ECON4150 - Introductory Econometrics

Lecture 16: Instrumental variables

Monique de Haan
(moniqued@econ.uio.no)

Stock and Watson Chapter 12

Lecture outline

- OLS assumptions and when they are violated
- Instrumental variable approach
    - 1 endogenous regressor & 1 instrument
    - IV assumptions:
        - instrument relevance
        - instrument exogeneity
    - 1 endogenous regressor, 1 instrument & control variables
    - 1 endogenous regressor & multiple instruments
    - multiple endogenous regressors & multiple instruments

Introduction

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

The 3 assumptions of an OLS regression model:

1. $E(u_i|X_i) = 0$
2. (X_i, Y_i), $i = 1, ...N$ are independently and identically distributed
3. Big outliers are unlikely.

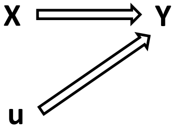
Threats to internal validity (violation of 1st OLS assumption):

- Omitted variables
- Functional form misspecification
- Measurement error
- Sample selection
- Simultaneous causality

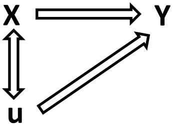
Introduction

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

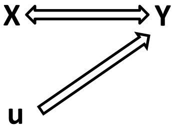
We **can** use OLS to obtain consistent estimate of the causal effect if



We **can't** use OLS to obtain consistent estimate of the causal effect if



and/or



Instrumental variables: 1 endogenous regressor & 1 instrument

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

- Potential solution if $E[u_i|X_i] \neq 0$: use an instrumental variable (Z_i)
- We want to split X_i into two parts:

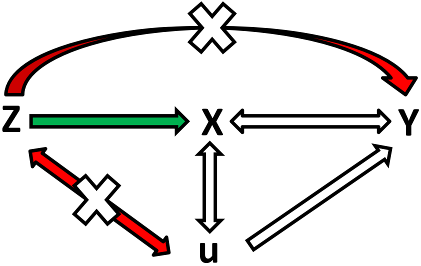
1. part that is correlated with the error term (causing $E[u_i|x_i] \neq 0$)
2. part that is uncorrelated with the error term

- If we can isolate the variation in X_i that is uncorrelated with u_i...
- ...we can use this to obtain a consistent estimate of the causal effect of X_i on Y_i

Instrumental variables: 1 endogenous regressor & 1 instrument

- In order to isolate the variation in X_i that is uncorrelated with u_i we can use an **instrumental variable** Z_i with the following properties:

1. **Instrument relevance:** Z_i is correlated with the endogenous regressor $Cov(Z_i, X_i) \neq 0$
2. **Instrument exogeneity:** Z_i is uncorrelated with the error term $Cov(Z_i, u_i) = 0$ and has no direct effect on Y_i



Instrumental variables: 1 endogenous regressor & 1 instrument

We can extend the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad\qquad X_i = \pi_0 + \pi_1 Z_i + v_i$$

We can estimate the causal effect of X_i on Y_i in two steps:

First stage: Regress X_i on Z_i & obtain predicted values $\widehat{X}_i = \widehat{\pi}_0 + \widehat{\pi}_1 Z_i$

- If $Cov(Z_i, u_i) = 0$, $\widehat{X}_i$ contains variation in X_i that is uncorrelated with u_i

Second stage: Regress Y_i on $\widehat{X}_i$ to obtain the Two Stage Least Squares estimator $\hat{\beta}_{2SLS}$:

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}) \left(\widehat{X}_i - \overline{\widehat{X}}\right)}{\sum_{i=1}^{n} \left(\widehat{X}_i - \overline{\widehat{X}}\right)^2}$$

Application: estimating the returns to education

- Data from the NLS Young Men Cohort collected in 1976 on (among others) wages and years of education for 3010 men.
- Data are provided by Professor David Card, he used the data in his article "Using Geographic Variation in College Proximity to Estimate the Return to Schooling"

```
. regress ln_wage education, robust

Linear regression                                      Number of obs =       3010
                                                          F(  1,  3008) =     321.16
                                                          Prob > F      =     0.0000
                                                          R-squared     =     0.0987
                                                          Root MSE      =     .42139

------------------------------------------------------------------------------
             |               Robust
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .0520942   .0029069    17.92   0.000     .0463946    .0577939
       _cons |   5.570882   .0390935   142.50   0.000      5.49423    5.647535
------------------------------------------------------------------------------
```

- OLS estimate of the returns to education likely inconsistent due to omitted variables and measurement error.

