models often suffer from substantial measurement error. This measurement error can lead to seriously biased inferences about the effects of health conditions on the outcome measure of interest, and the bias generally spills over into inferences about the effects of policy/treatment variables. We generalize an existing instrumental variables (IV) method to make it compatible with the types of instruments typically available in large datasets containing health diagnoses. In particular, we relax the classical IV assumption that the instruments must have uncorrelated measurement errors. We identify and estimate the covariance matrix of the measurement errors and then use this information to derive a correction term to mitigate or eliminate the bias associated with classical IV. Our Monte Carlo simulations suggest that this corrected IV method can produce estimates far superior to those produced by OLS or classical IV.

1 Introduction

Diagnostic categories are frequently used as explanatory variables in econometric models, where the outcome measure of interest is predicted for various diagnostic groups, policies/treatments, and other personal characteristics. However, as is well understood among clinical researchers, diagnosis indicators often suffer from substantial measurement error.¹ This is particularly true for mental health diagnoses. Reliance on inaccurate diagnosis indicators can substantially hinder

*Assistant professor of psychiatry, assistant professor of economics, associate professor of nursing, and professor of economics, respectively.

¹For the case of research instruments, there are many sources of measurement error including incorrectly answered questions, subjective decision-making or interpretation on the

researchers' and clinicians' ability to reliably evaluate treatment approaches, measure clinical progress, or estimate relationships between patient attributes, policy instruments, and various outcome measures of interest (such as functional status, average hospital length of stay, or success in work-fare). Assessing the generalizability of treatment interventions across patient types, for example, is problematic given limitations in the accuracy of diagnosis information. In the economics literature, it is commonplace to omit relevant mental and physical health variables in behavioral models, in part due to concern that such indicators are inaccurately measured.²

While "gold standard" references in the psychometric literature reflected in "reliability," "validity," and kappa scores provide useful benchmarks for making comparisons across diagnostic instruments (e.g., Fennig, et al. 1994; McGorry, et al. 1995; Roy, et al. 1997; Clarke, Smith, and Hermann 1998), even instruments meeting these gold standards are likely contaminated with sizable errors. The reliability and validity outcomes that appear in the literature point to substantial measurement error. Even if these measures showed greater agreement, they would still indicate measurement error, and most patients are not diagnosed with gold standard instruments. At any rate, these measures of agreement provide little assistance to mental health services researchers attempting to estimate the effects of health conditions and other variables on outcomes. Much effort has been put into developing better diagnosis instruments and into encouraging the use of standardized codes, but the nature of the measurement error problem is such that it will always exist as long as researchers must rely on clinical evaluation or research instruments.

The standard approach for handling measurement error in statistical models involves instrumental variables (IV) estimation.³ Suppose we wish to estimate the effects of health conditions on medical expenditures or average length of stay. In this context, a researcher might instrument one badly measured diagnosis indicator (a vector of possible diagnoses) with another in attempts to obtain consistent estimates of the health effects. In most widely used secondary datasets, however, multiple indicators of a patient's diagnosis are unlikely to be independent of each other as required for the classical IV method to work. For example, two diagnosers may partially rely on the same written medical history when forming their conclusions.

part of the professional administering the instrument, incomplete surveys, and the suboptimal weighting of answers to construct diagnosis measures. Even if there were no errors at the diagnosis stage, measurement error would still be prevalent given the practical necessity to aggregate different types of diagnoses into a relatively small classes. The discretization of mental health conditions (e.g., into binary "yes/no" diagnoses indicators) which in reality vary along continuums represents an additional source of measurement error.

²In the retirement literature, for example, prominent researchers have called the appropriate use of poorly measured health variables "the major unsettled issue in the empirical literature on the labor supply of older workers" (Anderson and Burkhauser, 1984).

³IV estimation has recently been used in health services research to control for endogenous regressors (see, for example, Newhouse and McClellan, 1998), though it does not appear that IV has been used in that field to address measurement error. While there are mathematical similarities between the endogenous regressors problem and the mismeasured regressors problem, they have different causes, different effects, and instrument selection rules differ.

The primary goal of this paper is to generalize the existing instrumental variables method to allow for correlated measurement errors. Our method first identifies and measures the biases resulting from dependent measurement errors, then offers a method for constructing estimators purged of these biases. Constructing such an estimator requires construction of the “covariance matrix” of the measurement errors, a rich measure of the magnitudes and directions of measurement errors in the diagnosis data. Our approach allows for very general error structures.⁴

Black et al. (1999) extend the IV literature in a different direction by providing parameter estimate bounds for the case of nonzero correlation between the true value of an explanatory variable and its measurement error. They maintain the classical IV assumption that the errors are independent across the mismeasured indicators. Their extension is especially attractive for the case in which a true explanatory variable measured with error is binary instead of continuous; in that case, the true variable and its measurement error will be negatively correlated.

Based on the results of our Monte Carlo simulations, we expect that our methods will allow mental health services researchers to obtain substantially improved estimates of health and intervention effects in many different contexts (e.g., in understanding real-world effectiveness of treatments and other interventions; understanding relationships between health, mental health, level of functional disability, and patient decisions, etc.). Thus far, we have been able to show that the classic IV method, which has not yet been introduced into the mental health services field, typically yields substantially improved estimates compared with the ordinary least squares technique. We have also been able to show, both mathematically and empirically, that our new generalized IV method further improves these estimates – sometimes dramatically.

2 Methodology

2.1 Intuition

We start off with an overly simplified example to provide the reader with some basic intuition. Once the basic intuition is developed, we discuss more interesting and realistic cases.

Consider a simple equation where some outcome variable for person i , y_i , is affected by whether person i has a psychiatric illness. Such an outcome variable could be medical expenses in a given year. We assume, for now, that the outcome variable is affected only by psychiatric illness according to the linear equation,

$$y_i = \beta q_i + \eta_i, \quad i = 1, 2, \dots, N \tag{1}$$

⁴At first glance, Fuller (1987 p. 151-153) proposes an “IV” estimator that looks similar to ours in that there is a “correction to the denominator.” However, Fuller explicitly assumes that his instrument is independent of the measurement error (Thm 2.4.1), and the correction is due to the fact that his IV estimator is really a LIML estimator.

where q_i is a continuous measure of person i 's illness, $q_i \in [0, 1]$, β is the effect of psychiatric illness on y_i , and η_i is a random error term with zero mean, uncorrelated with q_i . Assume that, instead of being able to observe q_i in the data, the investigator observes only an imperfect measure of q_i , x_{1i} , with

$$\begin{aligned}\Pr [x_{1i} = 1 \mid q_i] &= q_i; \\ \Pr [x_{1i} = 0 \mid q_i] &= 1 - q_i.^5\end{aligned}\tag{2}$$

Then, we can write

$$x_{1i} = q_i + v_{1i}\tag{3}$$

with

$$\begin{aligned}\Pr [v_{1i} = 1 - q_i \mid q_i] &= q_i; \\ \Pr [v_{1i} = -q_i \mid q_i] &= 1 - q_i;\end{aligned}$$

$$\begin{aligned}E [v_{1i} \mid q_i] &= q_i (1 - q_i) - (1 - q_i) q_i = 0; \\ Var [v_{1i} \mid q_i] &= q_i (1 - q_i)^2 + (1 - q_i) q_i^2 \\ &= q_i (1 - q_i).\end{aligned}$$

Since q is unobserved, the investigator might estimates equation (1) as

$$\begin{aligned}y_i &= \beta q_i + \eta_i \\ &= \beta x_{1i} + \eta_i - \beta v_{1i} \\ &= \beta x_{1i} + e_i\end{aligned}$$

using data on x_1 . Note, however, that the error is correlated with the explanatory variable. The ordinary least squares (OLS) estimator

$$\hat{\beta}_{OLS} = \frac{N^{-1} \sum_{i=1}^N x_{1i} y_i}{N^{-1} \sum_{i=1}^N x_{1i}^2}$$

converges (as $N \rightarrow \infty$) to

$$\begin{aligned}\text{plim} \hat{\beta}_{OLS} &= \frac{\text{plim} \frac{1}{n} \sum_i x_{ji} y_i}{\text{plim} \frac{1}{n} \sum_i x_{ji}^2} \\ &= \frac{\text{plim} \frac{1}{n} \sum_i (q_i + v_{ji}) (\beta q_i + \eta_i)}{\text{plim} \frac{1}{n} \sum_i (q_i + v_{ji})^2} \\ &= \frac{\beta \sigma_q^2}{\sigma_q^2 + \int q (1 - q) f(q) dq}\end{aligned}$$

where $\sigma_q^2 = \text{plim} N^{-1} \sum_{i=1}^N q_i^2$. Since $E q_i v_{1i} = E q_i \eta_i = E v_{1i} \eta_i = 0$; i.e., errors are uncorrelated with true diagnosis and with each other,

$$\text{plim} \hat{\beta}_{OLS} = \beta \frac{\sigma_q^2}{\sigma_q^2 + \int q (1 - q) f(q) dq}.$$

Since

$$0 < \frac{\sigma_q^2}{\sigma_q^2 + \int q(1-q)f(q) dq} < 1,$$

the OLS estimator converges to a number biased towards zero: $|\text{plim } \widehat{\beta}_{OLS}| < |\beta|$. Thus, the investigator will tend to underestimate the effect of the health condition on the outcome of interest when relying on OLS. This is the standard problem when explanatory variables are measured with error. The directions of bias cannot easily be determined a priori when equation (1) is generalized to allow for more than one poorly measured explanatory variable. Nevertheless, in general, OLS estimates will be biased in the presence of measurement error in explanatory variables.

Now, assume that there is another imperfect measure of q_i , x_{2i} , with the same properties as x_{1i} described in equation (2) but independent of x_{1i} ; i.e., v_{2i} and v_{1i} are independent. Then, the investigator can use x_{2i} as an instrument for x_{1i} in an instrumental variables (IV) estimator of β :

$$\widehat{b}_{IV} = \frac{N^{-1} \sum_{i=1}^N x_{2i} y_i}{N^{-1} \sum_{i=1}^N x_{2i} x_{1i}}.$$

Note that \widehat{b}_{IV} converges to

$$\begin{aligned} \text{plim } \widehat{b}_{IV} &= \frac{\text{plim } N^{-1} \sum_{i=1}^N x_{2i} y_i}{\text{plim } N^{-1} \sum_{i=1}^N x_{2i} x_{1i}} & (4) \\ &= \frac{\text{plim } N^{-1} \sum_{i=1}^N (q_i + v_{2i})(\beta q_i + \eta_i)}{\text{plim } N^{-1} \sum_{i=1}^N (q_i + v_{2i})(q_i + v_{1i})} \\ &= \frac{\text{plim } N^{-1} \sum_{i=1}^N [\beta q_i^2 + q_i(\beta v_{2i} + \eta_i) + v_{2i} \eta_i]}{\text{plim } N^{-1} \sum_{i=1}^N [q_i^2 + q_i(v_{2i} + v_{1i}) + v_{2i} v_{1i}]}. \end{aligned}$$

As long as $E q_i v_{1i} = E q_i v_{2i} = E v_{2i} \eta_i = E v_{2i} v_{1i} = 0$; i.e., errors are uncorrelated with true diagnosis and with each other, equation (4) simplifies to

$$\text{plim } \widehat{b}_{IV} = \frac{\text{plim } N^{-1} \sum_{i=1}^N \beta q_i^2}{\text{plim } N^{-1} \sum_{i=1}^N q_i^2} = \beta.$$

Thus, the IV estimator converges to the true effect. This method of dealing with measurement error in explanatory variables is well known in econometrics (e.g., Greene, 1997, p. 440).

