ECON 21110
Applied Microeconometrics
Winter 2022
Lecture 4
Instrumental Variables (IV)

Eyðfríð Juanna Schrøter Joensen

University of Chicago

February 1, 2022

This Lecture

- In Lecture 3, we showed how randomized experiments help us construct credible counterfactuals to do causal inference
  - But randomized experiments are often not feasible
- In Lecture 2, we discussed how we could include control variables using multiple regression
  - But often not feasible to measure *all* unobserved confounding variables
- Therefore, we have to consider alternative econometric methods
- In this lecture, we consider one such core econometric method: Instrumental Variables (IV)
  - **Wooldridge (2016) Chapter 15**-16
  - Angrist & Pischke (2015) Chapter 3
  - Angrist & Pischke (2009) Chapter 4

Introduction

- Instrumental Variables (IV) used when model has endogenous X
- That is, when $\mathbb{E}[U \mid X] \neq 0$
  - the OLS estimator ($\widehat{\beta}_{OLS}$) is biased and inconsistent
    the OLS estimator cannot be given a causal interpretation
- Cases where the assumption of $\mathbb{E}[U \mid X] = 0$ does not hold are:
  - omitted variables bias
  - selection bias
  - simultaneity of Y and X; i.e. X causes Y, Y causes X
  - errors-in-variables (CEV, classical measurement error in X)

Measurement error | Measurement error in X | Measurement error in X, CEV

Introduction

- Instrumental Variable (IV) estimators are consistent IF valid instruments exist
- That is, Instrumental Variables Z that are correlated with the regressor X, but uncorrelated with the error term, U
- Practically, it can be difficult to obtain valid instruments

Endogeneity

- An endogenous variable is one that is correlated with U
- An exogenous variable is one that is uncorrelated with U
- Example:

$$wage = \beta_0 + \beta_1 educ + U$$

- This simple regression model assumes that $\mathbb{E}[U \mid X] = 0$ and that the only effect of X on Y is through the term βX
  - $X \rightarrow Y$
  - $U \rightarrow Y$
  - $U \nrightarrow X$
- The OLS estimator $\widehat{\beta}_{OLS}$ is unbiased and consistent for β
- In the simple case, U includes *all* factors other than years of education that determine wages

Endogeneity

- BUT what about ability? Might induce a correlation between X and U
  - $X \rightarrow Y$
    $\uparrow$
  - $U \rightarrow Y$
  - X is endogenous $\rightarrow$ people with high ability are likely to have high education, i.e. high U, high X, and high Y
  - $\mathbb{E}[U \mid X] \neq 0$ implies that $Cov(U, X) \neq 0$ and $\widehat{\beta}_{OLS}$ is inconsistent $\rightarrow$ **endogeneity bias**
- One solution to this problem would be to add a regressor that controls for ability (Lecture 2)
- An alternative approach is IV:
  we must find an instrument for the endogenous variable X that is uncorrelated with the error term U, but highly correlated with X
- The identification problem: what will allow us to identify the *causal* effect of X on Y (e.g. the effect of schooling on wages)?

Endogeneity: IV Assumptions

- We must find an IV that satisfies two properties:

$$Cov(Z, U) = 0 \qquad (1: \text{ Exogeneity})$$

and

$$Cov(Z, X) \neq 0 \qquad (2: \text{ Relevance})$$

- If these conditions are fulfilled using Z as an *instrument* for X gives us a consistent estimate of β
- An instrument that is both *exogenous* and *relevant* (or *strong*) is called **valid**
- The Instrumental Variable detects movements in X that are uncorrelated with U, and uses these to estimate β

Endogeneity

We can test relevance, but not exogeneity

- We have to use common sense and economic theory to decide if it makes sense to assume $Cov(U, Z) = 0$
- The other correlation, $Cov(X, Z) \neq 0$, we can test in "first-stage" regression

$$\begin{array}{ccccc} Z & \rightarrow & X & \rightarrow & Y \\ & & \uparrow & & \\ & & U & \rightarrow & Y \end{array}$$

e.g. Z=college proximity (Card, 1993)

IV with one regressor and one instrument

$$Y = \beta X + U$$

- X is endogenous and correlated with the error term $Cov(U, X) \neq 0$
- We have an instrument, Z, which is *not* correlated with the error term $Cov(U, Z) = 0$
- Covariance between Y and Z:

$$\begin{aligned} Cov(Y, Z) &= Cov(\beta X + U, Z) \\ & \beta Cov(X, Z) + Cov(U, Z) \end{aligned}$$

- $Cov(U, Z) = 0$ (instrument exogeneity)

IV with one regressor and one instrument

$$\beta = \frac{Cov(Y, Z)}{Cov(X, Z)}$$

- A consistent estimate $\widehat{\beta}_{IV}$ of β can be obtained by replacing the covariances by their sample analogues:

$$\widehat{\beta}_{IV} = \frac{\widehat{Cov}(Y, Z)}{\widehat{Cov}(X, Z)} = \frac{\sum(Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum(X_i - \overline{X})(Z_i - \overline{Z})}$$

- This model is exactly identified since we have exactly one IV for exactly one endogenous variable
- IV is often obtained by 2SLS

IV with one regressor and one instrument

The 2SLS method has two stages:

(1) **First-stage** regression: Isolate the part of X that is correlated with the error term $\rightarrow$ regress X on Z using OLS

$$X = \pi Z + \upsilon$$

which gives us:

$$\widehat{\pi} = \frac{\widehat{Cov}(Z, X)}{\widehat{Var}(Z)}$$

use $\widehat{\pi}$ to compute the predicted value of X:

$$\widehat{X} = \widehat{\pi} Z$$

- the first stage basically decomposes X into two components:
  - One component υ that is correlated with the error term U
  - Another component πZ that is uncorrelated with U since Z is exogenous