Application: estimating the returns to education

- We want to isolate variation in years of education that is uncorrelated with the error term
- Card (1995) uses variation in college proximity as instrumental variable
- We have the following instrumental variable

near_college= 1 if individual grew up in area with a 4-year college
0 if individual grew up in area without a 4-year college

Step 1: First stage regression

```
. regress education near_college, robust

Linear regression                                   Number of obs =      3010
                                                       F(  1,  3008) =     60.37
                                                       Prob > F      =    0.0000
                                                       R-squared     =    0.0208
                                                       Root MSE      =    2.6494
```

education	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
near_college	.829019	.1066941	7.77	0.000	.6198182	1.03822
_cons	12.69801	.0902199	140.75	0.000	12.52112	12.87491

Application: estimating the returns to education

Step 2: Obtain the predicted values and perform the second stage regression

```
1 . predict pr_education, xb

2 . regress ln_wage pr_education, robust

Linear regression                                      Number of obs =      3010
                                                         F(  1,  3008) =     83.79
                                                         Prob > F      =    0.0000
                                                         R-squared     =    0.0268
                                                         Root MSE      =    .43789

------------------------------------------------------------------------------
             |               Robust
     ln_wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
pr_education |   .1880626   .0205454     9.15   0.000     .1477781    .2283472
       _cons |   3.767472   .2724927    13.83   0.000     3.233181    4.301763
------------------------------------------------------------------------------
```

Instrumental variables: 1 endogenous regressor & 1 instrument

Regression Y_i on $\widehat{X}_i$ gives the 2SLS estimator

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(\widehat{X}_i - \overline{\widehat{X}}\right)}{\sum_{i=1}^{n}\left(\widehat{X}_i - \overline{\widehat{X}}\right)^2}$$

If we substitute $\widehat{X}_i - \overline{\widehat{X}} = \left(\widehat{\pi}_0 + \widehat{\pi}_1 Z_i\right) - \left(\widehat{\pi}_0 + \widehat{\pi}_1 \overline{Z}\right) = \widehat{\pi}_1\left(Z_i - \overline{Z}\right)$ we get

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\widehat{\pi}_1\left(Z_i - \overline{Z}\right)}{\sum_{i=1}^{n}\widehat{\pi}_1^2\left(Z_i - \overline{Z}\right)^2} = \frac{1}{\widehat{\pi}_1} \times \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(Z_i - \overline{Z}\right)}{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2}$$

Since $\widehat{\pi}_1$ is the first stage OLS estimator:

$$\hat{\beta}_{2SLS} = \frac{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Z_i - \overline{Z}\right)} \times \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(Z_i - \overline{Z}\right)}{\sum_{i=1}^{n}\left(Z_i - \overline{Z}\right)^2}$$

Which gives the instrumental variable estimator

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)\left(Z_i - \overline{Z}\right)}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Z_i - \overline{Z}\right)}$$

Application: estimating the returns to education

- We can obtain the 2SLS estimator in two steps as we have seen
- However the standard errors reported in the second stage regression are incorrect
- Stata does not recognize that it is a second stage of a two stage process, it fails to take into account the uncertainty in the first stage estimation.
- Instead obtain the 2SLS-estimator in 1 step:

```
. ivregress 2sls ln_wage (education=near_college), robust

Instrumental variables (2SLS) regression               Number of obs =       3010
                                                       Wald chi2(    1)  =   51.78
                                                       Prob > chi2   =     0.0000
                                                       R-squared     =          .
                                                       Root MSE      =     .55667

------------------------------------------------------------------------------
             |               Robust
     ln_wage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .1880626   .0261339     7.20   0.000     .1368412    .2392841
       _cons |   3.767472   .3466268    10.87   0.000     3.088096    4.446848
------------------------------------------------------------------------------
Instrumented:  education
Instruments:   near_college
```