2.2 The Basic Problem with Independent Measurement Errors

It is now appropriate to add important characteristics of the problem specific to the application at hand. First of all, it may be the case that, instead of the

existence of a psychiatric illness being the explanatory variable of interest, it is the type of psychiatric illness that is of interest. We generalize equation (1) to

$$y_i = \sum_{k=1}^K \beta_k g(q_{ik}) + \sum_{k=1}^M z_{ik} \gamma_k + \eta_i \quad (5)$$

where $q_i = (q_{i1}, q_{i2}, \dots, q_{iK})'$ is a vector of size K measuring the severity/existence of different psychiatric conditions, $g(\bullet)$ is some specified function translating diagnoses into outcomes, and $z_i = (z_{i1}, z_{i2}, \dots, z_{iM})'$ is a vector of other exogenous covariates such as race, sex, and age to control for. We assume throughout that z_i is measured without error. Assume that x_{ijk} , $j = 1, 2$ is an independent discrete indicator of q_{ik} measured with error such that $\sum_k x_{ijk} = 1$ and

$$p_{ik}(q_i) \equiv \Pr[x_{ijk} = 1 \mid q_i]. \quad (6)$$

Constructing such a vector requires the investigator to commit to a method of aggregating psychiatric diagnoses into a relatively small class of K diagnoses. We show in Appendix A that we can generalize the specification of x_{ijk} to allow for multiple (comorbid) diagnoses. Although there are important issues associated with optimal aggregation that we begin to address in Appendix B, we assume them away here and condition on a chosen aggregation method. One possible specification for p is

$$p_{ik} = \frac{\exp\{q_{ik}\}}{\sum_l \exp\{q_{il}\}}. \quad (7)$$

As before, we can write

$$x_{ijk} = p_{ik} + v_{ijk}$$

with

$$\begin{aligned} \Pr[v_{ijk} = 1 - p_{ik} \mid q_{ik}] &= p_{ik}; \\ \Pr[v_{ijk} = -p_{ik} \mid q_{ik}] &= 1 - p_{ik}; \\ E[v_{ijk} \mid q_{ik}] &= 0; \\ \text{Cov}[v_{ijk}, v_{ijl} \mid q_i] &= \begin{cases} p_{ik}(1 - p_{ik}) & \text{for } k = l \\ -p_{ik}p_{il} & \text{for } k \neq l \end{cases}; \\ \text{Cov}[v_{ijk}, v_{ij'k'} \mid q_i] &= 0 \text{ for all } j \neq j'. \end{aligned} \quad (8)$$

Defining $x_{ij} = (x_{ij1}, x_{ij2}, \dots, x_{ijK})'$, $X'_{ij} = (x'_{ij}, z'_i)$, $\beta = (\beta_1, \beta_2, \dots, \beta_K)'$, $\gamma = (\gamma_1, \gamma_2, \dots, \gamma_M)$, and $\theta = (\beta', \gamma')'$,

$$\hat{\theta}_{IV} = \left[\frac{1}{n} \sum_i X_{i2} X'_{i1} \right]^{-1} \left[\frac{1}{n} \sum_i X_{i2} y_i \right] \quad (9)$$

with

$$\begin{aligned}
\text{plim} \widehat{\theta}_{IV} &= \left[\text{plim} \frac{1}{n} \sum_i X_{i2} X'_{i1} \right]^{-1} \left[\text{plim} \frac{1}{n} \sum_i X_{i2} y_i \right] \\
&= \left[\text{plim} \frac{1}{n} \sum_i \begin{pmatrix} (p_i + v_{i2}) (p_i + v_{i1})' & (p_i + v_{i2}) z'_i \\ z_i (p_i + v_{i1})' & z_i z'_i \end{pmatrix} \right]^{-1} \bullet \\
&\quad \left[\text{plim} \frac{1}{n} \sum_i \begin{pmatrix} (p_i + v_{i2}) [g(q_i) \beta + z_i \gamma + \eta_i] \\ z_i [g(q_i) \beta + z_i \gamma + \eta_i] \end{pmatrix} \right] \\
&= \left[\iint \begin{pmatrix} p(q) p'(q) & p(q) z' \\ z p'(q) & z z' \end{pmatrix} f(q, z) dq dz \right]^{-1} \bullet \\
&\quad \left[\iint \begin{pmatrix} p(q) g'(q) & p(q) z' \\ z g'(q) & z z' \end{pmatrix} f(q, z) dq dz \right] \begin{pmatrix} \beta \\ \gamma \end{pmatrix}.
\end{aligned}$$

If $g(q) = p(q)$, then $\text{plim} \widehat{\theta}_{IV} = \theta$. Otherwise, in general, there is a proportional bias equal to

$$\left[\iint \begin{pmatrix} p(q) p'(q) & p(q) z' \\ z p'(q) & z z' \end{pmatrix} f(q, z) dq dz \right]^{-1} \left[\iint \begin{pmatrix} p(q) g'(q) & p(q) z' \\ z g'(q) & z z' \end{pmatrix} f(q, z) dq dz \right]$$

which is the plim of a regression of $g(q)$ on $p(q)$ and z (CITE). In general, since $g(\bullet)$ cannot be identified without observing q , we assume that $g(q) = p(q)$ and recognize that results should be interpreted in light of the possibility that it is probably not so.

2.3 The Basic Problem with Correlated Measurement Errors

Now, consider the case where $\text{Cov}[v_{ijk}, v_{ij'k'} | q_i] \neq 0$ for all $j \neq j'$. For the specification of p in equation (7), for example, we could mechanically cause correlation among the measurement errors by adjusting the equation to

$$p_{ik} = \frac{\exp\{q_{ik} + u_{ik}\}}{\sum_l \exp\{q_{il} + u_{il}\}} \quad (10)$$

where $u_i = (u_{i1}, u_{i2}, \dots, u_{iK})'$ is a vector of errors constant across j . Also, consider the possibility that $\text{Cov}[v_{ijk}, z_i | q_i] \neq 0$ for all $j \neq j'$. For example, there is some concern (see, for example, [REF cite]) that blacks are overdiagnosed with schizophrenia and whites are underdiagnosed. Let

$$m(\theta) = \frac{1}{n} \sum_i X_{i2} (y_i - X'_{i1} \theta)$$

be the set of moment conditions for instrumental variables estimation. Then

$$\begin{aligned}
\text{plim } m(\theta) &= \text{plim} \frac{1}{n} \sum_i \begin{bmatrix} p(q_i) + v_{i2} \\ z_i \end{bmatrix} [p'(q_i)\beta + z'_i\gamma + \eta_i - p'(q_i)\beta - v'_{i1}\beta - z'_i\gamma]' \\
&= \text{plim} \frac{1}{n} \sum_i \begin{bmatrix} p(q_i) + v_{i2} \\ z_i \end{bmatrix} [-v'_{i1}\beta + \eta_i]' \\
&= \begin{pmatrix} -\text{plim} \frac{1}{n} \sum_i v_{i2}v'_{i1}\beta \\ -\text{plim} \frac{1}{n} \sum_i z_iv'_{i1}\beta \end{pmatrix} = \begin{pmatrix} -\Omega_{21} \\ -\Omega_{z1} \end{pmatrix} \beta
\end{aligned}$$

where

$$\begin{aligned}
\Omega_{21} &= \int E[v_{i2}v'_{i1} | q] f(q) dq; \\
\Omega_{z1} &= \int E[z_iv'_{i1} | q] f(q) dq.
\end{aligned}$$

If we can estimate Ω_{21} and Ω_{z1} consistently, then we can redefine our moment conditions as

$$\tilde{m}(\theta) = \frac{1}{n} \sum_i X_{i2}(y_i - X'_{i1}\theta) + \begin{pmatrix} \Omega_{21} \\ \Omega_{z1} \end{pmatrix} \beta, \quad (11)$$

and the value of θ that set $\tilde{m}(\theta) = 0$ will be a consistent estimate of θ . In particular,

$$\hat{\theta}_{IV} = \left[\frac{1}{n} \sum_i \begin{pmatrix} x_{i2}x'_{i1} & x_{i2}z'_i \\ z_ix'_{i1} & z_iz'_i \end{pmatrix} - \begin{pmatrix} \hat{\Omega}_{21} & 0 \\ \hat{\Omega}_{z1} & 0 \end{pmatrix} \right]^{-1} \left[\frac{1}{n} \sum_i \begin{pmatrix} x_{i2} \\ z_i \end{pmatrix} y_i \right]. \quad (12)$$

Note that, in general, the correction term in the inverse, $\begin{pmatrix} \hat{\Omega}_{21} & 0 \\ \hat{\Omega}_{z1} & 0 \end{pmatrix}$, decreases the inverse and therefore increases the magnitude of the IV estimator (counteracting the effect of correlated measurement error). Also note that if $\hat{\Omega}_{12} = \hat{\Omega}_{z1} = 0$, then the $\hat{\theta}_{IV}$ is equivalent to the IV estimator in equation (9) which is consistent when measurement errors are uncorrelated. Estimation error in the correction term can have serious effects on small sample bias and confidence interval sizes. We can mitigate these unfortunate effects by adjusting the correction by a proportionality factor a , where $a = 1$ corresponds to our generalized IV method just described and $a = 0$ corresponds to standard uncorrected IV. Using the mean squared error criterion, Appendix C shows that it is always optimal to introduce partial but not complete correction to the standard method: $0 < a < 1$.

2.4 Estimating the Measurement Error Covariance Matrices

The method discussed above requires having a consistent estimate of the covariance matrix of the diagnosis measurement errors, Ω . An important component

of our project is to collect information on the covariance of diagnosis errors across diagnosers so that we can estimate this matrix.⁶ Using the notation in Section 3.2, this covariance matrix is given by

$$\Omega = \begin{bmatrix} \text{cov}(v_{11}, v_{21}) & \text{cov}(v_{11}, v_{22}) & \cdots & \text{cov}(v_{11}, v_{2m}) \\ \text{cov}(v_{12}, v_{21}) & \text{cov}(v_{12}, v_{22}) & \cdots & \text{cov}(v_{12}, v_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(v_{1m}, v_{21}) & \text{cov}(v_{1m}, v_{22}) & \cdots & \text{cov}(v_{1m}, v_{2m}) \end{bmatrix}.$$

Once we estimate the terms in this matrix, we can use this information to undo the estimation bias associated with classical IV. Other researchers studying similar samples of patients could use our covariance estimates in their studies.

These error covariances across diagnoses cannot be directly observed in the data. Although we can easily estimate the covariances among the diagnoses themselves, this does not provide the requisite information about the error covariances since we cannot observe the true conditions (q). One approach to estimating the error covariances might be to survey a panel of medical experts about their impressions of these covariances. That approach is not very practical, however, because the relevant questions regarding error covariances would not be straightforward to interpret or answer. Instead, we present four conceptually feasible approaches for estimating the covariance matrix.

The first method is straightforward to implement but may not be financially viable for our particular application. We nevertheless discuss it to solidify ideas. For this approach, we would estimate the covariance matrix of the elements of the actual diagnoses x_{ij} :

$$V = \begin{pmatrix} \text{cov}(x_{11}, x_{21}) & \text{cov}(x_{11}, x_{22}) & \cdots & \text{cov}(x_{11}, x_{2m}) \\ \text{cov}(x_{12}, x_{21}) & \text{cov}(x_{12}, x_{22}) & \cdots & \text{cov}(x_{12}, x_{2m}) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(x_{1m}, x_{21}) & \text{cov}(x_{1m}, x_{22}) & \cdots & \text{cov}(x_{1m}, x_{2m}) \end{pmatrix} \quad (13)$$

using two separate data sources. The first source is our “estimation data” used to estimate the coefficients β in the outcome equation of interest. A “special data source” would be one in which patients have been independently diagnosed (independent in the sense that two or more diagnosers would not share information about the patient – at least information that might be subject to errors such as medical history). The ideal special data source would be one where we have N individuals each of whom is independently diagnosed by H “diagnosers.”⁷

An estimate of the covariance matrix Ω could then be obtained as the difference between the covariance matrix of diagnoses based on the estimation data

⁶Recall that there are many sources of diagnosis error even if diagnosers do not make mistakes in their assessments per se (e.g., the customary coding of health conditions in datasets as “yes/no” discards important information about the complete nature of the conditions).

⁷A diagnoser is a mental health professional (e.g., psychiatrist, psychologist, or social worker) who diagnoses individuals similar in characteristics to the diagnosers in the data used to estimate equation (5).

(in which diagnoses for a patient are not necessarily independent) and the covariance matrix of diagnoses based on the special data source. A consistent estimate of the covariance matrix for independent diagnoses when $H = 2$ is

$$\widehat{V} = \frac{1}{N} \sum_{i,h} \omega_i (x_{i2} - x_{\bullet 2})(x_{i1} - x_{\bullet 1})'$$

where ω_i is a weight given to observation i so that the weighted special data has the same distribution of observed explanatory variables as the estimation data and

$$x_{\bullet h} = \frac{1}{N} \sum_i \omega_i x_{ih}$$

is the average diagnosis in the special sample for diagnoser h .⁸ With a consistent estimate of V , we can compute the covariance matrix for equation (12) by

$$\widehat{\Omega} = \frac{1}{n} \sum_i \omega_i (x_{i2} - x_{\bullet 2})(x_{i1} - x_{\bullet 1})' - \widehat{V}. \quad (14)$$

Intuitively, there are three sources of variation in diagnoses: (a) variation due to variation in the true condition q_i , (b) variation due to measurement error independent across diagnosticians, and (c) measurement error common across diagnosticians. The covariance matrix of diagnoses using the estimation data includes variation due to variation of types (a), (b), and (c) while the covariance matrix of diagnoses using the special data source includes variation of types (a) and (b). Subtraction leaves only randomness due to variation of type (c). The problem with this approach is that we are currently unaware of any appropriate secondary dataset in which multiple diagnoses for patients can be considered independent.⁹ Thus, we would need to collect primary data. Based on our preliminary Monte Carlo experiments, it appears that we would need collect independent diagnoses on thousands of patients before obtaining reasonably precise estimates of Ω . The associated cost would likely be prohibitive.

