Econ 1123: Section 7

Elena Llaudet

November 1, 2010

Outline

Review - Based on Common Mistakes in PSet 7

Units, units, units

- Log-linear model: $ln(Y) = \beta_0 + \beta_1 X + u$
  What is the interpretation of β_1? A one unit change in X is associated with a 100 β_1 % change in Y (cp).
- Linear-log model: $Y = \beta_0 + \beta_1 ln(X) + u$
  What is the interpretation of β_1? A 1% change in X is associated with a change in Y of 0.01 β_1 (cp).
- Log-log model: $ln(Y) = \beta_0 + \beta_1 ln(X) + u$
  What is the interpretation of β_1? A 1% change in X is associated with a β_1 % change in Y (cp).
- Model with a Binary Dep. Var.: $P(Y = 1) = \beta_0 + \beta_1 X + u$
  What is the interpretation of β_1? A one unit increase in X is associated with 100 * β_1 *percentage points* increase in the probability that Y equals 1 (ceteris paribus).

- In the non-linear models with binary dependent variables, do not confuse estimating $\hat{Z}$ with estimating $\hat{Y}$.

  If we define $\hat{Z} = \hat{\beta}_0 + \hat{\beta}_1 X$, then:

  LPM: $P(Y=1|X) = \beta_0 + \beta_1 X$ $\quad \hat{\mathbf{Y}} = ?$ $\hat{Y} = \hat{Z}$

  Probit: $P(Y=1|X) = \phi(\beta_0 + \beta_1 X)$ $\quad \hat{\mathbf{Y}} = ?$ $\hat{Y} = \phi(\hat{Z})$

  Logit: $P(Y=1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$ $\quad \hat{\mathbf{Y}} = ?$ $\hat{Y} = \frac{1}{1+e^{-(\hat{Z})}}$

  where ϕ is the cumulative probability distribution function of the standard normal.
- In all of the three cases, we cannot use R^2 as a measure of fit. Instead, we need to use *the Fraction Correctly Predicted* or *Pseudo-R^2*.

IV Analysis

- **What happens when X is correlated with u (due to, for example, omitted variable bias or simultaneous causality)?:** The estimated coefficient of X is biased.
- IV analysis is a way to solve this problem.
- IV analyses uses instruments (Z) to isolate the movements in X that are uncorrelated with u. Then, uses the isolated part of X that is not correlated with u, to estimate an unbiased coefficient for X.
- The instrument (Z) must satisfy two conditions, in order for it to be *valid*:
  (a) It must be relevant: $corr(Z, X) \neq 0$
  (b) It must be exogenous: $corr(Z, u) = 0$

Two Stage Least Squares Estimator (TSLS)

- **The first stage** decomposes X into two components: a problematic component that may be correlated with the regression error, and another problem-free component that is uncorrelated with the error.

  First stage: $X_i = \pi_0 + \pi_1 Z_i + v_i \rightarrow \hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$

- **The second stage** uses the second component (the problem-free component that is uncorrelated with the error) to estimate β_1.

  Second stage: $Y_i = \beta_0 + \beta_1 \hat{X}_i + u_i$

Special Case of TSLS

- When we have a single regressor (X) and a single instrument (Z), there is a simple formula for the TSLS estimator:

  $$\hat{\beta}_1^{TSLS} = S_{ZY}/S_{ZX}$$

  where S means sample covariance.

- See pages 428-429 for details.

Example of the Special Case

- Suppose that we are trying to asses the effect that private schools have on the likelihood to attend college.
- We know that our estimates would suffer from omitted variable bias if we simply run:

  College Attendance $= \beta_0 + \beta_1$ Private School + u_i

- A way to solve this problem is to use instrumental variables (if we can find a valid instrument).

- Now, suppose that we organize a voucher program. Among those students who apply for the program, we randomly provide to some of them a voucher to attend a private school.

  A few years go by and we observe the following summary statistics from the group of applicants:

	Voucher=0	Voucher=1
Went to a Private School	.09	.13
Attended College	.4	0.42

- **Would the assignment of vouchers be a valid instrument?** Probably. For Z to be a valid instrument:

  (a) Z must be relevant: $corr(Z, X) \neq 0$.

  In this case, it seems reasonable to assume that *corr*(Vouchers, Private School)$\neq$ 0 given that those students who received a voucher were more likely to attend a private school.
  (To determine whether the difference is statistically significant we would run the following regression: Private School $= \pi_0 + \pi_1$ Voucher, and see whether the coefficient on Voucher is statistically significant)

  (b) Z must be exogenous: $corr(Z, u) = 0$

  In this case, it seems reasonable to assume that *corr*(Vouchers, u)= 0 because the vouchers were randomly assigned.