Instrumental variables: 1 endogenous regressor & 1 instrument

$$\hat{\beta}_{IV} = \frac{\sum_{i=1}^{n} (Y_i - \bar{Y}) \left(Z_i - \overline{Z}\right)}{\sum_{i=1}^{n} (X_i - \bar{X}) \left(Z_i - \overline{Z}\right)}$$

In large samples the IV-estimator converges to

$$plim(\hat{\beta}_{IV}) = \frac{Cov(Y_i, Z_i)}{Cov(X_i, Z_i)} = \frac{Cov(\beta_0 + \beta_1 X_i + u_i, Z_i)}{Cov(X_i, Z_i)} = \beta_1 + \frac{Cov(u_i, Z_i)}{Cov(X_i, Z_i)}$$

If the two IV-assumptions hold

1. **Instrument relevance:** $Cov(Z_i, X_i) \neq 0$
2. **Instrument exogeneity:** $Cov(Z_i, u_i) = 0$

The IV-estimator is consistent $plim(\hat{\beta}_{IV}) = \beta_1$, and is normally distributed in large samples

$$\hat{\beta}_{IV} \sim N\left(\beta_1, \ \frac{1}{n} \frac{Var\left[(Z_i - \mu_Z)\, u_i\right]}{\left[Cov\left(Z_i, X_i\right)\right]^2}\right)$$

Instrumental variables: 1 endogenous regressor & 1 instrument

The Instrumental Variables estimator is not unbiased

$$
\begin{aligned}
E\left[\hat{\beta}_{IV}\right] &= E\left[\frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(Z_i-\overline{Z}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Z_i-\overline{Z}\right)}\right] \\
&= E\left[\frac{\sum_{i=1}^{n}\left((\beta_0+\beta_1 X_i+u_i)-\left(\beta_0+\beta_1\overline{X}+\bar{u}\right)\right)\left(Z_i-\overline{Z}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Z_i-\overline{Z}\right)}\right] \\
&= E\left[\frac{\beta_1 \sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Z_i-\overline{Z}\right)+\sum_{i=1}^{n}(u_i-\bar{u})\left(Z_i-\overline{Z}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Z_i-\overline{Z}\right)}\right] \\
&= \beta_1 + E\left[\frac{\sum_{i=1}^{n}(u_i-\bar{u})\left(Z_i-\overline{Z}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Z_i-\overline{Z}\right)}\right] = \beta_1 + E\left[\frac{\sum_{i=1}^{n} u_i\left(Z_i-\overline{Z}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Z_i-\overline{Z}\right)}\right] \\
&= \beta_1 + E_{X,Z}\left[\frac{\sum_{i=1}^{n} E[u_i|Z_i,X_i]\left(Z_i-\overline{Z}\right)}{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)\left(Z_i-\overline{Z}\right)}\right] \\
&\neq \beta_1
\end{aligned}
$$

Instrument exogeneity implies $E[u_i|Z_i] = 0$ but not $E[u_i|Z_i, X_i] = 0$ (this would mean that $E[u_i|X_i] = 0$ and we would not need an instrument!)

Instrumental variables: 1 endogenous regressor & 1 instrument

How can we know whether the IV assumptions hold?

1. **Instrument relevance:** $Cov(Z_i, X_i) \neq 0$
   - We can check whether instrument relevance holds.
   - Note that $\pi_1 = \frac{Cov(Z_i, X_i)}{Var(Z_i)}$
   - We can therefore test $H_0 : \pi_1 = 0$ against $H_1 : \pi_1 \neq 0$
2. **Instrument exogeneity:** $Cov(Z_i, u_i) = 0$
   - We can't check whether this assumption holds.
   - We need to use economic theory, expert knowledge and intuition.

