IV worksheet

Ryan T. Godwin

Inconsistency of the LS estimator under missing variables

1. What is the definition of inconsistency?

   When an estimator fails to give the right answer when the sample size is infinite. Mathematically, an estimator is weakly consistent when:

   $$\lim_{n\to\infty}\left\{\Pr\cdot\left[\left|\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right|<\boldsymbol{\epsilon}\right]\right\}=1$$

   for all $\epsilon>0$, and is strongly consistent when:

   $$\begin{aligned}\text{Bias}\left(\widehat{\boldsymbol{\theta}}_n\right)&\to\mathbf{0};\ \text{ as } n\to\infty,\\ V\left(\widehat{\boldsymbol{\theta}}_n\right)&\to 0;\ \text{ as } n\to\infty.\end{aligned}$$

2. What are the two assumptions required to prove that the LS estimator is consistent?

   We need that $\text{plim}\left[n^{-1}X'\epsilon\right]=\mathbf{0}$ (which is an extension of A.5) and that $\text{plim}\left[n^{-1}X'X\right]=Q$ (which is an extension of A.2).

3. What does it mean if there is a missing variable?

   If there is a missing variable that is both correlated to X and $\boldsymbol{y}$, then LS is biased and inconsistent.

4. Prove that the LS estimator is inconsistent when $\text{plim}\left(\frac{1}{n}X'\boldsymbol{\epsilon}\right)\neq\mathbf{0}$.

   $$\begin{aligned}\boldsymbol{b}=\left(X'X\right)^{-1}X'\boldsymbol{y}&=\left(X'X\right)^{-1}X'(X\boldsymbol{\beta}+\boldsymbol{\epsilon})\\&=\boldsymbol{\beta}+\left(X'X\right)X'\boldsymbol{\epsilon}\\&=\boldsymbol{\beta}+\left[\frac{1}{n}\left(X'X\right)\right]^{-1}\left[\frac{1}{n}X'\boldsymbol{\epsilon}\right]\end{aligned}$$

   If $\text{plim}\left(\frac{1}{n}X'\boldsymbol{\epsilon}\right)\neq\mathbf{0}$, say $\text{plim}\left(\frac{1}{n}X'\boldsymbol{\epsilon}\right)=\boldsymbol{\gamma}$, then:

   $$\text{plim}(\boldsymbol{b})=\boldsymbol{\beta}+Q^{-1}\boldsymbol{\gamma}\neq\boldsymbol{\beta}$$

Instrumental variables

5. What two conditions must be satisfied in order for a variable to qualify as an "instrument"?

   (i) "Relevance" which means that the instrument needs to be correlated with the endogenous regressor; and (ii) the "exclusion restriction" which means the instrument needs to be uncorrelated with the missing variable.

6. Draw a diagram with y, X, missing variable M, and IV Z. Draw arrows between variables that are correlated.

7. What is the formula for the IV estimator?

   The generalized IV estimator (allowing for more instruments than endogenous regressors) is:

$$\boldsymbol{b}_{IV} = \left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}X'Z\left(Z'Z\right)^{-1}Z'\boldsymbol{y}$$

   or:

$$\boldsymbol{b}_{IV} = \left[X'P_Z X\right]^{-1}X'P_Z\boldsymbol{y}$$

   and the simple IV estimator is:

$$\boldsymbol{b}_{IV} = \left(Z'X\right)^{-1}Z'\boldsymbol{y}$$

8. Prove that the IV estimator is consistent.

   Full rank of the instrument matrix, the relevancy of the IV, and the exclusion restriction imply respectively that:

$$\begin{aligned}
\text{plim}\left(\frac{1}{n}Z'Z\right) &= Q_{ZZ}\\
\text{plim}\left(\frac{1}{n}Z'X\right) &= Q_{ZX}\\
\text{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right) &= \boldsymbol{0}
\end{aligned}$$