A feasible way to estimate Ω using the above logic involves hiring a group of diagnosticians (psychiatrists/psychologists/psychiatric nurses/social workers) to participate in a thought experiment called the ‘‘Delphi method’’ (e.g., Kahan, et al. 1994).¹⁰ The participants will be asked to imagine a large, representative

⁸When $H > 2$, one has to decide how to use the extra diagnosticians to increase precision of the estimator of V .

⁹We would need a large dataset with information on diagnoses made in a manner similar to diagnoses in large secondary datasets. We have considered the possibility of using field studies for the refinements/revisions to the DSM-IV or research studies that use two independent raters. However, the nature of the measurement error in these research instruments is significantly different from large secondary datasets. Large data sets such as Medicare may also have a subset of clients admitted to two different facilities, diagnosed by two different practitioners within a short time frame. A possible contaminating factor would be patient self-report of prior diagnosis or exposure by the second practitioner to the first practitioner’s diagnosis through medical records. However, medical records from different facilities are generally not available for admission diagnosis.

¹⁰We recognize that the resulting covariance matrix of measurement errors among diagnosticians will be limited to the population type from which the data are obtained.

sample of patients each with a mental illness receiving treatment (specified as either outpatient care or inpatient care). They will further be asked to imagine that we repeatedly sample a random patient along with two random diagnosticians who are asked to independently diagnose that patient. In appropriately worded language, we will ask the participants to use a computer to fill in the following table indicating the joint probability of diagnosis responses of two randomly mental health workers (denoted MHW1 and MHW2).¹¹

		MHW2			
		Diagnosis1	Diagnosis2	...	Diagnosis13
MHW1	Diagnosis1				
	Diagnosis2				
	⋮				
	Diagnosis13				

The computer will assist the participants by providing feedback on the consistency of their responses with published statistics such as prevalence rates and kappa statistics. We can measure how much confidence to put in the “Delphi estimate” of Ω by measuring the variance of the estimate across the experts in early rounds of the process.

Alternatively, we could exploit information from panel data in which patients are repeatedly diagnosed over a relatively long period of time. Treating a patient’s most recent diagnosis (made on the basis of a long medical history) as the gold standard, we can directly estimate the covariance structure of the errors in earlier diagnoses and thus obtain an estimate of Ω for other samples with similar types of patients.

{alternative approach: lower bound estimate using research instrument - sketch in Appendix D}

3 Monte Carlo Experiments

3.1 First Example

Our first example uses simulated data from a contrived example that is easy to manipulate and examine. We consider a model of the form in equation (5). In particular, $g(q_{ik}) = p_k(q_i)$ and is defined as

$$p_k(q_i) = \frac{\exp\{q_{ik}\}}{\sum_l \exp\{q_{il}\}}$$

¹¹Based on our previous work with mental health data, we have constructed 13 categories into which psychiatric diagnoses can be appropriately aggregated. These categories are also used in our Monte Carlo simulations.

with

$$\begin{aligned} q_{ik} &= u_{ik} + 3v_{ik}, \\ u_{ik} &\sim U(0, 1), \\ v_{ik} &= \begin{cases} 1 & \text{with probability } 1/13 \\ 0 & \text{with probability } 12/13 \end{cases} \end{aligned}$$

The specification for q_{ik} ensures that q_{ik} is continuous and yet that one diagnosis dominates the others. The other explanatory variables have a uniform distribution:

$$z_{ik} \sim U(0, 1).$$

Measurement error occurs according to

$$\Pr[x_{ijk} = 1 \mid q_i] = \frac{p_k(q_i) + \sigma_m \vartheta_{ik}}{\sigma_m + \sum_l p_l(q_i)}.$$

Note that ϑ_{ik} does not vary with diagnosers j , and therefore, as long as $\sigma_m > 0$, there is positive correlation in diagnosers' diagnoses. We set $\sigma_m = 0.8$. When $p_k(q_i)$ is directly observed, there is no measurement error, OLS and IV should both produce consistent estimates of θ , and OLS should be efficient. When only x_{ijk} is observed but $\sigma_m = 0$, there is uncorrelated measurement error, OLS should provide inconsistent estimates, and IV should provide consistent estimates. When only x_{ijk} is observed and $\sigma_m > 0$, there is correlated measurement error, both OLS and IV should provide inconsistent estimates. But the asymptotic bias for IV should be smaller than for OLS, and the corrected IV estimates should be consistent. Table 1 provides results for a Monte Carlo experiment with 4 z -variables, 13 x -variables, an error with a standard deviation of 0.05, and a sample size of 10,000. There are 100 independent draws of the data for each Monte Carlo experiment.

In the Panel A of Table 1, we see that, for this example, OLS provides unbiased and very precise estimates of the parameters when there is no measurement error. However, once we add measurement error, OLS estimates of the x variable coefficients become significantly biased towards zero as seen in Panel B. The estimates of the z variable coefficients are not biased because the z -variables are uncorrelated with the x -variables. Panel C shows that classical IV estimates (without correction for correlation) are also significantly biased towards zero, but that the bias is much smaller than in the OLS estimates.

Panels D, E, and F report results for IV corrected for correlation. In all three, asymptotic biases are very small and statistically insignificant. The three panels vary, however, with respect to characteristics of the special sample used to estimate Ω . In Panel D, $N = 1000$ and $H = 3$ ¹² i.e., the sample is of 1000 patients each independently diagnosed by 3 diagnosers. Even though the

¹²The three diagnosers are used so that the first one is a realization of x_{i1} and the next two are averaged for a realization of x_{i2} . The appropriate adjustment is made due to the fact that x_{i2} is an average. No weighting is necessary because the two data sets have the same distribution.

estimates in Panel D exhibit no bias, 80% confidence intervals are much larger than they were in previous panels. In Panels E and F, we try different methods to increase the accuracy of $\widehat{\Omega}$ and therefore our IV estimators. In Panel E, we increase H from 3 to 5, while, in Panel F, we increase N from 1000 to 5000. Increasing H has no appreciable effect, while increasing N does. This suggests that an optimal design for a special sample used to estimate V in equation (13) should have a large N but does not require a large H .

3.2 Second Example

Our second example uses simulated data to examine how inferences about the effects of mismeasured mental health conditions depend on the estimation approach. Our outcome measure of interest is hospital length of stay. We simulated a large sample ($N = 10,000$) of patients such that their characteristics match well with the set of 6498 patients actually admitted to a Virginia state psychiatric hospital in 1980. Characteristics of the data are reported in Table 2.¹³ In particular, our simulated sample matches the characteristics of a subset of 4893 patients for whom information was available on length of stay, initial and final diagnosis, race, gender, age, and health care facility.

A simulated patient is assumed to be diagnosed with one of 13 primary mental health conditions¹⁴ according to equation (10). We constructed a model of true diagnosis where q_{ik} in equation (5) is modeled as

$$q_{ik} = \frac{\exp\{z_i\alpha_k + e_{ik}\}}{\sum_l \exp\{z_i\alpha_l + e_{il}\}} \quad (15)$$

with $e_{ik} \sim iidN(0, \sigma_e^2)$. The existence of the e_{ik} in equation (15) allows for variation in true medical conditions even after conditioning on observed covariates z_i . Equations (10) and (15) are used to simulate x_{ij} , $j = 1, 2$, where it is assumed that $u_{ik} \sim iidN(0, \sigma_u^2)$. If u_{ik} varied independently over diagnostors j , then OLS estimates would still be inconsistent (because there is measurement error), but classical IV estimators would be consistent (because the measurement error in x_{i1k} would be independent of the measurement error in x_{i2k}).

“True” values of $(\alpha, \beta, \gamma, \sigma_u, \sigma_e)$ are estimated by matching moments of the simulated data to moments in the Virginia state psychiatric hospital data and matching simulated \varkappa (kappa) statistics¹⁵ to \varkappa statistics found in the literature (e.g., Stravynski, Lamontagne, and Lavallee 1986; Riskind et al. 1987; Clark, et al. 1993; Fennig, et al. 1994; Hiller et al. 1994; McGorry et al. 1995; Kelly

¹³See Holt, Merwin and Stern (1999) for a full description of the Virginia data.

¹⁴The possible conditions include substance abuse, alcoholism, organic, schizophrenia, schizo-affective, paranoia, other psychological disorders, bipolar, depression, personality, adjustment, dementia, and other.

¹⁵The \varkappa statistic is a measure of agreement and is equal to

$$\frac{p_o - p_c}{1 - p_c}$$

where p_o is the proportion of observations where there is agreement and p_c is the proportion of observations where there would have been agreement by chance. See Cohen (1960).

and Mann 1996; Parker et al. 1997; Rosenman, Korten, and Levings 1997; Roy et al. 1997; Usten et al. 1997; Clarke, Smith, and Hermann 1998).¹⁶ Estimates of the “true” parameters are reported in Table 3. Once we have “true values” of $(\alpha, \beta, \gamma, \sigma_u, \sigma_e)$, we can simulate data samples from the “true distribution” of observations, estimate the model parameters for each draw of the data, and compute the distribution of our estimators.

The results of this Monte Carlo experiment are recorded in Table 4. Median estimation bias for each coefficient in the length of stay equation, along with their 80 percent confidence intervals, are provided for the OLS, classical IV, and corrected IV methods. The OLS coefficient estimates are severely biased: the magnitude of the median bias is typically the same size as the “true value” of the parameter, and the 80% confidence interval does not include the “truth” for any of the mental health conditions. One-third of the OLS coefficients on the health conditions are not even the correct sign. In the first row, for example, the “true” average length of stay for schizophrenia was 66 percent shorter than for dementia (the excluded health condition), while OLS estimates the average length of stay to be longer for schizophrenia.¹⁷ In the next row, “true” average length of stay for substance abuse is 254 percent shorter than for dementia patients, with OLS estimating only a 74 percent difference. Notice in Figure 1 that OLS provides precise estimates, but they are precise estimates of wrong values.

Classical IV estimation will also produce biased estimates if diagnosis errors for a patient are correlated across physicians, which, we have argued, will be the case in most administrative data sets. As we have discussed, however, classic IV will still generally lead to better estimates than OLS. This is seen very clearly in Table 4 in that median biases are uniformly a fraction of the size of the OLS median biases. Note that the 80% confidence intervals now include the constructed truth for all variables. Across health variables, the average median bias under classical IV is less than 1/3 the bias associated with OLS. Notice, for example, that classical IV coefficient indicates a 213 percent difference in average length of stay between substance abuse and dementia patients, which is much closer to the true value (254 percent) than the OLS estimate (74 percent). Similarly, classical IV produces the correct sign and a reasonable magnitude for the schizophrenia coefficient. Even the non-diagnosis coefficients, which are not measured with error, are much improved under IV with an average bias of only 1/4 that of OLS.¹⁸

Our corrected IV approach should produce consistent estimates, but they can have reduced precision because the covariance matrix of measurement error

¹⁶ σ_u and σ_e are identified by deviations between correlations across diagnoses in the Virginia data and what they would be if $\sigma_u = 0$ and matching κ statistics. We assume $\kappa = 0.7$ for all conditions.

¹⁷Technically, our estimate of $\partial E \ln(\text{spell length}) / \partial(\text{schizophrenia})$ is -0.66 . For numbers close to zero, this derivative is approximately the percentage difference in the expected spell length for someone with substance abuse compared with dementia.

¹⁸Recall that biases associated with mismeasured variables spill over into the coefficient estimates on any variables correlated with the mismeasured variables. Some non-diagnosis variables, like health care facility indicators, are not shown to save space.

dependencies Ω must be estimated. As noted above, the optimal IV correction method involves scaling $\hat{\Omega}$ by a proportionality factor to reduce the size of the correction and its effect on the width of the confidence interval (see Appendix C). We chose a factor of 0.05, which keeps the 80 percent confidence intervals similar in length to the uncorrected IV case.¹⁹ As shown in the last column of Table 4, the biases in our corrected IV estimates are uniformly a fraction of the classical IV biases. The average median bias across the diagnosis variables is less than 1/2 that in the previous column (and about 1/6 that of OLS). For the remaining variables, the average bias is about 2/3 that of associated with classical IV (and again about 1/6 that associated with OLS). In broader specifications, this also implies that OLS is likely to produce severely biased estimates of treatment or other policy effects. Oftentimes it is accurate inferences about these treatment effects, and not about the impacts of the health variables per se, that is of paramount interest to researchers and policymakers.