Instrument relevance & weak instruments

- Clearly, an irrelevant instrumental variable has problems, recall that

$$\hat{\beta}_{2SLS} \rightarrow \frac{Cov(Y_i, Z_i)}{Cov(X_i, Z_i)}$$

- In case of an irrelevant (but exogenous) instrumental variable both the denominator and numerator are 0.

- If instrument is not irrelevant but $Cov(X_i, Z_i)$ is close to zero

  - The sampling distribution of $\hat{\beta}_{2SLS}$ is not normal
  - $\hat{\beta}_{2SLS}$ can be severely biased, in the direction of the OLS estimator, even in relatively large samples!

- We should therefore always check whether an instrument is relevant enough.

Instrument relevance & weak instruments

- Let F_{first} be the F-statistic resulting from the test $H_0 : \pi_1 = 0$ against $H_1 : \pi_1 \neq 0$

- Staiger & Stock (Econometrica, 1997) show that in a simple model $\frac{1}{F_{first}}$ provides approximate estimate of finite sample bias of $\widehat{\beta}_{2SLS}$ relative to $\widehat{\beta}_{OLS}$

- Stock & Yogo (2005) argue that instruments are weak if the IV Bias is more than 10% of the OLS Bias.

- **Rule of thumb**: the F-statistic for (joint) significance of the instrument(s) in the first-stage should exceed 10.

Application: estimating the returns to education

Do the instrumental variable assumptions hold for college proximity as an instrument to estimate the returns to education?

1. Instrument relevance/weak instruments

education	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
near_college	.829019	.1066941	7.77	0.000	.6198182	1.03822
_cons	12.69801	.0902199	140.75	0.000	12.52112	12.87491

```
. test near_college

 ( 1)  near_college = 0

       F(  1,  3008) =    60.37
            Prob > F =    0.0000
```

2. Instrument exogeneity:

- Is there a direct effect of living near a 4 year college on earnings?
- Is college proximity related to omitted variables that affect earnings?
  - What about area characteristics, such as living in a big city instead of a small village?

1 endogenous regressor, 1 instrument & control variables

- We can weaken the instrument exogeneity assumption by including area characteristics as control variables
- The Instrumental variables model is extended by including the control variables $W_{1i}, \ldots, W_{ri}$

$$Y_i = \beta_0 + \beta_1 X_i + \delta_1 W_{1i} +, \ldots, + \delta_r W_{ri} + u_i$$

$$X_i = \pi_0 + \pi_1 Z_i + \gamma_1 W_{1i} + \ldots + \gamma_r W_{ri} + v_i$$

- The Instrument exogeneity condition is now conditional on the included regressors $W_{1i}, \ldots, W_{ri}$

$$Cov\left(Z_i, u_i | W_{1i}, \ldots, W_{ri}\right) = 0$$

- In the returns to education example we will include the following control variables:
  - age and age squared
  - *south* equals 1 if an individuals lives in the southern part of the U.S.
  - *smsa* equals 1 if an individual lives in a Standard Metropolitan Statistical Area

Application: estimating the returns to education

Control variables must also be included in the first stage regression:

```
1 . regress education near_college age age2 south smsa, robust

  Linear regression                                     Number of obs =      3010
                                                           F(  5,  3004) =     40.82
                                                           Prob > F      =    0.0000
                                                           R-squared     =    0.0710
                                                           Root MSE      =    2.5822

  ------------------------------------------------------------------------------------
               |               Robust
     education |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------------
  near_college |   .3567396   .1117581     3.19   0.001      .1376095    .5758696
           age |   1.077846   .3044035     3.54   0.000      .4809854    1.674706
          age2 |  -.0189181   .0052999    -3.57   0.000     -.0293099   -.0085264
         south |  -.8953645   .0987761    -9.06   0.000     -1.08904    -.7016888
          smsa |   .7962275   .1156382     6.89   0.000      .5694895    1.022965
         _cons |  -2.349802   4.329293    -0.54   0.587     -10.83848    6.138875
  ------------------------------------------------------------------------------------

2 . test near_college

   ( 1)  near_college = 0

         F(  1,  3004) =    10.19
              Prob > F =    0.0014
```

Don’t use the overall F-statistic, this also tests whether the coefficients on the control variables equal zero!