IV with one regressor and one instrument

Main idea of IV is to only use the "exogenous" (or "problem-free") component of X to estimate β

(2) The second stage consists of replacing X by $\widehat{X}$ in the main regression:

$$Y = \beta \widehat{X} + U$$

$$\begin{aligned}
\widehat{\beta}_{2SLS} &= \frac{\widehat{Cov}(Y, \widehat{X})}{\widehat{Var}(\widehat{X})} = \frac{\widehat{Cov}(Y, \widehat{\pi} Z)}{\widehat{Var}(\widehat{\pi} Z)} = \frac{\widehat{\pi}\,\widehat{Cov}(Y, Z)}{\widehat{\pi}^2\,\widehat{Var}(Z)} \\
&= \frac{\widehat{Cov}(Y, Z)}{\widehat{\pi}\,\widehat{Var}(Z)} = \frac{\widehat{Cov}(Y, Z)}{\frac{\widehat{Cov}(Z,X)}{\widehat{Var}(Z)}\widehat{Var}(Z)} = \frac{\widehat{Cov}(Y, Z)}{\widehat{Cov}(Z, X)} = \widehat{\beta}_{IV}
\end{aligned}$$

IV with one regressor and one instrument

- We can also rewrite this as

$$\widehat{\beta}_{IV} = \beta + \frac{\widehat{Cov}(Z, U)}{\widehat{Cov}(X, Z)}$$

- The sample covariance is a consistent estimator of the population covariance $\Rightarrow$ the IV estimator is consistent:

$$\operatorname*{plim}_{n \to \infty} \widehat{\beta}^{IV} = \beta$$

- $\widehat{\beta}^{IV}$ is inconsistent when $Cov(U, Z) \neq 0$
  - $\widehat{\beta}^{OLS}$ is inconsistent if $Cov(U, X) \neq 0$
- In large samples, the IV estimator is normally distributed
  - inference such as hypothesis tests and confidence interval are conducted as before

IV with one regressor and one instrument

- The standard errors differ between OLS and IV
- The stronger the correlation between Z and X, the smaller the IV standard errors
- An important cost of doing IV when a regressor is not endogenous is that the asymptotic variance of $\widehat{\beta}_{IV}$ is always larger than the corresponding $\widehat{\beta}_{OLS}$ estimator
  - IV estimates will always have larger standard errors than the corresponding OLS estimates

IV and omitted variable bias

$$Y_i = \beta_0 + \beta_1 X_{1i} + U_i$$

- X_{2i} is an omitted variable in the regression and is part of the error term

$$U_i = \beta_2 X_{2i} + \upsilon_i$$

- We find an instrument (Z) that fulfills the conditions:
  - $Cov(U, Z) = 0$ (thus, uncorrelated with the omitted variable X_{2i})
  - $Cov(X_1, Z) \neq 0$ (can be tested)

IV and omitted variable bias

$$
\begin{aligned}
Cov(Y_i, Z_i) &= Cov(\beta_0 + \beta_1 X_{1i} + U_i, Z_i) \\
& \quad Cov(X_{1i}, Z_i) + \underbrace{Cov(U_i, Z_i)}_{=0} \\
\beta_1 &= \frac{Cov(Y_i, Z_i)}{Cov(Z_i, X_i)}
\end{aligned}
$$

- $\widehat{\beta}_1$ is a consistent estimator of the true β_1

IV Jargon: Reduced Form vs. IV Estimation

- The structural equation (causal relationship of interest) is:

$$Y = \beta_0 + \beta_1 X + U$$

- The reduced form equation for the outcome is:

$$Y = \gamma_0 + \gamma_1 Z + V$$

- The reduced form equation for the first-stage is:

$$X = \pi_0 + \pi_1 Z + \varepsilon$$

- 2SLS "second-stage" is:

$$Y = \beta_0 + \beta_1 \widehat{X} + \upsilon$$

Reduced Form vs. IV

- We have that:

$$\beta_1 = \frac{Cov(Z, Y)}{Cov(Z, X)}$$

$$\gamma_1 = \frac{Cov(Z, Y)}{Var(Z)}$$

$$\pi_1 = \frac{Cov(Z, X)}{Var(Z)}$$

- Note that the only difference between γ_1 (reduced form effect) and β_1 (the structural parameter) lies in the denominator

Reduced Form vs. IV

- From the previous slide, we also see that:

$$\frac{\gamma_1}{\pi_1} = \frac{\frac{Cov(Z,Y)}{Var(Z)}}{\frac{Cov(Z,X)}{Var(Z)}} = \frac{Cov(Z,Y)}{Cov(Z,X)} = \beta_1$$

- The IV-estimate (β_1) is thus equal to the ratio of the reduced-form estimate for the outcome (γ_1) and the first stage (π_1)
- We can think of the first stage estimate (π_1) as "scaling" the reduced form estimate for the outcome (γ_1) so as to obtain β_1

Reduced Form vs. IV

- When estimating the structural equation, we assume that the instrument (Z) only affects the outcome (Y) via its effect on the endogenous variable (X)
- To see this, note that $\gamma_1 = \pi_1 \beta_1$
- That is, the reduced form effect on the outcome (γ_1) reflects the effect of the instrument on the endogenous variable (π_1) multiplied by the effect of the endogenous variable on the outcome (β_1)

Reduced Form vs. IV

- The differences in scaling means that γ_1 and β_1 answer different questions
- Typically, we are more interested in the structural parameter, but not always
- For example, suppose we have a certain policy (Z) that functions as an instrument for education (X). Then we may be interested both in knowing the effect of education on, say, wages (β_1), but also the direct effect of the policy (γ_1)