4 Conclusion

The analytical results in this paper suggest a feasible generalized instrumental variables approach for mitigating the bias associated with measurement error in explanatory variables. We allow for the possibility of correlated measurement errors across instruments, thus expanding the set of potentially valid instruments. Once refined, our approach will hopefully be useful in a broad range of policy-relevant applications for researchers in many fields.

Our Monte Carlo experiments illustrate the method’s ability to substantially mitigate the bias associated with measurement error. In these experiments, we found that OLS produces severely biased estimates of the effects of mental health conditions on average log length of stay (ALOS) in a Virginia mental hospital. In fact, one-third of the OLS coefficients on mental health conditions have the wrong signs in our simulation. Classical IV performed much better than OLS for each of our explanatory variables, with an average estimation bias only 1/3 that associated with OLS. Our new corrected IV approach performed better yet, with an average estimation bias of about 1/2 that of classical IV and 1/6 that of OLS. Even the coefficients on variables not measured with error (but correlated with the variables measured with error), such as race and gender, were severely biased under OLS compared with our corrected IV method. Again, there was about a six-fold difference in the degree of bias. In broader specifications, this also suggests that our method is likely to produce substantially improved estimates of treatment or other policy effects. Oftentimes it is accurate inferences about these treatment effects, and not about the impacts of the health variables per se, that is of paramount interest to researchers and policymakers.

Currently, our results are limited to standard linear specifications of the type typically estimated using OLS or classical IV. We have begun work to apply our

¹⁹Further analysis of small sample properties of the IV correction could lead to a rule of thumb for the adjustment factor.

generalized approach to nonlinear models (which cannot be consistently estimated using OLS or classical IV in any case) such as those involving binary outcome variables (e.g., leaving against medical advice or labor force participation) or duration (e.g., time to recovery or community tenure). An estimator for a nonlinear version of the outcome model can be derived based on Taylor's expansion of the outcome equation with respect to the coefficient vector. The linear approximation error introduces an additional source of bias into the estimation. We expect to be able to derive an analytical correction for this bias which may interact with the estimated covariance matrix.

This version of the paper does not provide details for the asymptotic properties of our estimators. In particular, we need to formally prove consistency of the estimator, find the rate at which it converges to the true parameter as the sample size gets large, and derive its asymptotic distribution. We will also analyze the rate at which the covariance matrix adjustment factor converges to one as the sample size goes to infinity in order to maximize the rate of convergence of the corrected IV estimator. From a practical perspective, it will be important to develop rules of thumb conditional on the sample size and other characteristics of the data for choosing an appropriate adjustment factor for any particular estimation problem. We also need to develop estimators of standard errors for our corrected IV procedure. We anticipate being able to show that our estimators are asymptotically normal with an estimable covariance matrix; at any rate, we intend to explore bootstrapping methods for small samples.

References

- [1] Black, Dan, Mark Berger, and Frank Scott (2000). "Bounding Parameter Estimates with Non-Classical Measurement Error." *Journal of the American Statistical Association*. 95: 739-48.
- [2] Clark, L., J. McEwen, L. Collard, and L. Hickok (1993). "Symptoms and Traits of Personality Disorder: Two New Methods for Their Assessment." *Psychological Assessments*. 5: 81-91.
- [3] Clarke, D., G. Smith, and H. Hermann (1998). "Monash Interview for Liaison Psychiatry (MILP): Development, Reliability, and Procedural Validity." *Psychosomatics*. 39: 318-328.
- [4] Cohen, Jacob (1960). "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement*. 20(1): 37-46.
- [5] Fennig, S., J. Craig, M. Tanenberg-Karant, and E. Bromet (1994). "Comparison of Facility and Research Diagnoses in First-Admission Psychotic Patients." *American Journal of Psychiatry*. 151(10): 1423-1429.
- [6] Fuller, Wayne (1987). *Measurement Error Models*. New York: John Wiley & Sons.

- [7] Greene, William (1997). *Econometric Analysis*. 3rd Edition. Saddle River, N.J.: Prentice Hall.
- [8] Hiller, W., G. Dichtl, H. Hecht, W. Hundt, and D. Zerssen (1994). "Testing the Comparability of Psychiatric Diagnoses in ICD-10 and DSM-III-R." *Psychopathology*. 27: 19-28.
- [9] Holt, Fredrick, Elizabeth Merwin, and Steven Stern (1999). "title." University of Virginia, unpublished manuscript.
- [10] Kahan, J., S. Bernstein, L. Leape, L. Hilborne, R. Park, L. Parker, C. Kamberg, and R. Brook (1994). "Measuring the Necessity of Medical Procedures." *Medical Care*. 32(4): 357-365.
- [11] Kelly, T. and J. Mann (1996). "Validity of DSM-III-R Diagnosis by Psychological Autopsy: A Comparison with Clinician Ante-Mortem Diagnosis." *Acta Psychiatrica Scandinavica*. 94: 337-343.
- [12] McGorry, P., C. Mihalopoulos, L. Henry, J. Dakis, H. Jackson, M. Flaum, S. Harrigan, D. McKenzie, J. Kulkarni, and R. Karoly (1995). "Spurious Precision: Procedural Validity of Diagnostic Assessment in Psychotic Disorders." *American Journal of Psychiatry*. 152:2.
- [13] Newhouse, Joseph and Mark McClellan (1998). "Econometrics in Outcomes Research: The Use of Instrumental Variables." *Annual Review of Public Health*. 19: 17-34.
- [14] Parker, T., P. May, M. Maviglia, S. Petrakis, S. Sunde, and S. Gloyd (1997) "PRIME-MD: Its Utility in detecting Mental Disorders in American Indians." *International Journal of Psychiatry in Medicine*. 27(2): 107-128.
- [15] Riskind, J., A. Beck, R. Berchick, G. Brown, and R. Steer (1987). "Reliability of DSM-III Diagnoses for Major Depression and Generalized Anxiety Disorder Using the Structured Clinical Interview for DSM-III." *Archives of General Psychiatry*. 44: 817-820.
- [16] Rosenman, S., A. Korten, and C. Levings (1997). "Computerized Diagnosis in Acute Psychiatry: Validity of CIDI-Auto Against Routine Clinical Diagnosis." *Journal of Psychist. Res.* 31(5): 581-592.
- [17] Roy, M., G. Lanctot, C. Merette, D. Cliché, J. Fournier, P. Boutin, C. Rodrigue, L. Charron, M. Turgeon, M. Hamel, N. Montgrain, L. Nicole, A. Pires, H. Wallot, A. Ponton, Y. Garneau, C. Dion, J. Lavalée, A. Potvin, P. Szatmari, and M. Maziade (1997). "Clinical and Methodological Factors Related to Reliability of the Best-Estimate Diagnostic Procedure." *American Journal of Psychiatry*. 154(12): 1726-1733.
- [18] Stravynski, A., Y. Lamontagne, and Y. Lavalée (1986). "Clinical Phobias and Avoidant Personality Disorder Among Alcoholics Admitted to an Alcoholism Rehabilitation Setting." *Canadian Journal of Psychiatry*. 31: 714-719.

- [19] Usten, B., W. Compton, D. Mager, T. Babor, O. Baiyewu, S. Chatterji, L. Cottler, A. Gogus, V. Mavreas, L. Peters, C. Pull, J. Saunders, R. Smeets, M. Stipek, R. Vrasit, D. Hasin, R. Room, W. Wan den Brink, D. Regier, J. Blaine, B. Grant, and N. Sartorius (1997) “WHO Study on the Reliability and Validity of the Alcohol and Drug Use Disorder Instruments: Overview of Methods and Results.” *Drug and Alcohol Dependence*. 47: 161-169.

Table 1

Monte Carlo Results for the First Experiment

Panel A: OLS Estimates When There is no Measurement Error

Variable	True Value	Median Bias and 80% Confidence Interval	Variable	True Value	Median Bias and 80% Confidence Interval
z_1	0.050	0.000 (-0.003,0.002)	z_2	0.030	0.000 (-0.002,0.002)
z_3	0.120	0.000 (-0.003,0.002)	z_4	0.090	0.000 (-0.002,0.002)
x_1	0.110	0.000 (-0.005,0.005)	x_2	0.120	0.000 (-0.005,0.004)
x_3	0.130	-0.001 (-0.005,0.005)	x_4	0.140	0.000 (-0.005,0.005)
x_5	0.050	0.000 (-0.004,0.005)	x_6	0.060	0.001 (-0.003,0.004)
x_7	0.070	0.000 (-0.005,0.005)	x_8	0.080	-0.001 (-0.005,0.004)
x_9	-0.090	0.000 (-0.005,0.005)	x_{10}	-0.100	0.000 (-0.005,0.004)
x_{11}	-0.110	0.000 (-0.005,0.004)	x_{12}	-0.120	0.000 (-0.005,0.004)
x_{13}	-0.130	0.000 (-0.004,0.004)			

Panel B: OLS Estimates When There is Measurement Error

Variable	True Value	Median Bias and 80% Confidence Interval	Variable	True Value	Median Bias and 80% Confidence Interval
z_1	0.050	0.000 (-0.003,0.002)	z_2	0.030	0.000 (-0.004,0.003)
z_3	0.120	-0.001 (-0.003,0.003)	z_4	0.090	0.000 (-0.003,0.003)
x_1	0.110	-0.064 (-0.068,-0.061)	x_2	0.120	-0.071 (-0.076,-0.066)
x_3	0.130	-0.076 (-0.080,-0.071)	x_4	0.140	-0.084 (-0.089,-0.080)
x_5	0.050	-0.025 (-0.029,-0.021)	x_6	0.060	-0.030 (-0.034,-0.025)
x_7	0.070	-0.037 (-0.042,-0.033)	x_8	0.080	-0.045 (-0.050,-0.041)
x_9	-0.090	0.066 (0.062,0.070)	x_{10}	-0.100	0.075 (0.070,0.080)
x_{11}	-0.110	0.082 (0.077,0.086)	x_{12}	-0.120	0.086 (0.081,0.090)
x_{13}	-0.130	0.085 (0.082,0.089)			

Panel C: IV Estimates When There is Measurement Error Without Correction

Variable	True Value	Median Bias and 80% Confidence Interval	Variable	True Value	Median Bias and 80% Confidence Interval
z_1	0.050	0.000 (-0.003,0.004)	z_2	0.030	0.000 (-0.004,0.004)
z_3	0.120	-0.001 (-0.003,0.003)	z_4	0.090	0.001 (-0.003,0.004)
x_1	0.110	-0.027 (-0.035,-0.021)	x_2	0.120	-0.029 (-0.039,-0.021)
x_3	0.130	-0.031 (-0.038,-0.023)	x_4	0.140	-0.036 (-0.044,-0.030)
x_5	0.050	-0.012 (-0.018,-0.005)	x_6	0.060	-0.011 (-0.018,-0.005)
x_7	0.070	-0.015 (-0.023,-0.008)	x_8	0.080	-0.019 (-0.028,-0.012)
x_9	-0.090	0.028 (0.021,0.033)	x_{10}	-0.100	0.033 (0.024,0.041)
x_{11}	-0.110	0.035 (0.027,0.042)	x_{12}	-0.120	0.035 (0.027,0.044)
x_{13}	-0.130	0.033 (0.027,0.040)			

Panel D: IV Estimates When There is Measurement Error With
Correction
with $N = 1000$ and $H = 3$

Variable	True Value	Median Bias and 80% Confidence Interval	Variable	True Value	Median Bias and 80% Confidence Interval
z_1	0.050	0.000 (-0.004,0.005)	z_2	0.030	0.000 (-0.004,0.004)
z_3	0.120	0.000 (-0.004,0.004)	z_4	0.090	0.001 (-0.004,0.005)
x_1	0.110	-0.004 (-0.025,0.032)	x_2	0.120	0.001 (-0.031,0.040)
x_3	0.130	0.002 (-0.033,0.035)	x_4	0.140	-0.004 (-0.033,0.036)
x_5	0.050	0.001 (-0.024,0.026)	x_6	0.060	0.002 (-0.025,0.027)
x_7	0.070	0.002 (-0.026,0.029)	x_8	0.080	0.006 (-0.024,0.043)
x_9	-0.090	0.002 (-0.030,0.025)	x_{10}	-0.100	0.001 (-0.041,0.030)
x_{11}	-0.110	-0.006 (-0.034,0.033)	x_{12}	-0.120	-0.006 (-0.037,0.023)
x_{13}	-0.130	-0.002 (-0.034,0.019)			