Application: estimating the returns to education

IV estimates with control variables

```
. ivregress 2sls ln_wage (education=near_college) age age2 south smsa, robust

Instrumental variables (2SLS) regression               Number of obs =      3010
                                                       Wald chi2(    5)  =  757.69
                                                       Prob > chi2   =    0.0000
                                                       R-squared     =    0.1510
                                                       Root MSE      =    .40884
```

ln_wage	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
education	.0954681	.0481396	1.98	0.047	.0011163	.1898199
age	.0815643	.0702011	1.16	0.245	-.0560274	.2191559
age2	-.0007088	.0012218	-0.58	0.562	-.0031034	.0016859
south	-.1277804	.0478661	-2.67	0.008	-.2215962	-.0339646
smsa	.1038856	.0472	2.20	0.028	.0113752	.1963959
_cons	3.246947	.7048721	4.61	0.000	1.865423	4.628471

```
Instrumented:  education
Instruments:   age age2 south smsa near_college
```

- Estimated return to an additional year of education is now 9.5%
- Do we believe that instrument exogeneity holds now that we have included control variables?

1 endogenous regressor, multiple instruments

- Instead of 1 instrument we can also use $M > 1$ instruments
- We could calculate M different IV-estimates of β
- Since any linear combination of the Z_{mi} is again a valid instrument:
  - combine the Z_{mi} to get a more efficient estimator of β_1

$$Y_i = \beta_0 + \beta_1 X_i + \delta_1 W_{1i} +, \ldots, +\delta_r W_{ri} + u_i$$

$$X_i = \pi_0 + \pi_1 Z_{1i} + \ldots \pi_M Z_{Mi} + \gamma_1 W_{1i} + \ldots + \gamma_r W_{ri} + v_i$$

- Instrumental variable assumptions:

1. **Instrument relevance:** at least one of the instruments $Z_{1i}, \ldots, Z_{Mi}$ should have a nonzero coefficient in the population regression of X_i on $Z_{1i}, \ldots, Z_{Mi}$.
2. **Instrument exogeneity:** $Cov(Z_{1i}, u_i) = Cov(Z_{2i}, u_i) = \ldots = Cov(Z_{Mi}, u_i) = 0$

Application: estimating the returns to education

- The data set contains two potential instruments for years of education:

near_2yrcollege= 1 if individual grew up in area with a 2-year college
0 if individual grew up in area without a 2-year college

near_4yrcollege= 1 if individual grew up in area with a 4-year college
0 if individual grew up in area without a 4-year college

- To check for instrument relevance we should estimate the first stage regression, including both instruments
- And use an F-test to test for the joint significance of the two instruments.

Application: estimating the returns to education

```
Linear regression                               Number of obs =       3010
                                                  F(  6,  3003) =     34.03
                                                  Prob > F      =    0.0000
                                                  R-squared     =    0.0710
                                                  Root MSE      =    2.5827
```

education	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
near_4yrcollege	.3573365	.1121497	3.19	0.001	.1374385	.5772345
near_2yrcollege	-.0110908	.0976786	-0.11	0.910	-.2026145	.1804329
age	1.077147	.3045554	3.54	0.000	.4799884	1.674305
age2	-.0189051	.0053029	-3.57	0.000	-.0293028	-.0085074
south	-.8964387	.0991639	-9.04	0.000	-1.090875	-.7020027
smsa	.797801	.1167322	6.83	0.000	.5689179	1.026684
_cons	-2.336789	4.331927	-0.54	0.590	-10.83063	6.157055

```
2 . test near_4yrcollege=near_2yrcollege=0

 ( 1)  near_4yrcollege - near_2yrcollege = 0
 ( 2)  near_4yrcollege = 0

       F(  2,  3003) =     5.09
            Prob > F =    0.0062
```