- **What is the difference in probability of attending a private school between students who received the voucher and those who did not?**

  $\hat{P}$(Private School =1 | Voucher = 1)
  - $\hat{P}$(Private School =1 | Voucher = 0) = .04

- **Can you think of which regression we would have had to run on the data to estimate such a difference:?**

  Private School $= \pi_0 + \pi_1$ Voucher

  If we run the above regression, we would be estimating such a difference with π_1.

- **What is the difference in probability of attending college between students who received the voucher and those who did not?**

  $\hat{P}$(College =1 | Voucher = 1)
  - $\hat{P}$(College =1 | Voucher = 0) = .02

- **Can you think of which regression we would have had to run on the data to estimate such a difference:?**

  College $= \delta_0 + \delta_1$ Voucher

  If we run the above regression, we would be estimating such a difference with δ_1.

- Remember that in the case for a single regressor and a single instrument, the simple formula for the TSLS estimator is:

$$\hat{\beta}_1^{TSLS} = S_{ZY}/S_{ZX}$$

- **Can you think of a way for us to calculate $\hat{\beta}_1^{TSLS}$ with the summary statistics at hand?**

  (Hint: In models with a single regressor, the OLS estimator is equivalent to: S_{XY}/S_X^2):

$$\hat{\beta}_1^{TSLS} = S_{ZY}/S_{ZX} = \frac{S_{ZY}/S_Z^2}{S_{ZX}/S_Z^2} = \delta_1/\pi_1$$

  where δ_1 and π_1 are as defined in the previous slides.

General IV Regression Model

Setting:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \beta_{k+1} W_{1i} + ... + \beta_{k+r} W_{k+r} + u_i$$

where:

- Xs are endogenous regressors,
- Ws are included exogenous regressors,
- And, the instrumental variables are $Z_{1i}...Z_{mi}$.
- The coefficients are *overidentified* if m>k, *underidentified* if m<k, and *exactly identified* if m=k.
- Estimation of the IV regression model requires exact identification or overidentification.

Estimation:

1. For each X, we estimate a **fist-stage regression**:

   $X_{1i} = \pi_0 + \pi_1 Z_{1i} + ... + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + ... + \pi_{m+r} W_{ri} + v_i$
   $X_{2i} = \pi_0 + \pi_1 Z_{1i} + ... + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + ... + \pi_{m+r} W_{ri} + v_i$
   ...

   Using the estimated coefficients, we calculate all of the $\hat{X}$.

2. Then, we estimate **the second-stage regression**:
   $Y_i = \beta_0 + \beta_1 \hat{X}_{1i} + ... + \beta_k \hat{X}_{ki} + \beta_{k+1} W_{1i} + ... + \beta_{k+r} W_{k+r} + u_i$

Valid Instruments (Z):

- For IV analysis to generate consistent and asymptotically normal estimates, the instruments must be valid.
- **What are the two requirements that the instruments must comply with in order for them to be valid?**

  They must be both:

  (a) Relevant: Ideally, not only do they correlate with the Xs but they can explain a large portion of the variation of X. That is, they are not *weak* instruments.

  (b) Exogenous: They must be uncorrelated with the error term.

(a) Checking for Relevance

- **When can we check for the relevance of the instruments?** Always.
- We can test whether the instruments are relevant by performing an F-test on the coefficients of the instruments in the first-stage regression.
  $H_0 : \pi_1 = \pi_2 = ... = \pi_m = 0$
  H_A : At least one of them does not equal zero.
- Because we care about the instruments being strongly correlated with the Xs (not just simply correlated), simply rejecting the null is not good enough.
- Rule of thumb: We will say that our instruments are not weak if the first-stage F statistic exceeds 10.
- **If we are only using one instrument, the F-test would be equivalent to?** (The t-test of that instrument) [2].

(b) Checking for Exogeneity

- **When can we check for the exogeneity of the instruments?**

  Only whenever we have an *overidentified* model. That is, when we have more instruments than problematic variables ($m > k$).

  In all other cases, you can only justify the exogeneity of the instruments on logical or substantive grounds.

  Notice that once we have Ws in the regression, the requirement of exogeneity implies that $corr(Z_1, u) = 0$, after controlling for all the Ws.

- To check for the exogeneity of the instruments, we can use something called the *J test of overidentifying restrictions*.We will look at this next week.