Panel E: IV Estimates When There is Measurement Error With
Correction
with $N = 1000$ and $H = 5$

Variable	True Value	Median Bias and 80% Confidence Interval	Variable	True Value	Median Bias and 80% Confidence Interval
z_1	0.050	-0.001 (-0.005,0.005)	z_2	0.030	0.001 (-0.005,0.004)
z_3	0.120	0.000 (-0.005,0.004)	z_4	0.090	0.000 (-0.005,0.004)
x_1	0.110	0.006 (-0.023,0.048)	x_2	0.120	0.000 (-0.034,0.033)
x_3	0.130	0.004 (-0.022,0.035)	x_4	0.140	0.001 (-0.032,0.035)
x_5	0.050	0.001 (-0.017,0.022)	x_6	0.060	-0.002 (-0.025,0.023)
x_7	0.070	0.001 (-0.028,0.027)	x_8	0.080	-0.001 (-0.029,0.025)
x_9	-0.090	-0.001 (-0.027,0.025)	x_{10}	-0.100	-0.002 (-0.035,0.027)
x_{11}	-0.110	0.003 (-0.054,0.029)	x_{12}	-0.120	-0.002 (-0.041,0.032)
x_{13}	-0.130	-0.005 (-0.038,0.023)			

Panel F: IV Estimates When There is Measurement Error With
Correction
with $N = 5000$ and $H = 3$

Variable	True Value	Median Bias and 80% Confidence Interval	Variable	True Value	Median Bias and 80% Confidence Interval
z_1	0.050	0.000 (-0.005,0.004)	z_2	0.030	0.0000 (-0.005,0.004)
z_3	0.120	0.000 (-0.005,0.004)	z_4	0.090	0.000 (-0.003,0.005)
x_1	0.110	0.000 (-0.015,0.016)	x_2	0.120	0.002 (-0.015,0.020)
x_3	0.130	-0.002 (-0.015,0.017)	x_4	0.140	0.000 (-0.019,0.019)
x_5	0.050	0.000 (-0.013,0.013)	x_6	0.060	0.000 (-0.013,0.013)
x_7	0.070	0.001 (-0.011,0.016)	x_8	0.080	-0.001 (-0.016,0.017)
x_9	-0.090	-0.001 (-0.018,0.013)	x_{10}	-0.100	0.000 (-0.016,0.015)
x_{11}	-0.110	0.000 (-0.022,0.015)	x_{12}	-0.120	0.001 (-0.018,0.016)
x_{13}	-0.130	-0.002 (-0.017,0.011)			

Table 2

Characteristics of 1980 Virginia Data

Variable	Mean	St. Dev.	Definition
Ln(Spell)	3.27	1.56	Ln Spell Length
Black	0.31	0.46	Dummy for Black Race
Female	0.40	0.49	Dummy for Female
Facility1	0.16	0.37	Dummy for Facility 1
Facility2	0.10	0.30	Dummy for Facility 2
Facility3	0.12	0.32	Dummy for Facility 3
Facility4	0.24	0.43	Dummy for Facility 4
Facility5	0.21	0.41	Dummy for Facility 5
Facility6	0.10	0.30	Dummy for Facility 6
Facility7	0.05	0.23	Dummy for Facility 7
Facility8	0.02	0.13	Dummy for Facility 8
Married	0.15	0.36	Dummy for married
Age	39.11	15.19	Age of patient
Dementia	0.03	0.17	Dummy for diagnosis = Dementia
Substance Abuse	0.05	0.21	Dummy for diagnosis = Substance Abuse
Alcohol	0.15	0.36	Dummy for diagnosis = Alcohol Abuse
Organic	0.06	0.25	Dummy for diagnosis = Organic
Schizophrenia	0.22	0.42	Dummy for diagnosis = Schizophrenia
Schizoaffective	0.08	0.27	Dummy for diagnosis = Schizoaffective
Paranoid	0.01	0.10	Dummy for diagnosis = Paranoid
Other Psychotic	0.06	0.23	Dummy for diagnosis = Other Psychotic
Bipolar	0.11	0.31	Dummy for diagnosis = Bipolar
Depression	0.11	0.31	Dummy for diagnosis = Depression
Personality	0.02	0.14	Dummy for diagnosis = Personality problem
Adjustment	0.07	0.25	Dummy for diagnosis = Adjustment problem
Other Condition	0.04	0.19	Dummy for diagnosis = Other
NumPrevHosp	2.91	3.43	Number of Previous Hospital Stays since 1978

Table 3

IV Estimates of Model Parameters from 1980 Virginia Data
(coefficients are defined as “true” values in the subsequent simulations)

Variable	Estimate	Variable	Estimate
Black	-0.0571 (0.0504)	Schizophrenia	-0.659* (0.238)
Female	0.0211 (0.0465)	SubstAbuse	-2.53* (0.279)
Married	-0.279* (0.0593)	Alcohol	-3.36* (0.234)
Age	0.00782* (0.00200)	Organic	-0.0571 (0.341)
Facility1	4.21* (0.274)	Schizoaffective	-0.659* (0.238)
Facility2	4.07* (0.279)	Paranoid	-1.38* (0.633)
Facility3	4.16* (0.279)	OtherPsychotic	-0.855* (0.351)
Facility4	4.40* (0.271)	Bipolar	-1.13* (0.242)
Facility5	4.29* (0.278)	Depression	-0.957* (0.249)
Facility6	4.85* (0.280)	Personality	-1.21* (0.415)
Facility7	4.68 (0.270)	Adjustment	-2.90* (0.290)
Facility8	5.05* (0.302)	OtherCondition	-1.59* (0.361)
NumPrevHosp	0.0229* (0.00658)		
σ_u	1.99	σ_e	1.11

Notes:

1. Starred items are significant at the 5% level.
2. Numbers in parentheses are standard errors.
3. Omitted condition is dementia
4. No standard errors have yet been estimated for $\hat{\sigma}_u$ and $\hat{\sigma}_e$

Table 4
Results of Monte Carlo Experiment
(Dependent variable: Ln Length of Stay)

Variable	Truth	Median Bias and 80% Confidence Interval for Bias		
		OLS	Classical IV	Corrected IV
Schizophrenia	-0.659	0.760 (0.660,0.878)	0.267 (-0.131,0.617)	0.116 (-0.302,0.540)
Substance Abuse	-2.54	1.80 (1.66,1.93)	0.406 (-0.079,0.880)	0.256 (-0.259,0.763)
Alcohol	-3.36	2.04 (1.92,2.14)	0.394 (-0.034,0.778)	0.252 (-0.229,0.675)
Organic	-0.0571	0.425 (0.273,0.533)	0.200 (-0.204,0.725)	0.0564 (-0.404,0.659)
Schizoaffective	-0.656	0.693 (0.590,0.850)	0.224 (-0.221,0.652)	0.0855 (-0.424,0.574)
Paranoid	-1.38	1.18 (0.985,1.35)	0.333 (-0.314,1.027)	0.164 (-0.548,0.901)
Other Psychotic	-0.855	0.797 (0.692,0.937)	0.261 (-0.216,0.658)	0.120 (-0.432,0.542)
Bipolar	-1.13	0.983 (0.862,1.09)	0.265 (-0.0623,0.698)	0.125 (-0.261,0.593)
Depression	-0.957	0.894 (0.762,1.00)	0.274 (-0.138,0.663)	0.123 (-0.347,0.585)
Personality	-1.21	1.04 (0.865,1.20)	0.362 (-0.259,0.789)	0.217 (-0.462,0.683)
Adjustment	-2.90	1.97 (1.83,2.10)	0.406 (-0.0409,0.828)	0.245 (-0.228,0.720)
Other Condition	-1.59	1.241 (1.10,1.38)	0.334 (-0.0960,0.728)	0.199 (-0.304,0.651)
Black	-0.0571	0.201 (0.161, 0.240)	0.0425 (-0.0175, 0.0853)	0.0363 (-0.0236, 0.0816)
Female	0.0211	0.217 (0.175, 0.260)	0.0296 (-0.0142, 0.0877)	0.0294 (-0.0141,0.0886)
Married	-0.279	-0.0019 (-0.0447,0.0520)	-0.0010 (-0.0529,0.0397)	-0.0011 (-0.0526,0.0388)
Age	0.0078	0.0081 (0.0066,0.0094)	0.0017 (-0.0010,0.0038)	0.0011 (-0.0016,0.0033)

Notes:

1. Reported statistics are $\frac{\text{median bias}}{\text{(80\% confidence interval)}}$.
2. The omitted diagnosis category is dementia.
3. Results for some non-diagnosis variables are not reported to save space.

Appendix A. Effects of Allowing for Comorbidity

Instead of specifying a process like equation (7):

$$p_{ik} = \Pr [x_{ijk} = 1 | q_i] = \frac{\exp \{q_{ik}\}}{\sum_l \exp \{q_{il}\}},$$

we can specify a process like

$$p_{ik} = \Pr [x_{ijk} = 1 | q_i] = \frac{\exp \{q_{ik}\}}{1 + \exp \{q_{ik}\}}.$$

This will imply a very different covariance matrix. In particular, it will be diagonal. We can add off-diagonal terms by just allowing for some unobserved error correlated over diagnoses. For example,

$$p_{ik}(u_i) = \Pr [x_{ijk} = 1 | q_i] = \frac{\exp \{q_{ik} + u_{ijk}\}}{1 + \exp \{q_{ik} + u_{ijk}\}}$$

where u_i is a vector of errors with some covariance matrix Ω . Then the covariance matrix for x_i will be

$$\begin{aligned} E(x_i - Ep_i)(x_i - Ep_i)' &= \int \cdots \int [diag p_i(u_i) - p_i(u_i)p_i'(u_i)] f(u_i | \Omega) du_i \\ &\quad + \int \cdots \int [p_i(u_i) - Ep_i][p_i(u_i) - Ep_i]' f(u_i | \Omega) du_i \end{aligned}$$

where

$$Ep_i = \int \cdots \int p_i(u_i) f(u_i | \Omega) du_i.$$

The first term is the expected value of the conditional covariance matrix, and the second term is the covariance matrix of the conditional mean.

The IV estimator with and without correlated measurement error remains the same.

Estimating the covariance matrix of the measurement error changes somewhat. We can no longer ask an expert what is the probability of being diagnosed with condition x if you really have condition y . In fact, it is not obvious how to do this with the Delphi method. The other two methods have no problems.

Appendix B: Aggregation of Health Conditions

General Problem

For practical reasons, mental health services researchers are forced to combine different types of psychiatric conditions into relatively small classes and then treat diagnosis codes of the same class as identical in statistical models (cites). There are far too many different types of mental health conditions (and variants of these conditions) to allow each of them to have differential effects on outcomes; the researcher needs to commit to some aggregation scheme.

Unfortunately, such aggregation typically implies misspecification of the true relationships between mental health status and the outcome measure of interest. The purpose of this section is to learn about optimal aggregation schemes. The optimal method of aggregation is not obvious as various tradeoffs must be confronted. For example, utilizing a large number of codes allows for many differential impacts on the outcome measure, but it becomes more likely that the codes will be contaminated with substantial measurement error. Moreover, the researcher must be concerned with degrees of freedom. Alternately, a small number of codes mitigates the extent of measurement error but requires aggregation of diagnoses with differential impacts.

One fundamental difficulty confronting researchers is that available data are not likely to include the true set of all conditions affecting the outcome variable of interest. Diagnoses information will have already been aggregated in some fashion by the time the information is coded in the dataset, and some variants of mental health conditions have presumably not yet been discovered by the medical community. Since it is likely that *any* aggregation strategy will result in misspecification for a particular application, the goal should be to choose a strategy that is expected to lead to the most reliable statistical inferences given the available data. Among other preliminary results, we can show mathematically that the desirability of assigning different mental health conditions to the same aggregation group tends to rise if (a) the conditions have similar impacts on the outcome variable, or (b) they are relatively likely to be mistaken for each other (relative to other conditions) when diagnoses are made. Our estimated covariance matrices (described above) will provide information about (b).