- The first-stage F-statistic is well below 10, which indicates that we have weak instrument problems!
- It is better to drop the weakest instrument, *near_2yrcollege,* and use only 1 instrument *near_4yrcollege*

Overidentifying restrictions test (Sargan test, J-test)

- With more instruments than endogenous regressors we can test whether a subset of the instrument exogeneity conditions is valid.
- Suppose we have two instruments. Given our structural equation

$$Y_i = \beta_0 + \beta_1 X_i + \delta_1 W_{1i} +, \ldots, +\delta_r W_{ri} + u_i$$

  and assuming that $Cov(Z_{1i}, u_i) = 0$ we can test whether $Cov(Z_{2i}, u_i) = 0$ (or vice versa, but not both!)

- Intuition is as follows:
  - since $Cov(Z_{1i}, u_i) = 0 : \hat{\beta}_{2SLS}^{(Z_1)} \rightarrow \beta_1$
  - IF $Cov(Z_{2i}, u_i) = 0$ then also $\hat{\beta}_{2SLS}^{(z_2)} \rightarrow \beta_1$
- Testing whether $Cov(Z_{2i}, u_i) = 0$ is equivalent to testing $\hat{\beta}_{2SLS}^{(z_2)} = \hat{\beta}_{2SLS}^{(z_1)}$

Overidentifying restrictions test (Sargan test, J-test)

We can implement the test is as follows

1. Estimate $Y_i = \beta_0 + \beta_1 X_i + \delta_1 W_{1i} +, \ldots, +\delta_r W_{ri} + u_i$ by 2SLS using Z_{1i} and Z_{2i} as instruments
2. Obtain the residuals $\hat{u}_i^{2SLS} = Y_i - \widehat{\beta}_0 + \widehat{\beta}_1 X_i + \widehat{\delta}_1 W_{1i} +, \ldots, +\widehat{\delta}_r W_{ri}$
   - Note: use the true X_i and not the predicted value $\widehat{X}_i$
3. Estimate the following regression
$$\hat{u}_i^{2SLS} = \eta_0 + \eta_1 \cdot Z_{1i} + \eta_2 \cdot Z_{2i} + +\varphi_1 W_{1i} +, \ldots, +\varphi_r W_{ri} + e_i$$
4. And obtain the F-statistic of the test
$$H_0 : \eta_1 = \eta_2 = 0 \qquad versus \qquad H_0 : \eta_1 \neq 0 \; and/or \; \eta_2 \neq 0$$
5. Compute the J-test statistic
$$J = mF \sim \chi^2_q$$
   where q is number of instruments minus number of endogenous regressors (in this case 1)

Application: estimating the returns to education

```
1 . ivregress 2sls ln_wage (education=near_4yrcollege near_2yrcollege) age age2 south smsa,

  Instrumental variables (2SLS) regression               Number of obs =       3010
                                                          Wald chi2(    5)   =    766.83
                                                          Prob > chi2    =      0.0000
                                                          R-squared      =      0.1609
                                                          Root MSE       =      .40646

  ------------------------------------------------------------------------------------
                |               Robust
        ln_wage |      Coef.   Std. Err.       z     P>|z|      [95% Conf. Interval]
  --------------+---------------------------------------------------------------------
      education |   .0927438   .0477741     1.94   0.052     -.0008916     .1863792
            age |   .0844422   .0696594     1.21   0.225     -.0520878     .2209722
           age2 |  -.0007592   .0012123    -0.63   0.531     -.0031353     .0016169
          south |  -.1303678   .0475011    -2.74   0.006     -.2234683    -.0372672
           smsa |     .10638   .0468341     2.27   0.023      .0145869     .1981731
          _cons |   3.241778   .7006403     4.63   0.000      1.868548     4.615008
  ------------------------------------------------------------------------------------
  Instrumented:   education
  Instruments:    age age2 south smsa near_4yrcollege near_2yrcollege

2 . predict residuals, resid
```