IV and STATA

- `ivreg y (x=z) controls`
- Example: study the causal link between education and wages
  - Problem: need some exogenous variation in education outcome
  - Instrument: college proximity as an exogenous determinant of schooling
  - Data: NLS Young men cohort
  - First applied by David Card (1993)

IV and STATA: data description

```
. describe id lwage educ exper expersq black south smsa fatheduc motheduc nearc2
nearc4

              storage   display    value
variable name   type    format     label      variable label
-----------------------------------------------------------------------------------
-----------------------------------------------------------------------
id              int     %9.0g                 person identifier
lwage           float   %9.0g                 log(wage)
educ            byte    %9.0g                 years of schooling, 1976
exper           byte    %9.0g                 age - educ - 6
expersq         int     %9.0g                 exper^2
black           byte    %9.0g                 =1 if black
south           byte    %9.0g                 =1 if in south, 1976
smsa            byte    %9.0g                 =1 in in SMSA, 1976
fatheduc        byte    %9.0g                 father's schooling
motheduc        byte    %9.0g                 mother's schooling
nearc2          byte    %9.0g                 =1 if near 2 yr college, 1966
nearc4          byte    %9.0g                 =1 if near 4 yr college, 1966
```

IV and STATA: OLS regression

```
. reg lwage black south smsa exper expersq educ, robust

Linear regression                                      Number of obs =    3010
                                                       F(  6,  3003) =  217.74
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.2905
                                                       Root MSE      =  .37419

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.1896315   .0174324   -10.88   0.000    -.2238123   -.1554508
       south |  -.1248615   .0153508    -8.13   0.000    -.1549606   -.0947625
        smsa |    .161423   .0151751    10.64   0.000     .1316683    .1911776
       exper |   .0835958   .0067326    12.42   0.000     .0703948    .0967969
     expersq |  -.0022409   .0003181    -7.04   0.000    -.0028646   -.0016171
        educ |    .074009    .003642    20.32   0.000     .0668679    .0811501
       _cons |   4.733664   .0701577    67.47   0.000     4.596102    4.871226
------------------------------------------------------------------------------
```

IV and STATA: IV (2SLS) regression

```
. ivreg lwage black south smsa exper expersq (educ = nearc4), robust

Instrumental variables (2SLS) regression              Number of obs =    3010
                                                      F(  6,  3003) =  131.70
                                                      Prob > F      =  0.0000
                                                      R-squared     =  0.2252
                                                      Root MSE      =  .39103

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1322888   .0485779     2.72   0.007     .0370396    .2275381
       black |  -.1308019   .0515112    -2.54   0.011    -.2318027   -.0298011
       south |  -.1049005   .0229264    -4.58   0.000    -.1498535   -.0599475
        smsa |   .1313237    .029803     4.41   0.000     .0728872    .1897601
       exper |    .107498   .0211375     5.09   0.000     .0660525    .1489434
     expersq |  -.0022841   .0003467    -6.59   0.000    -.0029639   -.0016042
       _cons |   3.752781   .8177012     4.59   0.000      2.14947    5.356092
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   black south smsa exper expersq nearc4
------------------------------------------------------------------------------
```

IV and STATA: IV (2SLS) regression

- The slope estimate in IV is larger than the slope estimate in OLS
- We thought ability bias would make the OLS coefficient too large?
- the IV estimate is "too big" – few would argue that there is a 20% return to a year of education
- Two possiblities:
  1. the instrument is not valid (it is correlated with the error)
  2. the first stage is weak, which inflates the IV estimate

The general IV regression model

- We can extend to the multiple regression case where we have several included endogenous and exogenous variables as well as multiple instruments
- We need an instrument for each endogenous variable
- The model is called the structural model:

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

- $X_1, X_2, ..., X_k$: k endogenous variables (potentially correlated with U_i)
- $W_1, W_2, ..., W_r$: r exogenous variables (not correlated with U_i)
- $Z_1, Z_2, ..., Z_m$: m Instrumental Variables (excluded exogenous variables)

The general IV regression model

- Identification:
  - *exactly identified* ($m = k$) if the number of instruments, m, are equal to the number of endogenous variables, k. This means that there are just enough instruments to estimate the endogenous variables
  - *overidentified* if $m > k$. We now have more than enough instruments to estimate the endogenous variables. If this is the case, then one can test for "overidentifying restrictions"
  - *underidentified* if $m < k$. This means that there are too few instruments to estimate the endogenous parameters. In this case we must get hold of more instruments

The general IV regression model

- On top of the **order condition** $m \geq k$; i.e. at least as many instruments as endogenous variables, we also need:
  - Exogeneity: each of the instruments are uncorrelated with the error term conditional on *all* the other exogenous variables: $Cov(U, Z_j \mid Z_1, ..., Z_{j-1}, Z_{j+1}, ..., Z_m, W_1, ..., W_r) = 0$ for all $j \in \{1, ...m\}$
  - Relevance: the instruments must be partially correlated with the endogenous variable X once all other exogenous variables have been netted out (i.e. a **rank condition**)

IV and STATA: IV (2SLS) regression with multiple instruments

```
. ivreg lwage black south smsa exper expersq (educ = nearc2 nearc4), robust

Instrumental variables (2SLS) regression               Number of obs =    3010
                                                       F(  6,  3003) =  119.78
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1455
                                                       Root MSE      =  .41065

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1608487   .0485705     3.31   0.001     .0656139    .2560835
       black |  -.1019726   .0520797    -1.96   0.050    -.2040881    .0001429
       south |  -.0951187   .0234332    -4.06   0.000    -.1410654    -.049172
        smsa |   .1165736   .0302929     3.85   0.000     .0571767    .1759705
       exper |   .1192112   .0213279     5.59   0.000     .0773923      .16103
     expersq |  -.0023052   .0003691    -6.25   0.000    -.0030289   -.0015816
       _cons |   3.272102   .8178286     4.00   0.000     1.668541    4.875663
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   black south smsa exper expersq nearc2 nearc4
------------------------------------------------------------------------------
```