We first introduce some notation for the classification of psychiatric health conditions: Let $z_i^* = (z_{i1}^*, z_{i2}^*, \dots, z_{iJ_z}^*)$, with $\sum_j z_{ij}^* = 1$ for each patient i , represent the vector of J_z binary conditions that potentially have differential impacts on the outcome measure of interest y_i . Although the investigator does not observe z^* in the data, the true relationship between these conditions and the outcome measure is assumed to have the form

$$y_i = \sum_{j=1}^{J_z} \beta_j z_{ij}^* + u_i, \quad i = 1, 2, \dots, n \quad (16)$$

where β_j reflects the true impact of the j th condition on y and u is a random disturbance. (In the case of instrumental variables, we modeled true diagnoses as continuous because it was important to have measurement errors with zero mean. In this section, (a) it not necessary to have zero-mean measurement errors, and (b) it is easier to work with binary true diagnoses. Without loss of generality, we can allow some of the z^* s to represent nonmedical individual attributes such as gender, race, and health care facility. Instead of observing z_i^* , the researcher observes the set of classifications $x_i^* = (x_{i1}^*, x_{i2}^*, \dots, x_{iJ_x}^*)$ with $\sum_j x_{ij}^* = 1$. There is no requirement that $J_x = J_z$. The correspondence between z^* and x^* is not known to the researcher. With some probability, x_{ij}^* will be coded as the patient's true condition when in fact the true condition is z_{ij}^* . We

specify these probabilities (assumed invariant across individuals) as

$$p_{jk} = \Pr [x_{ij}^* = 1 \mid z_{ik}^* = 1].$$

Note that measurement error exists unless $p_{jj} = 1$ for all conditions j .

Assume that the number of x^* -categories J_x is a large number relative to the number of patients n so that some sort of aggregation is warranted. Consider an aggregation mechanism where x^* is partitioned into J_a subsets ($J_a < J_x$) where all the elements of any particular subset are aggregated together. Let $a = (a_1, a_2, \dots, a_{J_a})$ where a_k is a subset of the positive integers $\{1, 2, \dots, J_x\}$ such that for any two integers l and m , x_{il}^* and x_{im}^* are aggregated together if and only if l and m are both elements of the same set a_k . More formally, let a partition $(1, 2, \dots, J_x)$, and let

$$x_{ij} = 1 \text{ iff } \exists k \in a_j : x_{ik}^* = 1.$$

For example, if there are six observed categories in x^* denoted by $\{x_1^*, x_2^*, x_3^*, x_4^*, x_5^*, x_6^*\}$, then one possible aggregation mechanism involving three classes is $(x_1 = \{x_1^*, x_2^*, x_4^*\}, x_2 = \{x_5^*\}, x_3 = \{x_3^*, x_6^*\})$. That means, for example, that x_1^*, x_2^* , and x_4^* are coded identically in the researcher's aggregation strategy. This partition implies misspecification, however, unless (a) it is actually true that $\beta_1 = \beta_2 = \beta_4$ and $\beta_3 = \beta_6$ and (b) these codes include all of the relevant z^* s in (16).

To evaluate the consequences of a particular aggregation choice, we can define a new equation to estimate as

$$y_i = \sum_{j=1}^{J_a} b_j x_{ij} + e_i, \quad i = 1, 2, \dots, n, \quad (17)$$

which reflects the chosen aggregation along with any measurement error in the diagnoses. This equation can be written as

$$y_i = \sum_{j=1}^{J_a} b_j \sum_{l \in a_j} \sum_{k=1}^{J_z} r_{ikl} z_{ik}^* + e_i, \quad i = 1, 2, \dots, n \quad (18)$$

for appropriate binary indicators r_{ikl} .

Let $b = (b_1 b_2, \dots, b_{J_a})'$, and let \hat{b} be the ordinary least squares (OLS) estimator of b . We are interested in learning about the statistical properties of \hat{b} so that we can provide guidance about the most desirable aggregation strategy given the available data.

We have conducted some preliminary simulations to provide insights about optimal aggregation strategies. We are currently analyzing different formal criteria (such as mean squared error) for assessing which types of aggregation schemes provide the best estimates. The best way to present results will follow from our chosen criteria.

Best Case

In some cases, it will be fully appropriate and desirable to aggregate. In particular, if two or more conditions have the same true effect on the outcome variable, then they should be combined; such aggregation imposes valid constraints in the estimation which improves efficiency. In the case that all health conditions in each aggregation group truly have the same impact on y and there is no measurement error across elements of different groups, then OLS estimation of equation (17) will provide more efficient estimates for equation (16). Measurement error confined to a particular aggregation group has no negative consequences for estimation if all the conditions in that group have the same impact on the outcome.²⁰

[check this: β s don't belong to aggregation groups, and there may be extra true conditions z^ s not included in the x^* s - so the dimensions might not be the same. Also changed definition of p_{jk} so check if subscripts need to be switched in proof.]*

Theorem 1 *If for each aggregation group j ,*

$$\beta_k = \beta_j^* \forall k \in a_j \quad (19)$$

and

$$\sum_{k \in a_j} p_{kl} = 1 \forall l \in a_j, \quad (20)$$

then $\hat{b} = \hat{\beta}^* = (\hat{\beta}_1^*, \hat{\beta}_2^*, \dots, \hat{\beta}_{J_a}^*)'$.

P proof. Given equation (19), we can rewrite equation (16) as

$$y_i = \sum_{j=1}^{J_a} \beta_j^* 1[\exists k \in a_j : z_{ik}^* = 1] + u_i, \quad i = 1, 2, \dots, n \quad (21)$$

where $1[\cdot]$ is the indicator function. Also, given equation (20), equation (18) can be written as

$$y_i = \sum_{j=1}^{J_a} b_j \sum_{l \in a_j} \sum_{k \in a_j} p_{kl} z_{ik}^* + e_i, \quad i = 1, 2, \dots, n. \quad (22)$$

Note that, if $z_{ik}^* = 1$ for some $k \in a_j$, then $z_{il}^* = 0 \forall l \neq k$ and

$$\sum_{l \in a_j} p_{kl} z_{ik}^* = \sum_{l \in a_j} p_{kl} = 1,$$

which implies that

$$\sum_{l \in a_j} \sum_{m \in a_j} p_{ml} z_{im}^* = \sum_{l \in a_j} p_{kl} z_{ik}^* = 1.$$

²⁰Note that the effects of two conditions might be the same for some outcome variables but not for others.

Also, if $z_{ik}^* = 0 \forall k \in a_j$, then

$$\sum_{l \in a_j} \sum_{m \in a_j} p_{ml} z_{im}^* = 0.$$

Thus, equation (22) can be written as

$$y_i = \sum_{j=1}^{J_a} b_j 1[\exists k \in a_j : z_{ik}^* = 1] + e_i, \quad i = 1, 2, \dots, n$$

which has the same form as equation (21). Thus, $\hat{b} = \hat{\beta}^*$. ■

Aggregation Without Measurement Error

In the case that different conditions aggregated into the same class have differential impacts on the outcome measure and there is no measurement error, each element of b converges to a weighted average of the elements of β associated with that aggregation group where the weights are prevalence rates of the true conditions. We can write equation (18) in matrix form as

$$\begin{aligned} y &= Z^{**} R A b + e \\ &= X b + e \end{aligned}$$

where $y = (y_1, y_2, \dots, y_n)'$,

$$\begin{aligned} Z^{**} &= \begin{pmatrix} z_1^* & 0 & \cdots & 0 \\ 0 & z_2^* & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & z_n^* \end{pmatrix}, \\ z_i^* &= (z_{i1}^*, z_{i2}^*, \dots, z_{iJ_z}^*) \end{aligned}$$

$$\begin{aligned} R &= \begin{pmatrix} R_1 \\ R_2 \\ \vdots \\ R_n \end{pmatrix}, \\ R_i &= \begin{pmatrix} r_{i11} & r_{i12} & \cdots & r_{i1J_x} \\ r_{i21} & r_{i22} & \cdots & r_{i2J_x} \\ \vdots & \vdots & \ddots & \vdots \\ r_{iJ_z1} & r_{iJ_z2} & \cdots & r_{iJ_zJ_x} \end{pmatrix} \end{aligned}$$

where the j 'th row of R_i , $R_{ij} = (r_{ij1}, r_{ij2}, \dots, r_{ijJ_x})'$ is a vector of multinomial random variables (unobserved to the researcher) with probabilities $(p_{j1}, p_{j2}, \dots, p_{jJ_x})$, and

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1J_a} \\ a_{21} & a_{22} & \cdots & a_{2J_a} \\ \vdots & \vdots & \ddots & \vdots \\ a_{J_x1} & a_{J_x2} & \cdots & a_{J_xJ_a} \end{pmatrix}$$

where $a_{jk} = 1 (k \in a_j)$. The OLS estimate of b is

$$\begin{aligned}\widehat{b} &= [X'X]^{-1} [X'y] \\ &= [A'R'Z^{**'}Z^{**}RA]^{-1} [A'R'Z^{**'}(Z^{**}\beta + e)].\end{aligned}\tag{23}$$

First consider the case where there is no measurement error:

$$p_{jk} = 1 (j = k)$$

and $R_i = I_{J_x} \forall i$. Then, with

$$Z^* = \begin{pmatrix} z_1^* \\ z_2^* \\ \vdots \\ z_n^* \end{pmatrix},$$

equation (23) becomes

$$\begin{aligned}\widehat{b} &= [A'Z^{*'}Z^*A]^{-1} [A'Z^{*'}(Z^*\beta + e)] \\ &= \begin{pmatrix} \frac{\sum_{j \in a_1} n_{zj} \beta_j}{\sum_{j \in a_1} n_{zj}} \\ \frac{\sum_{j \in a_2} n_{zj} \beta_j}{\sum_{j \in a_2} n_{zj}} \\ \vdots \\ \frac{\sum_{j \in a_{J_a}} n_{zj} \beta_j}{\sum_{j \in a_{J_a}} n_{zj}} \end{pmatrix} + [A'Z^{*'}Z^*A]^{-1} [A'Z^{*'}e].\end{aligned}\tag{24}$$

Since $\text{plim } Z^{*'}e = 0$, each element of b converges to a weighted average of the elements of β associated with the aggregation group associated with the element of b .

Aggregation With Measurement Error

When measurement error exists, the OLS estimator for each element of b converges to a weighted average of β where the weights depend upon the aggregation scheme, the prevalence rates of the true conditions z^* , and the probabilities that particular z^* conditions are miscoded as particular x^* conditions. In general, define s_j as the true prevalence of condition j and p_{jk} as $\Pr(x_k^* = 1 | z_j^* = 1)$ and consider an aggregation mechanism with a subset a_m of the J_x observed diagnoses. Then

$$\text{plim } \widehat{b}_m = \frac{\sum_{j=1}^{J_z} \sum_{k \in a_m} p_{jk} s_j \beta_j}{\sum_{j=1}^{J_z} \sum_{k \in a_m} p_{jk} s_j}.$$

For example, suppose there are only two true conditions z_1^* and z_2^* (with “no condition” representing a third category). Suppose further that although these

conditions have different effects β_1 and β_2 , they are impossible to distinguish when making diagnoses so they are aggregated into one category. If the true prevalence rate of the first condition is twice that of the second, then the estimated effect of this group will converge to the value $(2/3)\beta_1 + (1/3)\beta_2$. This results is shown as follows:

When there is measurement error, equation (23) becomes

$$\begin{aligned} \hat{b} &= [A'R'Z^{**'}Z^{**}RA]^{-1}[A'R'Z^{**'}(Z^{**}\beta + e)] \\ &= \begin{pmatrix} \frac{\sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_1} r_{ijk} z_{ij}^* \beta_j}{\sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_1} r_{ijk} z_{ij}^*} \\ \frac{\sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_2} r_{ijk} z_{ij}^* \beta_j}{\sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_2} r_{ijk} z_{ij}^*} \\ \vdots \\ \frac{\sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_{J_a}} r_{ijk} z_{ij}^* \beta_j}{\sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_{J_a}} r_{ijk} z_{ij}^*} \end{pmatrix} + [A'R'Z^{**'}Z^{**}RA]^{-1}[A'R'Z^{**'}e] \end{aligned} \quad (26)$$

Consider the plim of an arbitrary element of equation (26):

Theorem 2 For an arbitrary element of \hat{b} ,

$$\text{plim} \hat{b}_m = \frac{\sum_{j=1}^{J_z} \sum_{k \in a_m} p_{jk} s_j \beta_j}{\sum_{j=1}^{J_z} \sum_{k \in a_m} p_{jk} s_j}$$

where $s_j = \text{plim} n^{-1} \sum_{i=1}^n z_{ij}^*$ is the proportion of the population with $z_{ij}^* = 1$.

