Econ 422 – Lecture Notes
Part I

(These notes are slightly modified versions of lecture notes provided by Stock and Watson, 2007. They are for instructional purposes only and are not to be distributed outside of the classroom.)

The General IV Regression Model

- So far we have considered IV regression with a single endogenous regressor (X) and a single instrument (Z).
- We need to extend this to:
  - multiple endogenous regressors ($X_1,\ldots,X_k$)
  - multiple included exogenous variables ($W_1,\ldots,W_r$)

    These need to be included for the usual OV reason
  - multiple instrumental variables ($Z_1,\ldots,Z_m$)

    More (relevant) instruments can produce a smaller variance of TSLS: the R^2 of the first stage increases, so you have more variation in $\hat{X}$.
- Terminology: identification & overidentification

Identification

- In general, a parameter is said to be ***identified*** if different values of the parameter would produce different distributions of the data.
- In IV regression, whether the coefficients are identified depends on the relation between the number of instruments (m) and the number of endogenous regressors (k)
- Intuitively, if there are fewer instruments than endogenous regressors, we can't estimate $\beta_1,\ldots,\beta_k$
  - For example, suppose $k = 1$ but $m = 0$ (no instruments)!

Identification, ctd.

The coefficients $\beta_1,\ldots,\beta_k$ are said to be:

- ***exactly identified*** if $m = k$.

  There are just enough instruments to estimate $\beta_1,\ldots,\beta_k$.

- ***overidentified*** if $m > k$.

  There are more than enough instruments to estimate $\beta_1,\ldots,\beta_k$. *If so, you can test whether the instruments are valid* (*a test of the "overidentifying restrictions*") – *we'll return to this later*

- ***underidentified*** if $m < k$.

  There are too few instruments to estimate $\beta_1,\ldots,\beta_k$. *If so, you need to get more instruments!*

The general IV regression model: Summary of jargon

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

- Y_i is the ***dependent variable***
- $X_{1i},\ldots, X_{ki}$ are the ***endogenous regressors*** (potentially correlated with u_i)
- $W_{1i},\ldots,W_{ri}$ are the ***included exogenous variables*** or ***included exogenous regressors*** (uncorrelated with u_i)
- $\beta_0, \beta_1,\ldots, \beta_{k+r}$ are the unknown regression coefficients
- $Z_{1i},\ldots,Z_{mi}$ are the m ***instrumental variables*** (the ***excluded exogenous variables***)
- The coefficients are ***overidentified*** if $m > k$; ***exactly identified*** if $m =$ k; and ***underidentified*** if $m < k$.

TSLS with a single endogenous regressor

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

- m instruments: $Z_{1i}, \ldots, Z_m$
- First stage
  - Regress X_1 on *all* the exogenous regressors: regress X_1 on $W_1, \ldots, W_r, Z_1, \ldots, Z_m$ by OLS
  - Compute predicted values $\hat{X}_{1i}$, $i = 1, \ldots, n$
- Second stage
  - Regress Y on $\hat{X}_1$, $W_1, \ldots, W_r$ by OLS
  - The coefficients from this second stage regression are the TSLS estimators, but *SE*s are wrong
- To get correct *SE*s, do this in a single step

Example: Demand for cigarettes

$$\ln(Q_i^{cigarettes}) = \beta_0 + \beta_1 \ln(P_i^{cigarettes}) + \beta_2 \ln(Income_i) + u_i$$

Z_{1i} = general sales tax$_i$

Z_{2i} = cigarette-specific tax$_i$

- Endogenous variable: $\ln(P_i^{cigarettes})$ ("one X")
- Included exogenous variable: $\ln(Income_i)$ ("one W")
- Instruments (excluded endogenous variables): general sales tax, cigarette-specific tax ("two Zs")
- *Is the demand elasticity β_1 overidentified, exactly identified, or underidentified?*

Example: Cigarette demand, one instrument

```
            Y       W          X          Z
. ivreg lpackpc lperinc (lravgprs = rtaxso) if year==1995, r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =    8.19
                                                      Prob > F      =  0.0009
                                                      R-squared     =  0.4189
                                                      Root MSE      =  .18957

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.143375   .3723025    -3.07   0.004    -1.893231   -.3935191
     lperinc |    .214515   .3117467     0.69   0.495     -.413375     .842405
       _cons |   9.430658   1.259392     7.49   0.000     6.894112    11.9672
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso        STATA lists ALL the exogenous regressors
                                     as instruments - slightly different
                                     terminology than we have been using
------------------------------------------------------------------------------
```

- Running IV as a single command yields correct *SE*s
- Use , r for heteroskedasticity-robust *SE*s

