Instrumental Variables and the Problem of Endogeneity

September 15, 2015

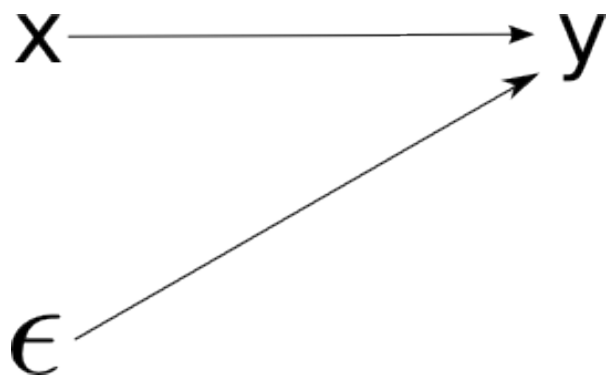
Exogeneity: Important Assumption of OLS

In a standard OLS framework,

$$\mathbf{y} = \mathbf{x}\beta + \epsilon \tag{1}$$

and for unbiasedness we need

$$E[\mathbf{x}'\epsilon] = 0_{[K\times 1]} \tag{2}$$



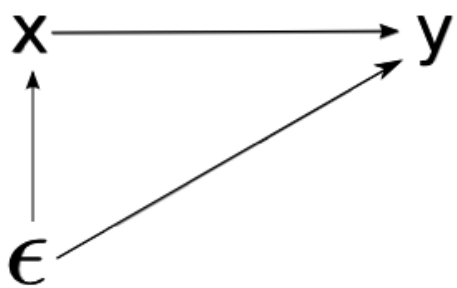
Endogeneity Defined

In a standard OLS framework,

$$\mathbf{y} = \mathbf{x}\beta + \epsilon \tag{3}$$

b is biased since $E[\mathbf{b}] \neq \beta$. This happens because

$$E[\mathbf{x}'\epsilon] \neq 0 \tag{4}$$



The Problem

- If ϵ imparts some effect on **x**, then we can't disentangle the direct impact of **x** on **y** (what we really want to know) with the indirect effect of ϵ on **y** via **x**.

Examples

  - Simultaneity: Aids funding in Africa and Aids incidence
  - Missing Variable Bias: Wage equation

- With biased estimates of β our model gives poor policy guidance.

Approaches

- Proxy Variables
  - Need to find a proxy correlated with the missing variable (or problematic part of the error).
  - Difficulties interpreting results, since the scale of estimated coefficient not necessarily informative for β
- Lagging **x** (ad hoc)
- Instrumental Variables
- Control Function

The Instrumental Variable Approach: Setup

Let $\mathbf{x}$ be an $N \times K$ matrix of the following form:

$$\mathbf{x} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1,K-1} & x_{1,K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i2} & \dots & x_{i,K-1} & x_{i,K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N2} & \dots & x_{N,K-1} & x_{N,K} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_2 & \dots & \mathbf{x}_{K-1} & \mathbf{x}_K \end{bmatrix} \tag{5}$$

- We believe that the K^{th} column is endogenous (it could be any column, 2 through K).
- Columns 1 to K-1 are exogenous ($\mathbf{x}_{-K} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_2 & \dots & \mathbf{x}_{K-1} \end{bmatrix}$)

Pick an Instrumental Variable (IV)

Find an instrumental Variable ($\mathbf{z}_K$) having the following properties:

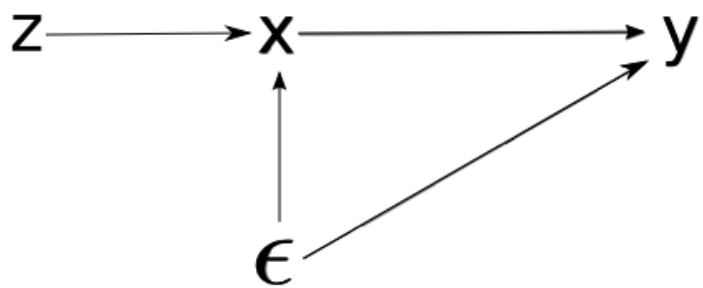
1. $E(\mathbf{z}'_K \epsilon) = 0$
2. Relevant:

$$\mathbf{x}_K = \delta_0 + \delta_1 \mathbf{x}_1 + \ldots + \delta_{K-1} \mathbf{x}_{K-1} + \theta_K \mathbf{z}_K + \mathbf{r}_K \qquad (6)$$