Application: estimating the returns to education

```
1 . regress residuals near_4yrcollege near_2yrcollege age age2 south smsa, robust

Linear regression                                      Number of obs =      3010
                                                          F(  6,  3003) =      0.42
                                                          Prob > F      =    0.8684
                                                          R-squared     =    0.0008
                                                          Root MSE      =    .40676

------------------------------------------------------------------------------------
                 |               Robust
       residuals |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-----------------+------------------------------------------------------------------
 near_4yrcollege |   -.0003358   .0170653    -0.02   0.984    -.0337967    .0331252
 near_2yrcollege |    .0242942   .0154024     1.58   0.115    -.0059061    .0544946
             age |    .0015897   .0486995     0.03   0.974     -.093898    .0970775
            age2 |   -.0000297   .0008437    -0.04   0.972    -.0016839    .0016245
           south |     .002501    .015634     0.16   0.873    -.0281535    .0331555
            smsa |    -.003772   .0174362    -0.22   0.829    -.0379601    .0304162
           _cons |   -.0297385   .6960319    -0.04   0.966    -1.394486    1.335009
------------------------------------------------------------------------------------

2 . test near_4yrcollege=near_2yrcollege=0

 ( 1)  near_4yrcollege - near_2yrcollege = 0
 ( 2)  near_4yrcollege = 0

       F(  2,  3003) =      1.24
            Prob > F =    0.2882
```

- $J = mF = 2 \cdot 1.24 = 2.48$
- $2.48 < 2.71$ (critical value of χ^2_1 at 10% significance level)
- So we do not reject the null hypothesis of instrument exogeneity.

Overidentifying restrictions test (Sargan test, J-test)

- Can we conclude that the two instruments satisfy instrument exogeneity? **NO!**
- Although the J-test seems a useful test there are 3 reasons to be very careful when using this test in practice

1. When we don't reject the null hypothesis this does not mean that we can accept it!
2. The power of the J-test can be low (probability of rejecting when H_o does not hold)
3. The J-test tests the joint hypothesis of instrument validity and correct functional form
    1. if the test rejects, the instruments might be valid but the functional form is wrong
    2. if the test rejects, the instruments might be valid but the effect of the regressor of interest is heterogeneous $\beta_{1i} \neq \beta_1$

The general IV regression model

- So far we considered the case with 1 endogenous variable, but we can extend the model to multiple endogenous variables

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_K X_{Ki} + \delta_1 W_{1i} +, \ldots, +\delta_r W_{ri} + u_i$$

$$X_{1i} = \pi_0^1 + \pi_1^1 Z_{1i} + \ldots + \pi_M^1 Z_{Mi} + \gamma_1^1 W_{1i} +, \ldots, +\gamma_r^1 W_{ri} + v_i^1$$
$$\vdots$$
$$X_{Ki} = \pi_0^K + \pi_1^K Z_{1i} + \ldots + \pi_M^K Z_{Mi} + \gamma_1^K W_{1i} +, \ldots, +\gamma_r^K W_{ri} + v_i^K$$

- The general IV regression model has 4 types of variables

1. The dependent variable Y_i
2. K (possibly) endogenous regressors $X_{1i}, \ldots, X_{Ki}$
3. r control variables $W_{1i}, \ldots, W_{ri}$ (not the variables of interest)
4. M instrumental variables $Z_{1i}, \ldots, Z_{Mi}$

The general IV regression model

- When there are multiple endogenous regressors the 2SLS algoritm is similar except that each endogenous regressor requires its own first stage.
- For IV regression to be possible there should be at least as many instruments as endogenous regressors
- The model is said to be