P proof.

$$\begin{aligned} \text{plim} \hat{b}_m &= \text{plim} \frac{n^{-1} \sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_m} r_{ijk} z_{ij}^* \beta_j}{n^{-1} \sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_m} r_{ijk} z_{ij}^*} \\ &= \frac{\text{plim} n^{-1} \sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_m} r_{ijk} z_{ij}^* \beta_j}{\text{plim} n^{-1} \sum_{i=1}^n \sum_{j=1}^{J_z} \sum_{k \in a_m} r_{ijk} z_{ij}^*} \\ &= \frac{\sum_{j=1}^{J_z} \sum_{k \in a_m} p_{jk} s_j \beta_j}{\sum_{j=1}^{J_z} \sum_{k \in a_m} p_{jk} s_j}. \end{aligned} \quad (27)$$

■

Optimal Aggregation With Measurement Error

Now we are ready to consider a criterion for choosing an optimal aggregation mechanism. Consider an objective function such as

$$L = E(\hat{b} - \gamma)' W(\hat{b} - \gamma);$$

the goal is to minimize L over choices of A given γ and W . First we must decide what is a good choice of γ . In the case without measurement error, the

researcher usually has in mind that for some aggregation group a_k , all of the elements of $\{\beta_j : j \in a_k\}$ are close, and then γ_k is some average measure of the elements of $\{\beta_j : j \in a_k\}$. An obvious measure would be

$$\gamma_k = \frac{\sum_{j \in a_k} s_j \beta_j}{\sum_{j \in a_k} s_j} \quad (28)$$

which corresponds to the spirit of equation (24). When there is measurement error, there are two problems with choosing γ according to equation (28): a) we have no consistent estimate of $s = (s_1, s_2, \dots, s_{J_2})$,²¹ and b) we may want to use information about the prevalence of measurement error summarized in P to construct an aggregation mechanism. The second point is important. Consider the extreme case where there exists a partition of $\{1, 2, \dots, J_x\}$, $(c_1, c_2, \dots, c_{J_c})$, such that

$$p_{jk} = \frac{1}{M_m} \mathbf{1}(j, k \in c_m) \quad (29)$$

where $M_m = \#c_m$. Equation (29) says that one has no information to distinguish between different diagnoses in any set c_m but one never mistakes a true diagnosis in c_m for one outside c_m . In this case, it makes no sense to construct an aggregation mechanism such that

$$\exists \text{ aggregation group } a_m : \exists j, k \in c_l \text{ with } j \in a_m \text{ and } k \notin a_m. \quad (30)$$

This can be formalized later once we have some precision associated with some of the concepts here.

In some applications, γ is given by the research problem. For example, we may be particularly interested in β_j or a particular weighted average of β .

Things to show: a) as the distance between β_j and β_k increases, the value of having j and k in the same aggregation group decreases; b) as p_{jk} and p_{kj} increase, the value of having j and k in the same aggregation group increases.

Intuition on point (a): Assume temporarily that there is no measurement error and write equation (16) as

$$\begin{aligned} y &= X^* \beta + u, \\ u &\sim N(0, \sigma^2 I), \end{aligned}$$

and assume

$$\Psi \beta - \phi = \varepsilon$$

where ε is a vector of small, unknown numbers (not equal to zero). As $\varepsilon \rightarrow 0$, then the efficient estimate of β becomes

$$\widehat{\beta}_R = (X^{*'} X^*)^{-1} (X^{*'} y) - (X^{*'} X^*)^{-1} \Psi' \left[\Psi (X^{*'} X^*)^{-1} \Psi' \right]^{-1} \left[\Psi (X^{*'} X^*)^{-1} (X^{*'} y) - \phi \right].$$

²¹This occurs because there is no guarantee that measurement error is unbiased which would be necessary for a Law of Large Numbers to apply. For the purposes of estimating s consistently, unbiasedness would require that $\sum_{k \neq j} p_{kj} = \sum_{k \neq j} p_{jk}$. Lack of consistency is, in itself, a significant problem in that it implies that one should view suspiciously any measures of population prevalence of any illness measured with large error.

For $\varepsilon \neq 0$,

$$E\widehat{\beta}_R = \beta - (X^{*'}X^*)^{-1}\Psi' \left[\Psi (X^{*'}X^*)^{-1} \Psi' \right]^{-1} \varepsilon,$$

and

$$\widehat{\beta}_R - E\widehat{\beta}_R = (X^{*'}X^*)^{-1} \left\{ I - \Psi' \left[\Psi (X^{*'}X^*)^{-1} \Psi' \right]^{-1} \Psi (X^{*'}X^*)^{-1} \right\} (X^{*'}u),$$

$$\begin{aligned} D(\widehat{\beta}_R) &= E(\widehat{\beta}_R - E\widehat{\beta}_R)(\widehat{\beta}_R - E\widehat{\beta}_R)' \\ &= \sigma^2 \left\{ (X^{*'}X^*)^{-1} - (X^{*'}X^*)^{-1} \Psi' \left[\Psi (X^{*'}X^*)^{-1} \Psi' \right]^{-1} \Psi (X^{*'}X^*)^{-1} \right\}. \end{aligned}$$

Thus, the mean squared error is

$$\begin{aligned} MSE(\widehat{\beta}_R) &= (E\widehat{\beta}_R - \beta)(E\widehat{\beta}_R - \beta)' + D(\widehat{\beta}_R) \\ &= (X^{*'}X^*)^{-1} \Psi' \left[\Psi (X^{*'}X^*)^{-1} \Psi' \right]^{-1} \varepsilon \varepsilon' \left[\Psi (X^{*'}X^*)^{-1} \Psi' \right]^{-1} \Psi (X^{*'}X^*)^{-1} \\ &\quad + \sigma^2 \left\{ (X^{*'}X^*)^{-1} - (X^{*'}X^*)^{-1} \Psi' \left[\Psi (X^{*'}X^*)^{-1} \Psi' \right]^{-1} \Psi (X^{*'}X^*)^{-1} \right\}. \end{aligned}$$

On the other hand, the mean squared error of

$$\widehat{\beta}_U = (X^{*'}X^*)^{-1} (X^{*'}y)$$

is $MSE(\widehat{\beta}_U) = \sigma^2 (X^{*'}X^*)^{-1}$. There is a critical value ε^* such that $MSE(\widehat{\beta}_R) < MSE(\widehat{\beta}_U)$ for all $\varepsilon < \varepsilon^*$ and $MSE(\widehat{\beta}_R) > MSE(\widehat{\beta}_U)$ for all $\varepsilon > \varepsilon^*$. In particular, if

$$\varepsilon \varepsilon' = \sigma^2 \Psi (X^{*'}X^*)^{-1} \Psi',$$

then $MSE(\widehat{\beta}_R) = MSE(\widehat{\beta}_U)$.

Intuition on point (b): Consider a simple case where $J_z = 3$ and

$$P = \begin{pmatrix} p_{11} & 1 - p_{11} & 0 \\ 1 - p_{22} & p_{22} & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Consider two aggregation mechanisms, $a^* = \{(1), (2), (3)\}$ and $a^{**} = \{(1, 2), (3)\}$. Under a^* , equation (17) becomes

$$\begin{aligned} y_i &= x_{i1}b_1^* + x_{i2}b_2^* + x_{i3}b_3^* + e_i^* \\ &= (r_{i11}z_{i1}^* + r_{i21}z_{i2}^*)b_1^* + (r_{i12}z_{i1}^* + r_{i22}z_{i2}^*)b_2^* + z_{i3}^*b_3^* + e_i^*, \end{aligned}$$

and, under a^{**} , it becomes

$$\begin{aligned} y_i &= 1(x_{i1} = 1 \text{ or } x_{i2} = 1)b_1^{**} + x_{i3}b_3^{**} + e_i^{**} \\ &= 1(z_{i1}^* = 1 \text{ or } z_{i2}^* = 1)b_1^{**} + z_{i3}^*b_3^{**} + e_i^{**}. \end{aligned}$$

Under a^* , from equation (27),

$$\text{plim} \begin{pmatrix} \widehat{b}_1^* \\ \widehat{b}_2^* \\ \widehat{b}_3^* \end{pmatrix} = \begin{pmatrix} \frac{p_{11}s_1\beta_1 + (1-p_{22})s_2\beta_2}{p_{11}s_1 + (1-p_{22})s_2} \\ \frac{(1-p_{11})s_1\beta_1 + p_{22}s_2\beta_2}{(1-p_{11})s_1 + p_{22}s_2} \\ \beta_3 \end{pmatrix},$$

and, under a^{**} ,

$$\text{plim} \begin{pmatrix} \widehat{b}_1^{**} \\ \widehat{b}_3^{**} \end{pmatrix} = \begin{pmatrix} \frac{s_1\beta_1 + s_2\beta_2}{s_1 + s_2} \\ \beta_3 \end{pmatrix}.$$

Note that, as $p_{11}, p_{22} \rightarrow 1$, $\text{plim}\widehat{b}^* \rightarrow \beta$, and, as $p_{11}, p_{22} \rightarrow \frac{1}{2}$,

$$\text{plim} \begin{pmatrix} \widehat{b}_1^* \\ \widehat{b}_2^* \\ \widehat{b}_3^* \end{pmatrix} \rightarrow \begin{pmatrix} \frac{s_1\beta_1 + s_2\beta_2}{s_1 + s_2} \\ \frac{s_1\beta_1 + s_2\beta_2}{s_1 + s_2} \\ \beta_3 \end{pmatrix}.$$

In this case, since $\widehat{b}_1^* - \widehat{b}_2^* \rightarrow 0$, there will be p^* such that if $p_{11}, p_{22} > p^*$, the efficiency gained from imposing the (untrue) constraint that $b_1^* = b_2^*$ will dominate the loss of consistency caused by $b_1^* \neq b_2^*$. Also note that, in the generic case (when $p_{11}, p_{22} < \frac{1}{2}$), if one knows s_1/s_2 and (p_{11}, p_{22}) , then one can identify β . However it is not trivial to observe s_1/s_2 and (p_{11}, p_{22}) in the presence of measurement error.

{Work out some numerical examples consistent with the IV simulations}

Appendix C. Adjustment Factor

Consider adding an adjustment factor a to the moment condition:

$$\widetilde{m}(\theta) = \frac{1}{n} \sum_i X_{i2} (y_i - X'_{i1}\widehat{\theta}) + a \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \widehat{\theta}. \quad (31)$$

When $a = 1$, we have the method we described, and when $a = 0$, we have standard IV. The issue is what is the optimal value of a . Consider an objective function such as mean squared error:

$$B[\widehat{\theta}(a)] + D[\widehat{\theta}(a)] \quad (32)$$

where

$$B[\widehat{\theta}(a)] = [E\widehat{\theta}(a) - \theta]' W_b [E\widehat{\theta}(a) - \theta],$$

$$D[\widehat{\theta}(a)] = \int \cdots \int [\widehat{\theta}(a) - E\widehat{\theta}(a)]' W_v [\widehat{\theta}(a) - E\widehat{\theta}(a)] f[\widehat{\theta}(a)] d\widehat{\theta}(a),$$

and

$$\begin{aligned}
E\widehat{\theta}(a) &= \int \cdots \int \widehat{\theta}(a) f[\widehat{\theta}(a)] d\widehat{\theta}(a) \\
&= E \left[\frac{1}{n} \sum_i X_{i2} X'_{i1} - a \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \left[\frac{1}{n} \sum_i X_{i2} y_i \right] \\
&\rightarrow \left[\int \cdots \int p(q) p'(q) f(q) dq + (1-a) \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \left[\int \cdots \int p(q) p'(q) f(q) dq \right] \theta,
\end{aligned}$$

and $|E\widehat{\theta}(a)| \leq \theta$. If $W_v = 0$, then $D[\widehat{\theta}(a)] = 0 \forall a$ and $a = 1$ minimizes equation (32). In general, $\frac{\partial}{\partial a} E\widehat{\theta}(a)$ can be found by taking the total derivative of equation (31) with respect to a and $\widehat{\theta}$:

$$\begin{aligned}
&-\frac{1}{n} \sum_i X_{i2} X'_{i1} d\widehat{\theta} + da \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \widehat{\theta} + a \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} d\widehat{\theta} = 0 \\
\Rightarrow \frac{\partial \widehat{\theta}}{\partial a} &= \left[\frac{1}{n} \sum_i X_{i2} X'_{i1} - a \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \widehat{\theta} \\
\rightarrow \frac{\partial E\widehat{\theta}}{\partial a} &= \left[\int \cdots \int p(q) p'(q) f(q) dq + (1-a) \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \theta \\
&= \left[\int \cdots \int p(q) p'(q) f(q) dq \right]^{-1} \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} E\widehat{\theta}
\end{aligned}$$

which moves in the same direction as $E\widehat{\theta}$. Therefore, since $\left[\int \cdots \int p(q) p'(q) f(q) dq \right]^{-1} \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix}$ is positive semidefinite, $\frac{\partial \widehat{\theta}}{\partial a}$ moves in the same direction as $E\widehat{\theta}(a)$, thus (usually) reducing attenuation bias. The derivative of the squared bias is

$$\frac{\partial}{\partial a} B[\widehat{\theta}(a)] = \frac{\partial}{\partial a} E\widehat{\theta}(a)' W_b [E\widehat{\theta}(a) - \theta] + [E\widehat{\theta}(a) - \theta]' W_b \frac{\partial}{\partial a} E\widehat{\theta}(a).$$

If $\frac{\partial}{\partial a} E\widehat{\theta}(a)$ has the same sign as $E\widehat{\theta}(a)$ and, therefore, the opposite sign of $E\widehat{\theta}(a) - \theta$, then $\frac{\partial}{\partial a} B[\widehat{\theta}(a)] < 0$. But, at $a = 1$, $E\widehat{\theta}(a) - \theta = 0$, so $\frac{\partial}{\partial a} B[\widehat{\theta}(a)] = 0$.