Intuition of the J test:

- When we have more instruments than endogenous regressors, for example, 2 instruments for 1 regressor, we can compare the estimates that we get using the two different instruments.
- If both instruments are exogenous, then the choice of the instrument used should not have a large impact on the results.
- If the results, however, are very different depending on the instrument used, then, we conclude that something is wrong with one or the other of the instruments - or both.

Calculation of the J test:

(1) Estimate the TSLS with *all* the instruments.

(2) Calculate $\hat{u}_i^{TSLS}$:

$$\hat{u}_i^{TSLS} = Y_i - \hat{Y}_i^{TSLS} = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + ... + \hat{\beta}_{k+r} W_{ri})$$

Notice that the true Xs and not the $\hat{X}$s are used.

(3) Regress $\hat{u}_i^{TSLS}$ on the Zs and the Ws (assuming homoskedasticity!)

$$\hat{u}_i = \delta_0 + \delta_1 Z_{1i} + ... + \delta_m Z_{mi} + \delta_{m+1} W_{1i} + ... + \delta_{m+k} W_{ri} + e_i$$

(4) Calculate the F-test of whether all of the coefficients on the Zs are zero. (H_0=$\delta_1 = ... = \delta_m = 0$.)

(5) $J = m\,F$.

Conclusions with the J test:

- Under the null, J is distributed χ^2_{m-k}.
- If we fail to reject the null, then the hypothesis of the exogeneity of the instruments is accepted.
- If we reject the null, then, you have some evidence that one or both of the instruments are endogenous.

Example of IV Analysis

Does an $\triangle$ of prisoners in jail ∇ the crime rate? [1]

Problem: Simultaneity/Endogeneity.

- $\triangle$ in the number of prisoners are likely to ∇ crime ($X \rightarrow Y$)
- $\triangle$ in crime rates also translate into $\triangle$ in prisoners ($Y \rightarrow X$)

Solution: IV analysis with prison overcrowding litigation as the instrument.

- It is likely to have a negative impact on prison populations
- It is unlikely to be related to fluctuations in the crime rate, except through its effect on prison populations.

[1] Source: S.D. Levitt (1996), "The Effect of Prison Population Size on Crime Rates: Evidence from Prison Overcrowding Legislation"

Data:
Annual, state-level observations.
All variables are changes from 1990 to 1991 for 51 US states:

variable	variable description
ln_crime	log of # of violent crimes per 100,000 residents
ln_pris	log of # of prisoners per 100,000 residents
cmetro	# of citizens in a metro area per 100,000 residents
ln_income	log of average per capita income in the state
police	# of policemen per 100,000 state residents
final1	=1 if overcrowding litigation reached this year
final2	=1 if overcrowding litigation reached in past 2 years

Option 1: Doing the IV Analysis Step by Step
(1) Run the first-stage regression

```
. reg ln_pris final1 final2 cmetro ln_income police , r

Linear regression                               Number of obs =     714
                                                F(  5,   708) =   10.48
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0312
                                                Root MSE      =  .06577

------------------------------------------------------------------------------
             |               Robust
    ln_pris |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
      final1 |   -.076315   .0122032    -6.25   0.000    -.1002737   -.0523562
      final2 |  -.0652256   .0200727    -3.25   0.001    -.1046347   -.0258165
      cmetro |  -1.129895    .907242    -1.25   0.213    -2.911102    .6513113
   ln_income |  -.1162396   .1249814    -0.93   0.353     -.361618    .1291389
      police |  -.0424221   .0336127    -1.26   0.207    -.1084147    .0235704
       _cons |   .0772898   .0078982     9.79   0.000     .0617832    .0927964
------------------------------------------------------------------------------
```

(2) Calculate the $\hat{X}$

```
. generate h_ln_pris = _b[_cons] *1 + _b[ final1] * final1 + _b[ final2] * final
> 2 + _b[ cmetro] * cmetro + _b[ ln_income] * ln_income + _b[ police] * police
```

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

(3) Run the second-stage regression

```
. reg ln_crime h_ln_pris cmetro ln_income police, r

Linear regression                               Number of obs =     714
                                                F(  4,   709) =    8.81
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0475
                                                Root MSE      =  .08667

------------------------------------------------------------------------------
             |               Robust
   ln_crime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
   h_ln_pris |  -1.092828   .2796472    -3.91   0.000    -1.641864   -.5437928
     cmetro |  -.7101292   1.470325    -0.48   0.629    -3.596842    2.176584
  ln_income |   .3784857   .1450369     2.61   0.009     .0937325    .6632389
     police |   .0773504    .057093     1.35   0.176    -.0347411    .1894419
      _cons |   .0697582   .0231102     3.02   0.003     .0243855    .1151309
------------------------------------------------------------------------------
```