IV Example: Levitt (1997) Effect of Police on Crime

- What is the effect of increasing the *police force* on the *crime rate*?
- A classic example of simultaneous causality: high crime areas need large police forces
- This leads to an incorrectly-signed (positive) OLS coefficient

$$
\begin{aligned}
(i)\ Crime &= \beta_0 + \beta_1 Police + U \\
(ii)\ Police &= \alpha_0 + \alpha_1 Crime + \upsilon
\end{aligned}
$$

- β_1 reflects a correlation, not causation if we estimate (i) without taking (ii) into account
- We want to estimate β_1. If $\alpha_1 > 0$ and $\beta_1 > 0$, then we have an endogeneity problem
- This means that $Cov(Police, U) > 0$
  - a spurious positive relationship between *Police* and *Crime*
  - this type of endogeneity problem is labeled **reverse causality**
- For causality, we need a scenario in which more police are hired for reasons completely unrelated to rising crime

IV Example: Levitt (1997) Effect of Police on Crime

- Basic idea is the fact that a popular re-election strategy among US mayors is to hire more police as an election approaches
- Uses the timing of US mayor elections as an instrument Variable
- Is this instrument valid?
- **Relevance**: police force increases in election years
- **Exogeneity**: elections cycles are pre-determined so election cycles per se are probably unrelated to crime
- The empirical strategy is to use IV to isolate the effect of extra police on crime

IV Example: Levitt (1997) Effect of Police on Crime

- Two-stage least squares (2SLS)
- First stage:

$$Police = \alpha_0 + \alpha Election + \upsilon$$

- Second stage

$$Crime = \beta_0 + \beta_1 \widehat{Police} + U$$

where $\widehat{Police} = \widehat{\alpha} Election$ is the predicted value of police from the first-stage regression

IV Example: Levitt (1997) Effect of Police on Crime

- **Finding**: an increased police force reduces violent crime (but has little effect on property crime)
- There are lots of examples of correlations that are interpreted as causal relations
  - It is important to be critical and question the plausibility of estimated effects
  - Try to seek alternative explanations and see to which extent each of them can contribute to the finding

IV Example: Angrist (1990) Veteran Draft Lottery

- In the following we will often refer to an example from Angrist's paper on the effects of military service on earnings
- Angrist (1990) uses the Vietnam draft lottery as in IV for military service
- In the 1960s and early 1970s, young American men were drafted for military service to serve in Vietnam
- Concerns about the fairness of the conscription policy lead to the introduction of a draft lottery in 1970
- From 1970 to 1972 random sequence numbers were randomly assigned to each birth date in cohorts of 19-year-olds
- Men with lottery numbers below a cutoff were drafted while men with numbers above the cutoff could not be drafted
- The draft did not perfectly determine military service:
    - Many draft-eligible men were exempted for health and other reasons
    - Exempted men volunteered for service

IV Example: Angrist (1990) Veteran Draft Lottery
Summary of Findings

- First-stage results:
  - Having a low lottery number (being eligible for the draft) increases veteran status by about 16 percentage points (the mean of veteran status is about 27 percent)
- Second-stage results:
  - Serving in the military lowers earnings by between \$2,050 and \$2,741 per year of service

IV with Heterogeneous Treatment Effects

- Up to this point we only considered models where the causal effect was the same for all individuals (homogenous treatment effects): $Y_{1i} - Y_{0i} = \rho$ for all i
- We now try to understand the IV estimates if treatment effects are heterogeneous
- This will inform us about two types of validity characterizing research designs:
  1. **Internal validity**: Does the design successfully uncover causal effects for the population studied?
  2. **External validity**: Do the study's results inform us about different populations?

IV with Heterogeneous Treatment Effects

- Variables used in this setup:
  - $Y_{Di}(Z)$ = potential outcome of individual i
  - D_i = treatment dummy
  - Z_i = instrument dummy
- Causal chain is:
  $Z_i \rightarrow D_i \rightarrow Y_i$
- Notation for D_i :
  - D_{1i} = i's treatment status when $Z_i = 1$
  - D_{0i} = i's treatment status when $Z_i = 0$
- Observed treatment status is therefore:

$$D_i = D_{0i} + (D_{1i} - D_{0i})Z_i = \pi_0 + \pi_{1i}Z_i + \zeta_i$$

  $\pi_0 = \mathbb{E}[D_{0i}]$
  $\pi_1 = \mathbb{E}[D_{1i} - D_{0i}]$ is the heterogeneous causal effect of the IV on D_i
- The average causal effect of Z_i on D_i is $\mathbb{E}[\pi_{1i}]$

Key Assumptions in the Heterogeneous Effects Framework

1. **Independence assumption:**

- The IV is independent of the vector of potential outcomes and potential treatment assignments (i.e. as good as randomly assigned): $\{Y_{Di}(1), Y_{Di}(0), D_{1i}, D_{0i}\} \perp Z_i$
- The independence assumption is sufficient for a causal interpretation of the reduced form:
  $\mathbb{E}[Y_i|Z_i = 1] - \mathbb{E}[Y_i|Z_i = 0]$
  $= \mathbb{E}[Y_{Di}(1)|Z_i = 1] - \mathbb{E}[Y_{Di}(0)|Z_i = 0]$
  $= \mathbb{E}[Y_{Di}(1)] - \mathbb{E}[Y_{Di}(0)]$
- Independence also means that the first stage captures the causal effect of Z_i on D_i:
  $\mathbb{E}[D_i|Z_i = 1] - \mathbb{E}[D_i|Z_i = 0]$
  $= \mathbb{E}[D_{1i}|Z_i = 1] - \mathbb{E}[D_{0i}|Z_i = 0]$
  $= \mathbb{E}[D_{1i} - D_{0i}]$