To find $\frac{\partial}{\partial a} D[\widehat{\theta}(a)]$, we must first write $D[\widehat{\theta}(a)]$ in terms of $D[\widehat{\Omega}]$. Define $vech\widehat{\Omega}$ to be the vector of independent elements of $\widehat{\Omega}$. Then the total derivative of equation (31) with respect to $\widehat{\theta}$ and $vech\widehat{\Omega}$ is

$$-\frac{1}{n} \sum_i X_{i2} X'_{i1} d\widehat{\theta} + a \frac{\partial}{\partial vech\widehat{\Omega}} \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \widehat{\theta} dvech\widehat{\Omega} + a \begin{pmatrix} \widehat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} d\widehat{\theta}$$

$$\begin{aligned}
&\Rightarrow \frac{\partial \hat{\theta}}{\partial \text{vech} \hat{\Omega}} = \left[\frac{1}{n} \sum_i X_{i2} X'_{i1} - a \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} a \frac{\partial}{\partial \text{vech} \hat{\Omega}} \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \hat{\theta} \\
&\rightarrow \left[\int \cdots \int p(q) p'(q) f(q) dq + (1-a) \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} a \frac{\partial}{\partial \text{vech} \hat{\Omega}} \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \hat{\theta} \\
&= \left[\int \cdots \int p(q) p'(q) f(q) dq + (1-a) \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} a [I \otimes \hat{\theta}] \frac{\partial \text{vec} \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix}}{\partial \text{vech} \hat{\Omega}} \\
&= [\Psi(a)]^{-1} a [I \otimes \hat{\theta}] \Gamma
\end{aligned}$$

where Γ is defined by

$$\text{vec} \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} = \Gamma \text{vech} \hat{\Omega}$$

and

$$\Psi(a) = \left[\int \cdots \int p(q) p'(q) f(q) dq + (1-a) \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right].$$

Note that $\Psi(a)$ is positive definite for $a \leq 1$ and Γ is a matrix of 1's and 0's.

Thus, $\frac{\partial \hat{\theta}}{\partial \text{vech} \hat{\Omega}}$ is usually moving in the same direction as $\hat{\theta}$. We can evaluate

$$\begin{aligned}
\frac{\partial^2 \hat{\theta}}{\partial a \partial \text{vech} \hat{\Omega}} &= [\Psi(a)]^{-1} [I \otimes \hat{\theta}] \Gamma + a [\Psi(a)]^{-1} \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} [\Psi(a)]^{-1} [I \otimes \hat{\theta}] \Gamma \\
&= \left\{ a^{-1} I + [\Psi(a)]^{-1} \begin{pmatrix} \hat{\Omega} & 0 \\ 0 & 0 \end{pmatrix} \right\} \frac{\partial \hat{\theta}}{\partial \text{vech} \hat{\Omega}}.
\end{aligned}$$

By similar reasoning, we expect $\frac{\partial^2 \hat{\theta}}{\partial a \partial \text{vech} \hat{\Omega}}$ to move in the same direction as $\hat{\theta}$ and $\frac{\partial \hat{\theta}}{\partial \text{vech} \hat{\Omega}}$. We can write

$$\text{Cov} [\hat{\theta}(a)] = \left(\frac{\partial \hat{\theta}}{\partial \text{vech} \hat{\Omega}} \right)' \text{Cov} [\text{vech} \hat{\Omega}] \left(\frac{\partial \hat{\theta}}{\partial \text{vech} \hat{\Omega}} \right)$$

and

$$\begin{aligned}
D [\hat{\theta}] &= E [\hat{\theta} - \theta]' W_v [\hat{\theta} - \theta] \\
&= \text{tr} E [\hat{\theta} - \theta]' W_v [\hat{\theta} - \theta] \\
&= \text{tr} W_v \text{Cov} [\hat{\theta}(a)]
\end{aligned}$$

with

$$\frac{\partial D [\hat{\theta}]}{\partial \text{Cov} [\hat{\theta}(a)]} = W_v.$$

The derivative is, therefore,

$$\begin{aligned}
\frac{\partial}{\partial a} D [\widehat{\theta}(a)] &= [trW_v] \frac{\partial}{\partial a} Cov [\widehat{\theta}(a)] \\
&= [trW_v] \left(\frac{\partial^2 \widehat{\theta}}{\partial a \partial vech \widehat{\Omega}} \right)' D [vech \widehat{\Omega}] \left(\frac{\partial \widehat{\theta}}{\partial vech \widehat{\Omega}} \right) \\
&\quad + [trW_v] \left(\frac{\partial \widehat{\theta}}{\partial vech \widehat{\Omega}} \right)' D [vech \widehat{\Omega}] \left(\frac{\partial^2 \widehat{\theta}}{\partial a \partial vech \widehat{\Omega}} \right).
\end{aligned}$$

Note that $D [vech \widehat{\Omega}]$ does not depend on a . Also, since $\frac{\partial^2 \widehat{\theta}}{\partial a \partial vech \widehat{\Omega}}$ is equal to a positive definite matrix times $\frac{\partial \widehat{\theta}}{\partial vech \widehat{\Omega}}$ and $trW_v > 0$, $\frac{\partial}{\partial a} D [\widehat{\theta}(a)]$ is positive for $a > 0$. When $a = 0$, $\left(\frac{\partial \widehat{\theta}}{\partial vech \widehat{\Omega}} \right) = 0$, implying that $\frac{\partial}{\partial a} D [\widehat{\theta}(a)] = 0$.

We want to pick the value of a that minimizes equation (32); i.e., the value where

$$-\frac{\partial}{\partial a} D [\widehat{\theta}(a)] = \frac{\partial}{\partial a} B [\widehat{\theta}(a)]. \quad (33)$$

Note that, at $a = 1$, $\frac{\partial}{\partial a} D [\widehat{\theta}(a)] > 0$ and $\frac{\partial}{\partial a} B [\widehat{\theta}(a)] = 0$. Also, at $a = 0$, $\frac{\partial}{\partial a} D [\widehat{\theta}(a)] = 0$ and $\frac{\partial}{\partial a} B [\widehat{\theta}(a)] < 0$. Thus, the optimal value of a must be $0 < a < 1$.

We can operationalize this by using estimates from the data. In particular, given $\widehat{\Omega}$, we can, in some cases, analytically construct an estimate of $D [vech \widehat{\Omega}]$ and, in all cases, simulate an estimate of $D [vech \widehat{\Omega}]$. Next, given $D [vech \widehat{\Omega}]$, we can simulate $\frac{\partial}{\partial a} D [\widehat{\theta}(a)]$ and $\frac{\partial}{\partial a} B [\widehat{\theta}(a)]$ and thus find the value of a that solves equation (33).

Appendix D: Lower Bound Estimate of the Covariance Matrix (sketch)

$$\text{plim} \widetilde{m}(\theta, \Omega_{12}) = \text{plim} \left\{ \frac{1}{n} \sum_i X_{i2} (y_i - X'_{i1} \theta) + \begin{pmatrix} \Omega_{12} \\ 0 \end{pmatrix} \beta \right\} = 0.$$

$$\begin{aligned}
\text{plim} \widetilde{m}(\theta, A) &= \text{plim} \left\{ \frac{1}{n} \sum_i X_{i2} (y_i - X'_{i1} \theta) + \begin{pmatrix} A \\ 0 \end{pmatrix} \beta \right\} \\
&= \begin{pmatrix} A - \Omega_{12} \\ 0 \end{pmatrix} \beta.
\end{aligned}$$

$$\begin{aligned} \frac{1}{n} \sum_i X_{i2} y_i - \frac{1}{n} \sum_i X_{i2} X'_{i1} \theta + \begin{pmatrix} A \\ 0 \end{pmatrix} \beta &= 0 \\ \left[\frac{1}{n} \sum_i X_{i2} X'_{i1} - \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \right] \theta &= \frac{1}{n} \sum_i X_{i2} y_i \\ \hat{\theta}(L) &= \left[\frac{1}{n} \sum_i X_{i2} X'_{i1} - \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \left[\frac{1}{n} \sum_i X_{i2} y_i \right] \end{aligned}$$

$$\begin{aligned} \text{plim} \left[\frac{1}{n} \sum_i X_{i2} X'_{i1} - \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \right] \text{plim} \hat{\theta}(A) &= \text{plim} \frac{1}{n} \sum_i X_{i2} y_i \\ \left[P + \begin{pmatrix} \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} A & 0 \\ 0 & 0 \end{pmatrix} \right] \text{plim} \hat{\theta}(A) &= P\theta \\ P^{-1} \begin{pmatrix} A - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \text{plim} \hat{\theta}(L) &= \text{plim} \hat{\theta}(L) - \theta \\ \Omega_{12} - L \text{ is psd} &\Rightarrow P^{-1} \begin{pmatrix} L - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \text{ is nsd} \end{aligned} \quad (34)$$

Write equation (34) as

$$P^{-1} \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \text{plim} \hat{\theta}(aL) = \text{plim} \hat{\theta}(aL) - \theta \quad (35)$$

where $0 \leq a \leq 1$ is a scalar and L is a psd lower bound estimate of Ω_{12} . When $a = 0$, $\hat{\theta}(aL)$ is the uncorrected IV estimate, and when $a = 1$, $\hat{\theta}(aL)$ is the IV estimate using L as a correction factor. Consider taking the full derivative of equation (35) with respect to a and $\text{plim} \hat{\theta}$:

$$\begin{aligned} P^{-1} \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \text{plim} \hat{\theta} da + P^{-1} \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} d\text{plim} \hat{\theta} &= d\text{plim} \hat{\theta} \\ \left[I - P^{-1} \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \right] d\text{plim} \hat{\theta} &= P^{-1} \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \text{plim} \hat{\theta} da \\ \frac{d\text{plim} \hat{\theta}}{da} &= \left[I - P^{-1} \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} P^{-1} \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \text{plim} \hat{\theta} \\ &= \left[P - \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \text{plim} \hat{\theta} \end{aligned} \quad (36)$$

which, in the leading case, has the same sign as $\text{plim} \hat{\theta}$ because $aL - \Omega_{12}$ is nsd and L is psd.

Consider the quadratic form,

$$\left[\text{plim}\hat{\theta} - \theta \right]' W \left[\text{plim}\hat{\theta} - \theta \right],$$

and then differentiate with respect to a :

$$\begin{aligned} & \frac{\partial}{\partial a} \left[\text{plim}\hat{\theta} - \theta \right]' W \left[\text{plim}\hat{\theta} - \theta \right] \\ = & \left[\text{plim}\hat{\theta} - \theta \right]' W \left[P - \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \text{plim}\hat{\theta} \\ & + \text{plim}\tilde{\theta}' \begin{pmatrix} L' & 0 \\ 0 & 0 \end{pmatrix} \left[P' - \begin{pmatrix} aL' - \Omega'_{12} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} W \left[\text{plim}\hat{\theta} - \theta \right]. \end{aligned}$$

Note that

$$\begin{aligned} & \left[\text{plim}\hat{\theta} - \theta \right]' W \left[P - \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \text{plim}\hat{\theta} \tag{37} \\ = & \text{plim}\tilde{\theta}' \begin{pmatrix} aL' - \Omega'_{12} & 0 \\ 0 & 0 \end{pmatrix} P^{-1} W \left[P - \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \right]^{-1} \begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix} \text{plim}\hat{\theta} \end{aligned}$$

using equation (35). Note that $\begin{pmatrix} aL' - \Omega'_{12} & 0 \\ 0 & 0 \end{pmatrix}$ is nsd, and P^{-1} , W , $\left[P - \begin{pmatrix} aL - \Omega_{12} & 0 \\ 0 & 0 \end{pmatrix} \right]$, and $\begin{pmatrix} L & 0 \\ 0 & 0 \end{pmatrix}$ are psd. This implies that equation (37) plus its transpose is always negative.