   Estimate the relevancy equation and test
   - H_0 : $\mathbf{z}_K$ not relevant: $\theta_K = 0$
   - H_1 : $\mathbf{z}_K$ is relevant: $\theta_K \neq 0$
3. $\mathbf{z_k}$ does not directly impact $\mathbf{y}$

Endogeneity and Instrumental Variables: An illustration

These two conditions ensure the causality is running in this direction:



In particular:

1. $\mathbf{z}_k$ is uncorrelated with ϵ
2. $\mathbf{z}_k$ has no direct impact on $\mathbf{y}$ (sometimes called the exclusion restriction)

Examples

- Demand estimation: price is endogenously determined by demand shifts (and random shocks to demand)
  - Look for something correlated (+/-) with price but not directly related to quantity demanded or error.
  - Things only impacting supply have this property (for food commodities: weather)

- Returns to education: educational attainment likely correlated with errors in wage equation
  - Look for something correlated with educational attainment but not related to wage or error
    - Distance from school/university
    - Month of birth

Implementing the IV Approach: Step 1

Define the matrix $\mathbf{z}$ as

$$\mathbf{z} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1,K-1} & z_{1,K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i2} & \dots & x_{i,K-1} & z_{i,K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N2} & \dots & x_{N,K-1} & z_{N,K} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_2 & \dots & \mathbf{x}_{K-1} & \mathbf{z}_K \end{bmatrix} \tag{7}$$

How do we use the IV (found in $\mathbf{z}$) to estimate β

Idea: Define $\mathbf{b}^{iv}$ as $(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$

Substitute the "Relevancy equation" into our original estimating equation:

$$\begin{aligned}
\mathbf{y} =& \beta_1 + \beta_2\mathbf{x}_2 + \ldots + \beta_{K-1}\mathbf{x}_{K-1} \\
&+ \beta_K(\delta_1 + \delta_2\mathbf{x}_2 + \ldots + \delta_{K-1}\mathbf{x}_{K-1} + \theta_K\mathbf{z}_K + \mathbf{r}) + \epsilon \\
=& (\beta_1 + \beta_K\delta_1) + (\beta_2 + \beta_K\delta_2)\mathbf{x}_2 + \ldots + (\beta_{K-1} + \beta_K\delta_{K-1})\mathbf{x}_{K-1} \\
&+ (\beta_K\theta_K)\mathbf{z}_k + (\beta_K\mathbf{r}_k + \epsilon) \\
=& \alpha_1 + \alpha_2\mathbf{x}_2 + \ldots + \alpha_{K-1}\mathbf{x}_{K-1} + \alpha_K\mathbf{z}_k + \mathbf{v}
\end{aligned}$$

If we run this regression, we obtain $\mathbf{a} = (\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{y}$

So merely substitute our instrument in for $\mathbf{x}_K$ and recovering parameters $\mathbf{a}$ will give you estimates where:

- $\alpha_k \neq \beta_k$ for *every* parameter you estimate, not just the endogenous one, β_K
- The variance/covariance matrix of the errors ($\mathbf{v}$) is not $N(0, \sigma^2\mathbf{I})$
- $E[\mathbf{a}] \neq \beta$, so this is not a good IV estimator.
- Given an estimate for $\mathbf{a}$, we can't solve for the K estimates for β because we have K equations in $2 \times K$ unknowns.