Key Assumptions in the Heterogeneous Effects Framework

2. **Exclusion restriction:**

- $Y_{Di}(Z)$ is a function of D only. Formally: $Y_{Di}(0) = Y_{Di}(1) = Y_{Di}$ for $D = 0, 1$
- In the Vietnam draft lottery example: an individual's earnings potential as a veteran or non-veteran are assumed to be unchanged by draft eligibility status
- The exclusion restriction would be violated if low lottery numbers may have affected schooling (e.g. to avoid the draft). If this was the case the lottery number would be correlated with earnings for at least two cases:
  1. through its effect on military service
  2. through its effect on educational attainment
- The fact that the lottery number is randomly assigned (and therefore satisfies the independence assumption) does not ensure that the exclusion restriction is satisfied

Key Assumptions in the Heterogeneous Effects Framework

- Using the exclusion restriction we can define potential outcomes indexed solely against treatment status:
  $Y_{1i} = Y_{1i}(1) = Y_{1i}(0)$
  $Y_{0i} = Y_{0i}(1) = Y_{0i}(0)$
- In terms of potential outcomes we can write:
  $Y_i = Y_{0i}(Z_i) + [Y_{1i}(Z_i) - Y_{0i}(Z_i)]D_i$
  $Y_i = Y_{0i} + [Y_{1i} - Y_{0i}]D_i$
- Random coefficients notation for this is:
  $Y_i = \alpha_0 + \rho_i D_i$
  with $\alpha_0 = \mathbb{E}[Y_{0i}]$and $\rho_i = Y_{1i} - Y_{0i}$

Key Assumptions in the Heterogeneous Effects Framework

3. **First Stage:**

- We need that the instrument has to have a significant effect on treatment:
  $\mathbb{E}[D_{1i} - D_{0i}] \neq 0$

Key Assumptions in the Heterogeneous Effects Framework

4. **Monotonicity:**

- Either $\pi_{1i} \geq 0$ for all i or $\pi_{1i} \leq 0$ for all i
  - While the instrument may have no effect on some people, all those who are affected are affected in the same way
  - In the draft lottery example: draft eligibility may have had no effect on the probability of military service. But there should also be no one who was kept out of the military by being draft eligible
    $\rightarrow$ this is likely satisfied
- Without monotonicity, IV estimators are not guaranteed to estimate a weighted average of the underlying causal effects of the affected group, $Y_{1i} - Y_{0i}$

What does IV estimate under these 4 Assumptions?

- These conditions imply the **Local Average Treatment Effect** ($LATE$) parameter:

$$
\begin{aligned}
& \mathbb{E}[Y \mid Z=1] - \mathbb{E}[Y \mid Z=0] \\
= \;& \mathbb{E}[Y_0 + D(Y_1 - Y_0) \mid Z=1] - \mathbb{E}[Y_0 + D(Y_1 - Y_0) \mid Z=0] \\
= \;& \mathbb{E}[D(Y_1 - Y_0) \mid Z=1] - \mathbb{E}[D(Y_1 - Y_0) \mid Z=0] \\
= \;& \mathbb{E}[(D_1 - D_0)(Y_1 - Y_0)] \quad \text{(Independence)} \\
= \;& \begin{array}{l} \mathbb{E}[Y_1 - Y_0 \mid D_1 - D_0 = 1]\, P(D_1 - D_0 = 1) \\ -\mathbb{E}[Y_1 - Y_0 \mid D_1 - D_0 = -1]\, P(D_1 - D_0 = -1) \end{array} \quad \text{(LIE)} \\
= \;& \begin{array}{l} \mathbb{E}[Y_1 - Y_0 \mid D_1 = 1, D_0 = 0]\, P(D_1 = 1, D_0 = 0) \\ -\mathbb{E}[Y_1 - Y_0 \mid D_1 = 0, D_0 = 1]\, P(D_1 = 0, D_0 = 1) \end{array}
\end{aligned}
$$

What does IV estimate under these 4 Assumptions?

- Invoke monotonicity condition
- Suppose that $P(D_1 = 0, D_0 = 1) = 0$. Then,

$$\begin{aligned} & \mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0] \\ = \; & \mathbb{E}[Y_1 - Y_0 \mid D_1 = 1, D_0 = 0]\, P(D_1 = 1, D_0 = 0) \end{aligned}$$

- The Local Average Treatment Effect is then given by:

$$\begin{aligned} LATE &= \frac{\mathbb{E}[Y \mid Z = 1] - \mathbb{E}[Y \mid Z = 0]}{P(D = 1 \mid Z = 1) - P(D = 1 \mid Z = 0)} \\ &= \mathbb{E}[Y_1 - Y_0 \mid D_1 = 1, D_0 = 0] \end{aligned}$$

- The average gain to those induced to switch from $D = 0$ to $D = 1$ by a change in Z from 0 to 1

What does IV estimate under these 4 Assumptions?