Example: Cigarette demand, two instruments

Y *W* *X* Z_1 Z_2

```
. ivreg lpackpc lperinc (lravgprs = rtaxso rtax) if year==1995, r;

IV (2SLS) regression with robust standard errors      Number of obs =      48
                                                      F(  2,    45) =   16.17
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.4294
                                                      Root MSE      =  .18786

------------------------------------------------------------------------------
             |               Robust
     lpackpc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    lravgprs |  -1.277424   .2496099    -5.12   0.000    -1.780164   -.7746837
     lperinc |   .2804045   .2538894     1.10   0.275     -.230955    .7917641
       _cons |   9.894955   .9592169    10.32   0.000     7.962993    11.82692
------------------------------------------------------------------------------
Instrumented:  lravgprs
Instruments:   lperinc rtaxso rtax    STATA lists ALL the exogenous regressors
                                      as "instruments" - slightly different
                                      terminology than we have been using
------------------------------------------------------------------------------
```

TSLS estimates, Z = sales tax ($m = 1$)

$$\widehat{\ln(Q_i^{cigarettes})} = 9.43 - 1.14\widehat{\ln(P_i^{cigarettes})} + 0.21\ln(Income_i)$$
$$(1.26) \quad (0.37) \qquad\qquad (0.31)$$

TSLS estimates, Z = sales tax, cig-only tax ($m = 2$)

$$\widehat{\ln(Q_i^{cigarettes})} = 9.89 - 1.28\widehat{\ln(P_i^{cigarettes})} + 0.28\ln(Income_i)$$
$$(0.96) \quad (0.25) \qquad\qquad (0.25)$$

- Smaller *SEs* for $m = 2$. Using 2 instruments gives more information – more "as-if random variation".
- Low income elasticity (not a luxury good); income elasticity not statistically significantly different from 0
- Surprisingly high price elasticity

The General Instrument Validity Assumptions

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

(1) ***Instrument exogeneity***: $\text{corr}(Z_{1i}, u_i) = 0, \ldots, \text{corr}(Z_{mi}, u_i) = 0$

(2) ***Instrument relevance***: *General case, multiple X's*

Suppose the second stage regression could be run using the predicted values from the *population* first stage regression. Then: there is no perfect multicollinearity in this (infeasible) second stage regression.

- Multicollinearity interpretation…
- *Special case of one X*: the general assumption is equivalent to (a) at least one instrument must enter the population counterpart of the first stage regression, and (b) the *W*'s are not perfectly multicollinear.

The IV Regression Assumptions

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

1. $E(u_i | W_{1i}, \ldots, W_{ri}) = 0$
   - #1 says "the exogenous regressors are exogenous."
2. $(Y_i, X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{mi})$ are i.i.d.
   - #2 is not new
3. The X's, W's, Z's, and Y have nonzero, finite 4^{th} moments
   - #3 is not new
4. The instruments $(Z_{1i}, \ldots, Z_{mi})$ are valid.
   - We have discussed this

- Under 1-4, TSLS and its t-statistic are normally distributed
- The critical requirement is that the instruments be valid…

Checking Instrument Validity

Recall the two requirements for valid instruments:

1. *Relevance* (special case of one X)

   At least one instrument must enter the population counterpart of the first stage regression.

2. *Exogeneity*

   All the instruments must be uncorrelated with the error term: $\text{corr}(Z_{1i},u_i) = 0,\ldots, \text{corr}(Z_{mi},u_i) = 0$

What happens if one of these requirements isn't satisfied?

How can you check? What do you do?

If you have multiple instruments, which should you use?

Checking Assumption #1: Instrument Relevance

We will focus on a single included endogenous regressor:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \ldots + \beta_{1+r} W_{ri} + u_i$$

First stage regression:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \ldots + \pi_{m+k} W_{ki} + u_i$$

- The instruments are relevant if at least one of $\pi_1, \ldots, \pi_m$ are nonzero.
- The instruments are said to be ***weak*** if all the $\pi_1, \ldots, \pi_m$ are either zero or nearly zero.
- ***Weak instruments*** explain very little of the variation in X, beyond that explained by the W's

What are the consequences of weak instruments?