Implementing the IV Approach: Step 2

Given an estimate using the instrumental variables approach ($\mathbf{b}^{iv}$), we can define the predicted model error as

$$\mathbf{e}^{iv} = y - \mathbf{x}\mathbf{b}^{iv} \tag{8}$$

This error must have the property that[1]

$$E[\mathbf{z}'\epsilon] \Rightarrow \mathbf{z}'\mathbf{e}^{iv} = 0 \tag{9}$$

[1] So long as it can be shown that b^{iv} is a consistent estimate of β

Implementing the IV Approach: Step 2

Simplify

$$
\begin{aligned}
0 =& \mathbf{z}'(\mathbf{y} - \mathbf{x}\mathbf{b}^{iv}) && (10)\\
=& \mathbf{z}'\mathbf{y} - \mathbf{z}'\mathbf{x}\mathbf{b}^{iv} && (11)\\
\Rightarrow& \mathbf{b}^{iv} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y} && (12)
\end{aligned}
$$

Implementing the IV Approach: Step 2a (alternative way)

An equivalent way to think about IV regression:

1. Run the following regression (relevancy equation):

$$\mathbf{x}_K = \delta_1 + \delta_2 \mathbf{x}_2 + \ldots + \delta_{K-1} \mathbf{x}_{K-1} + \theta_K \mathbf{z}_K + \mathbf{r} \tag{13}$$

2. With parameter estimates ($\mathbf{d}$ and t_k) calculate $\hat{\mathbf{x}}_K$:

$$\hat{\mathbf{x}}_K = d_1 + d_2 \mathbf{x}_2 + \ldots + d_{K-1} \mathbf{x}_{K-1} + t_K \mathbf{z}_K \tag{14}$$

Implementing the IV Approach: Step 2a (alternative way),cont.

Defining

$$\hat{\mathbf{x}} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1,K-1} & \hat{x}_{1,K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i2} & \dots & x_{i,K-1} & \hat{x}_{i,K} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{N2} & \dots & x_{N,K-1} & \hat{x}_{N,K} \end{bmatrix} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_2 & \dots & \mathbf{x}_{K-1} & \hat{\mathbf{x}}_K \end{bmatrix} \tag{15}$$

We can write the IV estimator as

$$\mathbf{b}^{iv} = (\hat{\mathbf{x}}'\hat{\mathbf{x}})^{-1}\hat{\mathbf{x}}'\mathbf{y} \tag{16}$$

Intuition of IV estimators

Intuition

$\hat{\mathbf{x}}$ contains only exogenous information. By using the predicted value we take the part of $\mathbf{x}_k$ not correlated with ϵ and use it to estimate β.

It can also be shown that for 1 IV and 1 endogenous variable:

$$\mathbf{b}^{iv} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y} = (\hat{\mathbf{x}}'\hat{\mathbf{x}})^{-1}\hat{\mathbf{x}}'\mathbf{y} \tag{17}$$

We can identify the β parameters using IV Regression

Assumptions

1. $\mathbf{z}$ is not correlated with the "true" model error, or $E[\mathbf{z}'\epsilon] = 0$ (we can't test this)
2. $\mathbf{z}$ imparts an effect on $\mathbf{y}$ only via x, not directly (difficult to test)
3. It is relevant and strong (we can test this)
4. rank($\mathbf{z}$)=K

If these conditions hold, then we have consistent estimates for β. In this case, since we have 1 IV and 1 endogenous variable, we say the model is exactly identified.

b^{iv} is LATE

The instrumental variable estimator gives us estimates of locally averaged treatment (causal effect) effects (LATE).
This means it doesn't help identify behavioral changes that may occur for non-treatment groups.

- Consider using an earthquake event (=1,0) to instrument for price of agricultrual commodities. The $\mathbf{b}^{iv}$ we get gives us the correct estimate of the estimate on price (β_p) telling us
  - about those affected by price experiencing the earthquake
  - very little about those not affected by the earthquake
- These sorts of problems loom larger when your instrument is a dummy variable.

So long as your treatment group is representative of everyone, no issues.