  Underidentified if $M < K$, we cannot estimate the model, the number of instruments is then smaller that the number of endogenous regressors

  Exactly identified if $M = K$, the number of instruments equals the number of endogenous regressors

  Overidentified if $M > K$, the number of instruments exceeds the number of endogenous regressors

The general IV regression model

Assumptions of the general IV-model

1. Instrument exogeneity:

$$Cov(Z_{1i}, u_i) = Cov(Z_{2i}, u_i) = \ldots = Cov(Z_{Mi}, u_i) = 0$$

2. Instrument relevance:

   - for each endogenous regressor $X_{1i}, \ldots, X_{Ki}$, at least one of the instruments $Z_{1i}, \ldots, Z_{Mi}$ should have a nonzero coefficient in the population regression of the endogenous regressor on the instruments.
   - The predicted values and the control variables $(\widehat{X}_{1i}, \ldots, \widehat{X}_{Ki}, W_{1i}, \ldots, W_{ri}, 1)$ should not be perfectly multicollinear.

3. $(X_{1i}, \ldots, X_{Ki}, W_{1i}, \ldots, W_{ri}, Z_{1i}, \ldots, Z_{Mi}, Y_i)$ should be iid draws from their joint distribution.

4. Large outliers are unlikely: the $X's$, $W's$, $Z's$ and Y have finite fourth moments.

Application: estimating the returns to education

Summary of results using college proximity as instrument:

	OLS	1 IV without controls	1 IV with controls	2 IV's with controls
IV results, log(earnings) as dependent variable				
Education	0.052***	0.188***	0.095**	0.093*
	(0.003)	(0.021)	(0.048)	(0.048)
First stage regression				
near 4yr college		0.829***	0.357***	0.357***
		(0.107)	(0.112)	(0.112)
near 2yr college				-0.011
				(0.098)
First stage F		60.37	10.19	5.09

* significant at 10%, ** significant at 5%, *** significant at 1%

- Is college proximity a valid instrument?

Application: estimating the returns to education

- Another possible instrument for education is compulsory schooling laws
- Between 1925 and 1970 there were quite some changes in the minimum school leaving age in the US
  - these changes varied between states
- Oreopoulos (AER,2006) uses variation in minimum school leaving age as instrument for years of schooling
- Main assumptions
  - Changes in minimum school leaving age uncorrelated with unobserved variables affecting education (such as ability)
  - No direct effect of changes in minimum school leaving age on wages
  - Minimum school leaving age has a nonzero impact of years of education

Estimating returns to education

- Oreopoulos estimates the following first stage and second stage equations:

$$Y_{ist} = \beta X_{ist} + \gamma_s + \gamma_t + V'_{ist}\theta + W'_{st}\lambda + \varepsilon_{ist}$$

$$X_{ist} = \pi Z_{st} + \delta_s + \delta_t + V'_{ist}\rho + W'_{st}\kappa + \mu_{ist}$$

- Y_{ist} is log wage of individual i living in state s in year t at age 14
- X_{ist} is years of schooling of individual i living in state s in year t at age 14
- Z_{st} is the minimum school leaving age in state s in year t
- γ_s and δ_s are state fixed effects, γ_t and δ_t are year fixed effects
- V'_{ist} are individual characteristics and W'_{st} are state characteristics

Estimating returns to education

Results from Oreopoulos (2006)

	OLS	First stage	IV
	Earnings	Education	Earnings
Years of education	0.078***		0.142***
	(0.0005)		(0.012)
Minimum school leaving age		0.110***	
		(0.007)	

- First stage F-statistic: $F_{first} = t^2 = \left(\frac{0.110}{0.007}\right)^2 = 246.9$
- IV estimate almost twice as high as OLS estimate, not what we expect on basis of positive ability bias story
- Possible explanations:
  - downward bias in OLS due to measurement error
  - heterogeneity in the returns to education (IV estimates local average treatment effect)