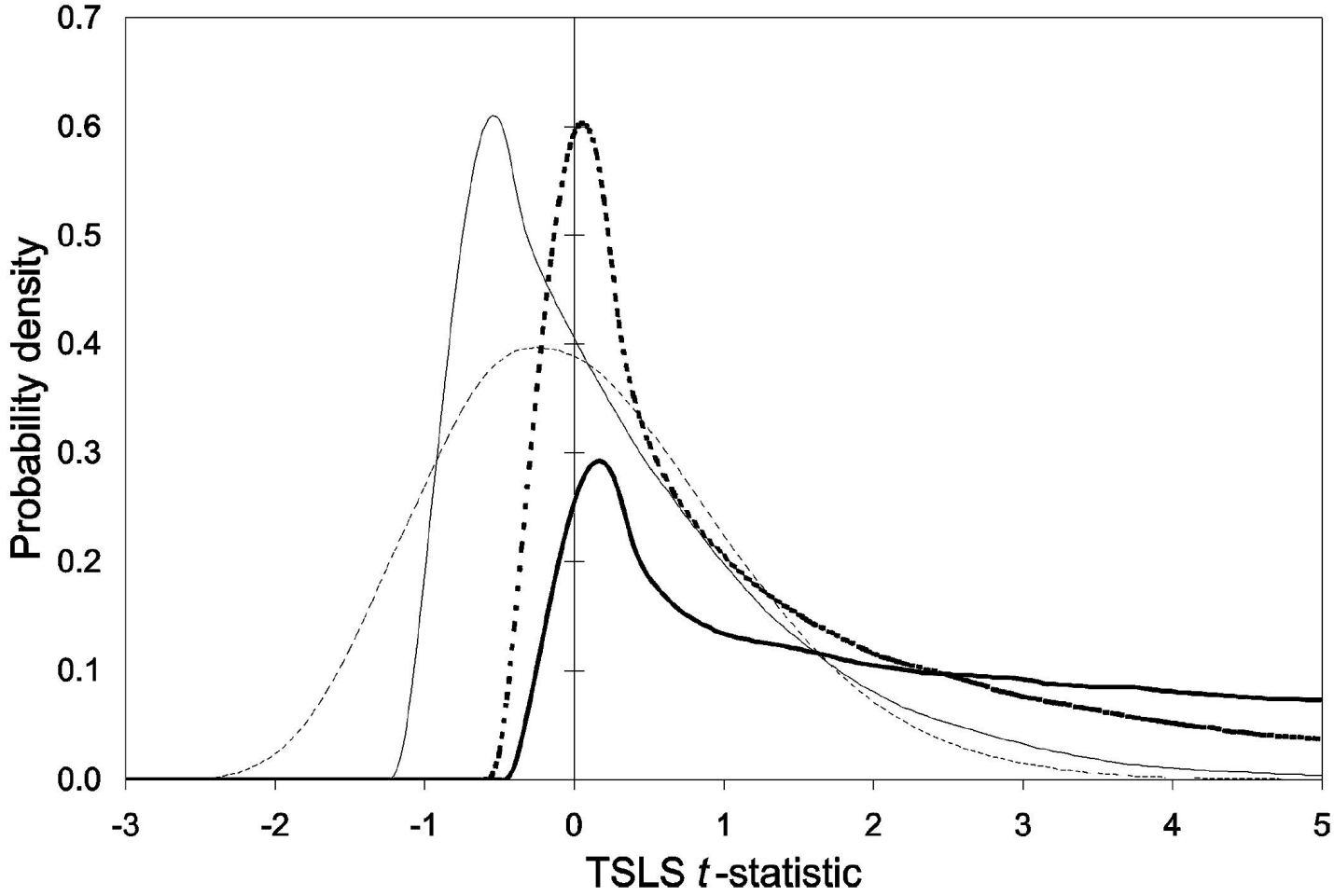
If instruments are weak, the sampling distribution of TSLS and its *t*-statistic are not (at all) normal, even with *n* large. Consider the simplest case:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \pi_0 + \pi_1 Z_i + u_i$$

- The IV estimator is $\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$
- If cov(X,Z) is zero or small, then s_{XZ} will be small: With weak instruments, the denominator is nearly zero.
- If so, the sampling distribution of $\hat{\beta}_1^{TSLS}$ (and its *t*-statistic) is not well approximated by its large-*n* normal approximation…

An example: the sampling distribution of the TSLS *t*-statistic with weak instruments



Dark line = irrelevant instruments

Dashed light line = strong instruments

Why does our trusty normal approximation fail us?

$$\hat{\beta}_1^{TSLS} = \frac{s_{YZ}}{s_{XZ}}$$

- If cov(X,Z) is small, small changes in s_{XZ} (from one sample to the next) can induce big changes in $\hat{\beta}_1^{TSLS}$
- Suppose in one sample you calculate s_{XZ} = .00001...
- Thus the large-n normal approximation is a poor approximation to the sampling distribution of $\hat{\beta}_1^{TSLS}$
- A better approximation is that $\hat{\beta}_1^{TSLS}$ is distributed as the *ratio* of two correlated normal random variables (see SW App. 12.4)
- If instruments are weak, the usual methods of inference are unreliable – potentially very unreliable.

Measuring the strength of instruments in practice: The first-stage *F*-statistic

- The first stage regression (one X):

  Regress X on $Z_1,..,Z_m,W_1,\ldots,W_k$.

- Totally irrelevant instruments $\Leftrightarrow$ *all* the coefficients on $Z_1,\ldots,Z_m$ are zero.
- The ***first-stage F-statistic*** tests the hypothesis that $Z_1,\ldots,Z_m$ do not enter the first stage regression.
- Weak instruments imply a small first stage F-statistic.