Relevancy Test

Run the relevancy regression:

$$\mathbf{x}_K = \delta_1 + \delta_2 \mathbf{x}_2 + \ldots + \delta_{K-1} \mathbf{x}_{K-1} + \theta_K \mathbf{z}_K + \mathbf{r} \tag{18}$$

and test:

Relevancy Test

$H_0 : \theta_k = 0$: The instrument $\mathbf{z}_K$ is not relevant
$H_1 : \theta_k \neq 0$: The instrument $\mathbf{z}_K$ is relevant

Furthermore, the F-test with one degree of freedom can tell us about "strong" instruments. Rejecting the NULL hypothesis, $\mathbf{z}_K$ is strong if the F-statistic exceeds 10.

Testing for the endogeneity of $\mathbf{x}_K$

Another standard test is to see if, in fact $\mathbf{x}_K$ is endogenous. This test proceeds from the observation that if $\mathbf{x}_K$ is endogenous and we have a strong and relevant IV meeting the assumptions above, then the OLS estimate is biased and inconsistent whereas the IV estimate is consistent:

$$E[\mathbf{b}^{OLS}] \neq E[\mathbf{b}^{IV}] = \beta \tag{19}$$

Consequently, in the case of an endogenous $\mathbf{x}_K$, we can test for a meaningful difference between the two sets of estimates.

Hausman Test

The Hausman test is widely used for testing differences in parameter estimates. In an IV setting, this is called the Hausman-Wu test, having

Hausman-Wu Endogeneity Test

$H_0 : \mathbf{b}^{IV} - \mathbf{b}^{OLS} = 0 : \mathbf{x}_K$ is exogenous
$H_1 : \mathbf{b}^{IV} - \mathbf{b}^{OLS} \neq 0 : \mathbf{x}_K$ is endogenous

Where the test statistic is distributed F with 1 degree of freedom.

Implementing the Hausman-Wu Test

1. Run the Relevancy Equation:

$$\mathbf{x}_K = \delta_1 + \delta_2 \mathbf{x}_2 + \ldots + \delta_{K-1} \mathbf{x}_{K-1} + \theta_K \mathbf{z}_K + \mathbf{r} \tag{20}$$

2. Recover the predicted residuals from this regression ($\hat{\mathbf{r}}$)
3. Run this regression:

$$\mathbf{y} = \mathbf{x}\tau + \mu\hat{\mathbf{r}} + u \tag{21}$$

Hausman-Wu Endogeneity Test

$H_0 : \mathbf{b}^{IV} - \mathbf{b}^{OLS} = 0 : \mu = 0 :$ $\mathbf{x}_K$ is exogenous
$H_1 : \mathbf{b}^{IV} - \mathbf{b}^{OLS} \neq 0 : \mu \neq 0:$ $\mathbf{x}_K$ is endogenous

Where the test statistic is distributed F with 1 degree of freedom.

Intuition of the Test

- The residual $\hat{\mathbf{r}}$ should only include the endogenous part of $\mathbf{x}_K$ (if any exists), since we have controlled for all exogenous information at our disposal ($\mathbf{x}_{-K}$ and $\mathbf{z}_K$).
- If this endogenous part of $\mathbf{x}_K$ is useful for predicting $\mathbf{y}$ after we control for our full set of original regressors ($\mathbf{x}$), then this provides evidence of significance differences in our regressors because there is a part of $\mathbf{x}_K$ that is correlated with $\mathbf{y}$ via the error term.

Variance/Covariance Matrix of Parameters

- It can be shown that Var($\mathbf{b}^{IV}$) is

$$E[Var(\mathbf{b}^{IV})] = \sigma^2 \left((\mathbf{z}'\mathbf{x})^{-\mathbf{1}} \mathbf{z}'\mathbf{z} (\mathbf{z}'\mathbf{x})^{-\mathbf{1}\prime} \right) \tag{22}$$

- With the robust version:

$$E[Var(\mathbf{b}^{IV})]_{Robust} = (\mathbf{z}'\mathbf{x})^{-\mathbf{1}} \mathbf{z}'\mathbf{V}\mathbf{z} (\mathbf{z}'\mathbf{x})^{-\mathbf{1}\prime} \tag{23}$$