- Observe $LATE = ATE$ if $P(D = 1 \mid Z = 1) = 1$ while $P(D = 1 \mid Z = 0) = 0$; i.e. the whole population switches treatment choice and "complies" to the instrument
- "*Identification at infinity*" plays a crucial role throughout the entire literature on policy evaluation
- In general, $LATE \neq ATE = \mathbb{E}[Y_1 - Y_0]$
- Not treatment on the treated: $\mathbb{E}[Y_1 - Y_0 \mid D = 1]$
- Different instruments identify different parameters
- Having a wealth of different strong instruments does not improve the precision of the estimate of any particular parameter

IV Estimates LATE

- If all 4 assumptions are satisfied, IV estimates the LATE
- LATE is the average effect of X on Y for those whose treatment status has been changed by the instrument Z
- In the draft lottery example: IV estimates the average effect of military service on earnings for the subpopulation ("**compliers**") who enrolled in military service because of the draft, but would *not* have served otherwise
  - For example, this excludes volunteers ("**always-takers**") and men who were exempted from military service for medical reasons ("**never-takers**")
- We have reviewed the properties of IV with heterogeneous treatment effects using a dummy endogenous variable and no additional controls. The intuition of LATE generalizes to most cases where we have continuous endogenous variables and instruments, and additional control variables

Some LATE Framework Jargon

- The LATE framework partitions any population with an instrument into potentially 4 groups:
  1. **Compliers**: The subpopulation with $D_{1i} = 1$ and $D_{0i} = 0$.
     Their treatment status is affected by the instrument in the "right" direction
  2. **Always-takers**: The subpopulation with $D_{1i} = D_{0i} = 1$
     They always take the treatment independently of Z
  3. **Never-takers**: The subpopulation with $D_{1i} = D_{0i} = 0$
     They never take the treatment independently of Z
  4. **Defiers**: The subpopulation with $D_{1i} = 0$ and $D_{0i} = 1$.
     Their treatment status is affected by the instrument in the "wrong" direction
- These terms come from an analogy to the medical literature where the treatment, for example, is taking a pill

Monotonicity or (Uniformity)

- **Monotonicity** ensures that there are *no defiers*
- Why is it important to not have defiers?
  - If there were defiers, effects on compliers could be (partly) cancelled out by opposite effects on defiers
  - One could then observe a reduced form which is close to 0 even though treatment effects are positive for everyone, but the compliers are pushed in one direction by the instrument and the defiers in the other direction

What Does IV Estimate and What Not? LATE and ATE

- As outlined above, with all 4 assumptions satisfied IV estimates the average treatment effect for compliers
- Without further assumptions (e.g. constant causal effects) LATE is not informative about effects on never-takers or always-takers because the instrument does not affect their treatment status

Other Potentially Interesting Treatment Effects

- Another effect which we may potentially be interested in estimating is the average treatment effect on the treated (TT)
- LATE is not the same as TT

$$
\begin{aligned}
\underbrace{\mathbb{E}\left[Y_{1i}-Y_{0i}|D_i=1\right]}_{\text{Effect on the treated}} &= \underbrace{\mathbb{E}\left[Y_{1i}-Y_{0i}|D_i=1\right]}_{\text{Effect on always takers}} P\left[D_{0i}=1|D_i=1\right] \\
&\quad + \underbrace{E\left[Y_{1i}-Y_{0i}|D_{1i}>D_{0i}\right]}_{\text{Effect on compliers}} \\
&\quad \times P\left[D_{1i}>D_{0i}, Z_i=1|D_i=1\right]
\end{aligned}
$$

- The average treatment effect on the treated is a weighted average of the effects on always-takers and compliers
- If there are no always takers we can, however, estimate TT which is equal to LATE in that case

IV in Randomized Trials

- The use of IV methods can also be useful when evaluating a randomized trial
- In many randomized trials, participation is voluntary among those randomly assigned to treatment
- On the other hand people in the control group usually do not have access to treatment
  $\rightarrow$ only those who are particularly likely to benefit from treatment will actually take up treatment (leads almost always to positive selection bias
  $\rightarrow$ if you just compare means between treated and untreated individuals (using OLS) you will obtain biased treatment effects
- Solution: Instrument for treatment with whether you were offered treatment
  $\rightarrow$ you estimate LATE

Weak Instruments

- If you have weak instruments, IV is consistent but not unbiased
- For a long time researchers estimating IV models never cared much about the small sample bias
- In the early 1990s a number of papers, however, highlighted that IV can be severely biased in particular if instruments are weak (i.e. the first stage relationship is weak) and if you use many instruments to instrument for one endogenous variable (i.e. there are many overidentifying restrictions)
- In the worst case, if the instruments are so weak that there is no first stage the 2SLS sampling distribution is centered on the probability limit of OLS

Weak Instruments – Bias Towards OLS

- The OLS bias is: $\mathbb{E}[\widehat{\beta}^{OLS} - \beta] = \frac{Cov(X,U)}{Var(X)}$
- If U and ε are correlated, then the OLS bias is $\frac{\sigma_{\varepsilon U}}{\sigma_X^2}$
- It can be shown that the bias of 2SLS is approximately:
  $\mathbb{E}[\widehat{\beta}^{2SLS} - \beta] \approx \frac{\sigma_{\varepsilon U}}{\sigma_\varepsilon^2}\frac{1}{F+1}$
  where F is the population analogue of the F-statistic for the joint significance of the instruments in the first-stage regression
- If the first-stage is weak ($F \to 0$) the bias of 2SLS approaches $\frac{\sigma_{\varepsilon U}}{\sigma_\varepsilon^2}$
- This is exactly the OLS bias IF $\pi = 0$; i.e. "no first-stage", since this implies $\sigma_X^2 = \sigma_\varepsilon^2$ and therefore the OLS bias becomes $\frac{\sigma_{\varepsilon U}}{\sigma_\varepsilon^2}$
- If the first-stage is very strong ($F \to \infty$) the 2SLS bias goes to 0

Weak Instruments – Bias Towards OLS
Adding More Instruments

- Adding more weak instruments will increase the bias of 2SLS. By adding further instruments without predictive power the first stage F-statistic goes towards 0 and the bias increases

- If the model is just identified, weak instrument bias is less of a problem: approximately unbiased IF the first stage is not 0