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

Option 2: Doing the IV Analysis All at Once

(1) Use ivreg in STATA (or other similar statistical packages)

```
. ivreg ln_crime  cmetro  ln_income   police  (ln_pris=final1 final2), r

Instrumental variables (2SLS) regression        Number of obs =     714
                                                F(  4,   709) =    5.56
                                                Prob > F      =  0.0002
                                                R-squared     =       .
                                                Root MSE      =  .10529

------------------------------------------------------------------------------
             |               Robust
   ln_crime |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
    ln_pris |  -1.092829   .3094878    -3.53   0.000    -1.700451   -.4852063
     cmetro |  -.7101293   1.691211    -0.42   0.675    -4.03051    2.610251
  ln_income |   .3784857   .2043234     1.85   0.064    -.0226656    .779637
     police |   .0773504   .0624596     1.24   0.216    -.0452775    .1999782
      _cons |   .0697582   .0261249     2.67   0.008     .0184668    .1210496
------------------------------------------------------------------------------
Instrumented:  ln_pris
Instruments:   cmetro ln_income police final1 final2
------------------------------------------------------------------------------
```

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

Between Option 1 and Option 2

- **How do the estimated coefficients compare?** They are the same.
- **How do the SE compare?** The SE of option 2 are larger.
- **Which SE are correct?** Those of option 2.

  When the estimation process is done by stages (as opposed to simultaneously) the second-stage OLS standard errors are incorrect because they fail to adjust for the fact that the second-stage regression uses the predicted values of the included endogenous variables.

If we want to assess whether the instruments are valid, which two questions shall we answer? We need to answer the following two questions:

- (a) Are the instruments relevant?
- (b) Are the instruments exogenous?

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

(a) Testing Instrument Relevance
From the 1^{st} *stage*, we can calculate the following F-test:

```
. test final1 final2

 ( 1)  final1 = 0
 ( 2)  final2 = 0

       F(  2,   708) =   24.46
            Prob > F =    0.0000
```

Based on our results, are our instruments relevant and not weak? Yes. Remember the rule of thumb is: We will say that our instruments are not weak if the first-stage F statistic (H_0: coefficient of all instruments = 0) exceeds 10.

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

(b) Testing Instrument Exogeneity We can use the *J*-test of overidentifying restrictions to test whether both instruments are exogenous:

H_0: final1=0 and final2=0
H_A: final1≠0 and/or final2≠0

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

```
. ivreg ln_crime  cmetro  ln_income   police  (ln_pris=final1 final2), r
```

...(regression output)

```
. predict residual, resid

. reg residual final1 final2  cmetro ln_income  police

      Source |       SS       df       MS              Number of obs =     714
-------------+------------------------------           F(  5,   708) =    0.01
       Model |  .000741396     5  .000148279            Prob > F      =  0.9999
    Residual |  7.85853827   708  .01109963             R-squared     =  0.0001
-------------+------------------------------           Adj R-squared = -0.0070
       Total |  7.85927967   713  .011022833            Root MSE      =  .10535

------------------------------------------------------------------------------
    residual |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      final1 |   .0086237   .0434347     0.20   0.843    -.0766525    .0938999
      final2 |  -.0050312    .030721    -0.16   0.870    -.0653464    .0552839
     cmetro |  -.0102384   1.285999    -0.01   0.994    -2.535066    2.514589
   ln_income |   .0024082   .1391296     0.02   0.986    -.2707479    .2755643
      police |   .0014565   .0730567     0.02   0.984    -.1419771    .1448902
       _cons |  -.0001269    .00963    -0.01   0.989    -.0190338    .0187799
------------------------------------------------------------------------------
```

```
. test final1 final2

 ( 1)  final1 = 0
 ( 2)  final2 = 0

       F(  2,   708) =    0.03
            Prob > F =    0.9672

. sca J = r(F)*2

. sca pval= chiprob(1,J)

. display J
.06679469

. display pval
.79606218
```

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

Based on our results, can we reject the null hypothesis at the conventional levels of significance? No.

Based on our results, then, are the instruments exogenous? Based on the J-test, we have no evidence to the contrary.

Econ 1123: Section 7

Review
IV Analysis
TSLS
Special Case TSLS
General IV Regression Model
Example
STATA Help

STATA help for Problem Set 7

To do an IV analysis on STATA, the command is:

ivreg Y controls (X= instruments), r