More IV's than endogenous variables

Now consider a case where for our population regression, column K continues to be suspected as endogenous and we want to have M instrumental Variables rather than only 1. Redefine $\mathbf{z}$ as:

$$\mathbf{z} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1,K-1} & z_{1,1} & z_{1,2} & \dots & z_{1,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{i2} & \dots & x_{i,K-1} & z_{i,1} & z_{i,2} & \dots & z_{i,M} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\ 1 & x_{N2} & \dots & x_{N,K-1} & z_{N,1} & z_{N,2} & \dots & z_{N,M} \end{bmatrix} \tag{24}$$

$$= \begin{bmatrix} \mathbf{1} & \mathbf{x}_2 & \dots & \mathbf{x}_{K-1} & \mathbf{z}_1 & \mathbf{z}_2 & \dots & \mathbf{z}_M \end{bmatrix} \tag{25}$$

Some Terminology

Stata uses a slightly different terminology:

	$\mathbf{x}_{-K}$	$\mathbf{z}_k$	$\mathbf{x}_k$
My Terminology	Exogenous Independent Variables	Instrumental Variable(s)	Endogenous Independent Variable(s)
Stata's Terminology	Instruments	Instruments	Instrumented

Rationale for more than one IV

- We put alot of hopes on our lone IV ($\mathbf{z}_K$)
  - It is uncorrelated with ϵ
  - It is correlated with $\mathbf{y}$ only via $\mathbf{x}_K$
  - It is relevant. While relevancy is a good thing, it doesn't ensure the "best" IV
- With this limitation in mind, extend the IV model to include more instruments for $\mathbf{x}_K$

Rationale for more than one IV: The dark side

- But including more IV's risks violating the exogeneity assumption:

$$E[\mathbf{z}'\epsilon] \neq 0 \tag{26}$$

- Or, we might have redundant or near redundant information in $\mathbf{z}$.

Combined, or taken separately this complicates pinning down a unique and consistent estimate for β.

Deriving the IV Estimator

Using the Method of Moments approach outlined above, we need to find an estimate for $\mathbf{b}^{IV}$ satisfying

$$E[\mathbf{z}'\epsilon] = 0 \tag{27}$$

If β^{IV} is a consistent estimate for β, then this condition becomes

$$0 = \mathbf{z}'e \tag{28}$$

Deriving the IV Estimator

Using the Method of Moments approach outlined above, we need to find an estimate for $\mathbf{b}^{IV}$ satisfying

$$E[\mathbf{z}'\epsilon] = 0 \tag{27}$$

If β^{IV} is a consistent estimate for β, then this condition becomes

$$0 = \mathbf{z}'e \tag{28}$$

$$0 = \mathbf{z}'(\mathbf{y} - \mathbf{x}\mathbf{b}^{\mathbf{IV}}) \tag{29}$$

$$0 = \mathbf{z}'\mathbf{y} - \mathbf{z}'\mathbf{x}\mathbf{b}^{\mathbf{IV}} \tag{30}$$

Rearranging yields $\mathbf{b}^{IV} = (\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}$ as before....Or does it. Check dimensionality of $(\mathbf{z}'\mathbf{x})^{-1}\mathbf{z}'\mathbf{y}$ versus what we expect the dimensionality of b^{IV} to be.

Deriving the Estimator, 2SLS

As in the case above, from the relevancy equation calculate predicted $\mathbf{x}_K$ as a function of our exogenous variables:

$$\hat{\mathbf{x}}_K = d_0 + d_2\mathbf{x_2} + d_3\mathbf{x_3} + \ldots + d_{K-1}\mathbf{x_{K-1}} + t_1\mathbf{z_1} + \ldots + t_M\mathbf{z_M} \quad (31)$$

by running an OLS regression. Denoting $\hat{\mathbf{x}} = \begin{bmatrix} 1 & \mathbf{x_2} & \ldots & \mathbf{x_{K-1}} & \hat{\mathbf{x}}_K \end{bmatrix}$, the two stage least squares estimator (2SLS) is

$$\hat{\beta}^{2SLS} = (\hat{\mathbf{x}}'\hat{\mathbf{x}})^{-1}\hat{\mathbf{x}}'\mathbf{y} \quad (32)$$