Checking for weak instruments with a single X

- Compute the first-stage F-statistic.

  Rule-of-thumb: If the first stage F-statistic is less than 10, then the set of instruments is weak.

- If so, the TSLS estimator will become more severely biased, and statistical inferences (standard errors, hypothesis tests, confidence intervals) can be misleading.
- Note that simply rejecting the null hypothesis that the coefficients on the Z's are zero isn't enough – you actually need substantial predictive content for the normal approximation to be a good one.
- There are more sophisticated things to do than just compare F to 10 but they are beyond this course.

What to do if you have weak instruments?

- Get better instruments (!)
- If you have a number of instruments and perhaps some are much weaker than others; then, one approach is to drop some of the weaker ones (dropping an irrelevant instrument will increase the first-stage *F*). This leads to a possible pre-testing problem, however.
- If you only have a few instruments, and all are weak, then you need to do some IV analysis other than TSLS…
    - Separate the problem of estimation of β_1 and construction of confidence intervals
    - This seems odd, but if TSLS isn't normally distributed, it makes sense (right?)

Estimation with weak instruments

- TSLS estimator become more severely biased if instruments are weak or irrelevant.
- However, some estimators have a distribution more centered around β_1 than does TSLS
- One such estimator is the limited information maximum likelihood estimator (LIML)
- The LIML estimator
  - can be derived as a maximum likelihood estimator
  - is the value of β_1 that minimizes the p-value of the AR test(!)

Checking Assumption #2: Instrument Exogeneity

- Instrument exogeneity: ***All*** the instruments are uncorrelated with the error term: $\text{corr}(Z_{1i},u_i) = 0,\ldots,$ $\text{corr}(Z_{mi},u_i) = 0$
- If the instruments are correlated with the error term, the first stage of TSLS doesn't successfully isolate a component of X that is uncorrelated with the error term, so $\hat{X}$ is correlated with u and TSLS is inconsistent.
- If there are more instruments than endogenous regressors, it is possible to test – *partially* – for instrument exogeneity.

Testing overidentifying restrictions

Consider the simplest case:

$$Y_i = \beta_0 + \beta_1 X_i + u_i,$$

- Suppose there are two valid instruments: Z_{1i}, Z_{2i}
- Then you could compute two separate TSLS estimates.
- Intuitively, if these 2 TSLS estimates are very different from each other, then something must be wrong: one or the other (or both) of the instruments must be invalid.
- The *J*-test of overidentifying restrictions makes this comparison in a statistically precise way.
- This can only be done if #*Z*'s > #*X*'s (overidentified).

Suppose #instruments = $m >$ # X's = k (overidentified)

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

The *J*-test of overidentifying restrictions

The *J*-test can be constructed as follows:

1. First estimate the equation of interest using TSLS and all m instruments; compute the predicted values $\hat{Y}_i$, using the *actual* X's (not the $\hat{X}$'s used to estimate the second stage)
2. Compute the residuals $\hat{u}_i = Y_i - \hat{Y}_i$
3. Regress $\hat{u}_i$ against $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$
4. Compute the F-statistic testing the hypothesis that the coefficients on $Z_{1i},\ldots,Z_{mi}$ are all zero;
5. The ***J-statistic*** is $J = mF$

$J = mF$, where F = the F-statistic testing the coefficients on $Z_{1i},\ldots,Z_{mi}$ in a regression of the TSLS residuals against $Z_{1i},\ldots,Z_{mi}$, $W_{1i},\ldots,W_{ri}$.

Distribution of the *J*-statistic

- Under the null hypothesis that all the instruments are exogeneous, J has a chi-squared distribution with $m–k$ degrees of freedom
- If $m = k$, $J = 0$ (*does this make sense?*)
- If some instruments are exogenous and others are endogenous, the J statistic will be large, and the null hypothesis that all instruments are exogenous will be rejected.

Checking Instrument Validity: Summary

The two requirements for valid instruments:

1. ***Relevance*** (special case of one X)
   - At least one instrument must enter the population counterpart of the first stage regression.
   - If instruments are weak, then the TSLS estimator will be more severely biased and the *t*-statistic has a non-normal distribution
   - To check for weak instruments with a single included endogenous regressor, check the first-stage *F*
     - If $F>10$, instruments are strong – use TSLS
     - If $F<10$, weak instruments – take some action

2. *Exogeneity*

- ***All*** the instruments must be uncorrelated with the error term: $\text{corr}(Z_{1i},u_i) = 0,\ldots, \text{corr}(Z_{mi},u_i) = 0$
- We can partially test for exogeneity: if $m>1$, we can test the hypothesis that all are exogenous, against the alternative that as many as $m-1$ are endogenous (correlated with u)
- The test is the J-test, constructed using the TSLS residuals.