   Then, the IV estimator is consistent:

$$\begin{aligned}
\boldsymbol{b}_{IV} &= \left(Z'X\right)^{-1}Z'\boldsymbol{y} = \left(Z'X\right)^{-1}Z'(X\boldsymbol{\beta}+\boldsymbol{\epsilon})\\
&= \left(Z'X\right)^{-1}Z'X\boldsymbol{\beta} + \left(Z'X\right)^{-1}Z'\boldsymbol{\epsilon}\\
&= \boldsymbol{\beta} + \left(Z'X\right)^{-1}Z'\boldsymbol{\epsilon}\\
&= \boldsymbol{\beta} + \left(\frac{1}{n}Z'X\right)^{-1}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right)\\
\text{plim}\left(\boldsymbol{b}_{IV}\right) &= \boldsymbol{\beta} + \left[\text{plim}\left(\frac{1}{n}Z'X\right)\right]^{-1}\text{plim}\left(\frac{1}{n}Z'\boldsymbol{\epsilon}\right)\\
&= \boldsymbol{\beta} + Q_{ZX}^{-1}\boldsymbol{0} = \boldsymbol{\beta}
\end{aligned}$$

2SLS

9. What are the two stages of 2SLS?

   In the first stage, we regress each column of X on Z using LS, and get $\hat{X}$. That is, we get $\hat{X} = P_Z X$. In the second stage, we estimate the model: $y = \hat{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = P_Z X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, using LS.

10. Derive the *generalized* (over-identified) IV estimator using the 2SLS method.

    The LS estimator is applied to $y = \hat{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ where $\hat{X} = Z\left(Z'Z\right)^{-1}Z'X$. So:

$$
\begin{aligned}
\boldsymbol{b}_{IV} &= (\hat{X}'\hat{X})^{-1}\hat{X}'\boldsymbol{y} \\
&= \left[X'Z(Z'Z)^{-1}Z'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}X'Z(Z'Z)^{-1}Z'\boldsymbol{y} \\
&= \left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}X'Z\left(Z'Z\right)^{-1}Z'\boldsymbol{y}
\end{aligned}
$$

11. Show that the generalized estimator collapses to the simple IV estimator when the number of instruments equals the number of "endogenous" variables (an X variable is said to be endogenous if it is correlated to a missing variable that impacts y).

    If the dimensions of Z and X match then the inverse in the square brackets above becomes:

$$
\left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1} = (Z'X)^{-1}(Z'Z)(X'Z)^{-1}
$$

    and the IV estimator becomes:

$$
\begin{aligned}
\boldsymbol{b}_{IV} &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z\left(Z'Z\right)^{-1}Z'\boldsymbol{y} \\
&= (Z'X)^{-1}(Z'Z)\left(Z'Z\right)^{-1}Z'\boldsymbol{y} \\
&= (Z'X)^{-1}Z'\boldsymbol{y}
\end{aligned}
$$

Properties and testing

12. Use the asymptotic variance of the IV estimator to see that the higher the correlation between Z and X, the better.

    The asymptotic distribution of the IV estimator is:

$$
\sqrt{n}\left(\boldsymbol{b_{IV}} - \boldsymbol{\beta}\right) \xrightarrow{d} N\left[\mathbf{0}, \sigma^2 Q_{ZX}^{-1} Q_{ZZ} Q_{XZ}^{-1}\right]
$$

    The stronger the correlation between Z and X, the "larger" will be Q_{ZX}, and the smaller will be the asymptotic variance.

13. What are *weak* instruments?

    Weak instruments are when the instruments are weakly correlated to the X. The IV estimator can be worse than the LS estimator.

14. Describe the Hausman test to see if IV is needed.

    The null is:

$$
H_0 : \text{plim}\left(\frac{1}{n}X'\epsilon\right) = \mathbf{0}
$$

    Both the LS and IV estimators are calculated. Under the null, both estimators are consistent and so there should be little difference between the two. If there is a big enough difference, we reject the null. The formal test statistic is:

$$
H = (\boldsymbol{b}_{IV} - \boldsymbol{b})'\left[\hat{V}\left(\boldsymbol{b}_{IV}\right) - \hat{V}(\boldsymbol{b})\right]^{-1}(\boldsymbol{b}_{IV} - \boldsymbol{b})
$$

    which is $H \xrightarrow{d} \chi_J^2$ if H_0 is true, and where J are the number of endogenous regressors (usually one).

15. Explain why the exclusion restriction is difficult to test.

    The exclusion restriction says the instrument must be uncorrelated with the error term. Since the error term is unobservable, this restriction is difficult to test.