Note: In a more general setting where there are more than 1 endogenous variables, define $\hat{\mathbf{x}}$ as

$$\hat{\mathbf{x}} = \mathbf{z}(\mathbf{z}'\mathbf{z})^{-1}\mathbf{z}'\mathbf{x} \quad (33)$$

Deriving the Estimator, GMM

We have more equations than unknowns making it unlikely that $\mathbf{z}'\mathbf{e} = 0$ for every column of $\mathbf{z}$. For estimation purposes $\mathbf{b}^{IV}$ is found by minimizing

$$\min_{b^{IV}} \frac{\mathbf{e}'\mathbf{z}\mathbf{W}\mathbf{z}'\mathbf{e}}{N} \tag{34}$$

which is a scalar value. $\mathbf{W}$ is a weighting matrix.
Typically $\mathbf{W}$ contains similar information to $\mathbf{V}$, the matrix we used to correct for robust standard errors in the OLS chapter. Setting $\mathbf{W} = \mathbf{I}$ restricts the GMM estimate for b^{IV} to be equal to the 2SLS estimate.

Deriving the Estimator, GMM

While it is almost always possible to find a $\mathbf{b}^{IV}$ that minimizes this condition, it does not impose the orthogonality condition for each column of $\mathbf{z}$. Thus there is the possibility of *overidentification*.

GMM versus 2SLS, which to use?

In addition to GMM and 2SLS, there are additional methods one could use to estimate our estimate for β in an IV framework. Which to use?

Stata Name	Description	Notes
2SLS	Two Stage Least Squares	Useful for understanding classical IV regression
GMM	Generalized Method of Moments	Probably most useful method for most settings. Must understand implications for various choices of **W** you might use. Defaults not bad.
LIML	Limited Information Maximum Likelihood	Uses ML methods, has the best small sample properties
3SLS	Three Stage Least Squares	More efficient than 2SLS but not used as often in an IV framework because not robust to specification error

In most cases you want to use GMM with default weighting matrix unless you have good reasons to do otherwise.

More IV's than endogenous columns of **x**: Steps

1. Contemplate endogeneity problem for each regressor in **x**
2. If some elements of **x** could be endogenous, contemplate instrumental variables, noting that the number of instruments must be greater than or equal to the number of endogenous variables.
3. Test for the relevance of your instruments
4. Test for overidentification
5. Test for the endogeneity of your suspect columns of **x**

The Overidentification Test (Sargon Test): Manual Method

Steps for 2SLS:

1. Recover predicted errors from IV regression ($\mathbf{e}^{iv} = \mathbf{y} - \mathbf{x}\mathbf{b}^{iv}$)
2. Regress the predicted errors $\mathbf{e}^{iv}$ on all exogenous regressors and instruments ($\mathbf{x}_{-K}$ and the instrumental variables $\mathbf{z}_1$ through $\mathbf{z}_M$), which we defined previously as $\mathbf{z}$

$$\mathbf{e}^{iv} = \mathbf{z}\mu + \psi \tag{35}$$

3. Conduct the joint test of $R^2 \times N$ distributed $\chi^2(M-1)$:
   - $H_0 = \mu = 0 \implies (\mathbf{z}'\mathbf{e}^{\mathbf{iv}} = 0)$
   - $H_1 = \mu \neq 0 \implies (\mathbf{z}'\mathbf{e}^{\mathbf{iv}} \neq 0)$

Note: The manual steps outlined above are not recommended, instead use the stata command overid.

Variance/Covariance Matrix for $\mathbf{b}_{iv}$ (standard errors)

Standard errors in the multiple IV 2SLS framework can be calculated as above, except they are inefficient (and will not match what stata reports).

Why?

- Two stage least squares involves estimation of $\hat{\mathbf{x}}_k$ used to estimate $\mathbf{b}_{iv}$
- The correct standard errors take into acccount that $\hat{\mathbf{x}}_k$ is a random variable.

Bottom Line: Let stata calculate standard errors, and don't use manually calculated se's.