- Bound, Jaeger & Baker (1995) highlighted this problem for the Angrist & Krueger (1991) study using different sets of instruments:
    - quarter of birth dummies $\rightarrow$ 3 instruments
    - quarter of birth + (quarter of birth) x (year of birth) dummies $\rightarrow$ 30 instruments
    - quarter of birth + (quarter of birth) x (year of birth) + (quarter of birth) x (state of birth) $\rightarrow$ 180 instruments

- Illustrate that adding more weak instruments reduced the first stage F-statistic and moves the coefficient towards the OLS coefficient

Weak Instruments – Bias Towards OLS
Adding More Instruments

Table 1. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings (standard errors of coefficients in parentheses)

	(1) OLS	(2) IV	(3) OLS	(4) IV	(5) OLS	(6) IV
Coefficient	.063 (.000)	.142 (.033)	.063 (.000)	.081 (.016)	.063 (.000)	.060 (.029)
F (excluded instruments)		13.486		4.747		1.613
Partial R^2 (excluded instruments, $\times$100)		.012		.043		.014
F (overidentification)		.932		.775		.725
Age Control Variables						
Age, Age^2	x	x			x	x
9 Year of birth dummies			x	x	x	x
Excluded Instruments						
Quarter of birth		x		x		x
Quarter of birth $\times$ year of birth				x		x
Number of excluded instruments		3		30		28

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), and 8 Regional dummies as control variables. *F* (first stage) and partial R^2 are for the instruments in the first stage of IV estimation. *F* (overidentification) is that suggested by Basmann (1960).

Weak Instruments – Bias Towards OLS
Adding More Instruments

Table 2. Estimated Effect of Completed Years of Education on Men's Log Weekly Earnings, Controlling for State of Birth (standard errors of coefficients in parentheses)

	(1) OLS	(2) IV	(3) OLS	(4) IV
Coefficient	.063	.083	.063	.081
	(.000)	(.009)	(.000)	(.011)
F (excluded instruments)		2.428		1.869
Partial R^2 (excluded instruments, $\times$100)		.133		.101
F (overidentification)		.919		.917
Age Control Variables				
Age, Age^2			x	x
9 Year of birth dummies	x	x	x	x
Excluded Instruments				
Quarter of birth		x		x
Quarter of birth $\times$ year of birth		x		x
Quarter of birth $\times$ state of birth		x		x
Number of excluded instruments		180		178

NOTE: Calculated from the 5% Public-Use Sample of the 1980 U.S. Census for men born 1930–1939. Sample size is 329,509. All specifications include Race (1 = black), SMSA (1 = central city), Married (1 = married, living with spouse), 8 Regional dummies, and 50 State of Birth dummies as control variables. F (first stage) and partial R^2 are for the instruments in the first stage of IV

Weak Instruments

- The exogeneity assumption that $Cov(Z, U) = 0$ is unlikely to be completely true
- What is the consequence of a small correlation between Z and U?
- When is OLS preferable?
- Let's first look at the probability limit for OLS; i.e. what $\widehat{\beta}_1^{OLS}$ converges to as sample size goes to infinity

Weak Instruments: Asymptotic Bias for OLS

- From Chapter 5, we know that:

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1^{OLS} = \beta_1 + \frac{Cov(X,U)}{Var(X)}$$

- Using the fact that:

$$\rho_{X,U} = \frac{Cov(X,U)}{\sigma_U \sigma_X} \;\Rightarrow\; Cov(X,U) = \rho_{X,U}\sigma_U\sigma_X$$

and the fact that $Var(X) = \sigma_X^2 = \sigma_X\sigma_X$, we get:

$$\begin{aligned}\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1^{OLS} &= \beta_1 + \frac{\rho_{X,U}\sigma_U\sigma_X}{\sigma_X\sigma_X} \\ &= \beta_1 + \rho_{X,U}\frac{\sigma_U}{\sigma_X}\end{aligned}$$

Weak Instruments: Asymptotic Bias for IV

- Similarly, for the IV estimator we have that:

$$\underset{n\to\infty}{\text{plim}}\ \widehat{\beta}_1^{IV} = \beta_1 + \frac{Cov(Z,U)}{Cov(Z,X)}$$

- Using the fact that:

$$\rho_{Z,U} = \frac{Cov(Z,U)}{\sigma_Z \sigma_U} \ \Rightarrow\ Cov(Z,U) = \rho_{Z,U}\sigma_Z\sigma_U$$
$$\rho_{Z,X} = \frac{Cov(Z,X)}{\sigma_Z \sigma_X} \ \Rightarrow\ Cov(Z,X) = \rho_{Z,X}\sigma_Z\sigma_X$$

we get:

$$\begin{aligned}\underset{n\to\infty}{\text{plim}}\ \widehat{\beta}_1^{IV} &= \beta_1 + \frac{\rho_{Z,U}\sigma_Z\sigma_U}{\rho_{Z,X}\sigma_Z\sigma_X} \\ &= \beta_1 + \frac{\rho_{Z,U}}{\rho_{Z,X}}\frac{\sigma_U}{\sigma_X}\end{aligned}$$

Weak Instruments: Comparing OLS and IV

- Let's compare the asymptotic biases for IV and OLS:

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1^{OLS} = \beta_1 + \rho_{X,U}\frac{\sigma_U}{\sigma_X}$$
$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1^{IV} = \beta_1 + \frac{\rho_{Z,U}}{\rho_{Z,X}}\frac{\sigma_U}{\sigma_X}$$

- The sign and relative strength of asymptotic bias in OLS and IV depends on $\rho_{X,U}$ vs. $\frac{\rho_{Z,U}}{\rho_{Z,X}}$
- Note that the biases from IV and OLS are not necessarily of the same sign

Weak Instruments: Comparing OLS and IV

- If $\rho_{X,U} > 0$ and $\frac{\rho_{Z,U}}{\rho_{Z,X}} > 0$, then both OLS and IV are biased upward
- Which bias is largest?
- If
$$\frac{\rho_{Z,U}}{\rho_{Z,X}} > \rho_{X,U}$$
then the bias is largest for IV and OLS is preferable
- If
$$\frac{\rho_{Z,U}}{\rho_{Z,X}} < \rho_{X,U}$$
then the bias is largest for OLS and IV is preferable

Weak Instruments: Comparing OLS and IV

- The important point here is that, that the lower $\rho_{Z,X}$ (i.e. the weaker the instrument) the smaller must $\rho_{Z,U}$ be in order for IV to be preferable
- For example, suppose we have an instrument for which $\rho_{Z,X} = 0.2$ (implying that regressing X on Z gives an R^2 of $0.2^2 = 0.04$) and that $\rho_{X,U} = 0.1$. Then IV is preferred to OLS only in case $\rho_{Z,U} < 0.2 * 0.1 = 0.02$
- Bottom line: A *small* correlation between the instrument and the error could cause a *large* bias if the instrument is *weak*
- As $\rho_{Z,X} \to 0$, the inconsistency of $\widehat{\beta}_1^{IV}$ goes to infinity for a given value of $\rho_{Z,U} \neq 0$

Checking Assumption 2: Instrument Relevance
Weak Instruments?

- Checking if $Cov(X, Z) \neq 0$
  - Relevant instruments: at least one of the estimated coefficients of the instruments in the first stage is non-zero
- Weak instruments: all the estimated coefficients in the first stage are zero or nearly zero, $Cov(X, Z) \approx 0$
  - Explain very little of the Variation in the endogenous variable
  - One consequence is that the standard errors will be large
- If the instrument is totally irrelevant, i.e. $Cov(X, Z) = 0$, then we cannot even identify the population parameter since we know that $\widehat{\beta}_{IV} = \frac{\widehat{Cov}(Y,Z)}{\widehat{Cov}(Z,X)}$

Checking Assumption 2: Instrument Relevance
Weak Instruments?

- If $Cov(X, Z)$ is small, then $\widehat{Cov}(Z, X)$ is small
- With weak instruments, the denominator is nearly zero
- The sampling distribution of $\widehat{\beta}_{IV}$ is not well approximated by the normal distribution, even if the sample is large
- If instruments are weak, the usual methods of inference are unreliable
- In the case of simple regression with a single instrument, the formula for the standard error is given by:

$$\widehat{\sigma}(\widehat{\beta}_{IV}) = \frac{\sigma}{SST_X R^2_{X,Z}}$$

- if $R^2_{X,Z} \to 0$, get very large standard errors
  if $R^2_{X,Z} \to 1$, OLS standard errors
- Worse is that the IV estimator can be badly biased if the instruments are weak AND not exogenous

Checking Assumption 2: Instrument Relevance
Weak Instruments?

- How relevant must the instruments be for IV to be reliable?
  - In the case of a single endogenous variable $\rightarrow$ look at the F-statistic from the first stage
  - This statistic tests the hypothesis that all the coefficients (instruments) are zero
  - Weak instruments imply a small first stage F-statistic
  - *Rule of thumb:* First stage F-statistic $< 10 \Rightarrow$ weak instruments
  - Weak instrument $\Rightarrow$ the 2SLS estimator more sensitive to bias, and statistical inferences (standard errors, hypothesis tests, confidence intervals etc) will be imprecise

Checking Assumption 2: Instrument Relevance
Weak Instruments?

- Example: evaluating the first stage:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 W_1 + \beta_3 W_2 + U$$

- First-stage equation:

$$X_1 = \pi_0 + \pi_1 W_1 + \pi_2 W_2 + \pi_3 Z_1 + \pi_4 Z_2 + \upsilon$$

- In order to identify (or estimate) β_1, we need *at least* one of the "excluded instruments" (Z_1 and/or Z_2) to be significantly related to X_1, controlling for W_1 and W_2
- Evaluate the first-stage regression:
  - t-values on the excluded (or identifying) instruments
  - F-test for the joint significance of the excluded instruments
- For multiple endogenous regressors, evaluate each first stage regression accordingly

Checking Assumption 2: Instrument Relevance Weak Instruments?

- Do NOT use instruments if they are not significant in the first stage
- The weak instrument problem is generally easy to avoid
- Econometric practice: it is very important to test whether the instruments are significant predictors of the potentially endogenous regressor
- Researchers often report the results of this test along with coefficient estimates in tables that show 2SLS estimators
- What to do if you have weak instruments?
  - Get better instruments!
  - If you have many instruments, drop the weaker ones (will increase credibility of exogeneity assumption and increase the first-stage F)

Example: IV and STATA (first-stage)

```
. reg educ black south smsa exper expersq nearc2 nearc4, robust

Linear regression                                      Number of obs =    3010
                                                       F(  7,  3002) =  522.11
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.4748
                                                       Root MSE      =  1.9421

------------------------------------------------------------------------------
             |               Robust
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   -1.01298   .0877548   -11.54   0.000    -1.185046   -.8409146
       south |  -.2786568   .0787816    -3.54   0.000    -.4331282   -.1241855
        smsa |   .3886608   .0856653     4.54   0.000     .2206922    .5566295
       exper |   -.409533   .0319212   -12.83   0.000    -.4721226   -.3469435
     expersq |   .0006956   .0016996     0.41   0.682     -.002637    .0040282
      nearc2 |   .1076585   .0731378     1.47   0.141    -.0357467    .2510637
      nearc4 |   .3312388   .0805747     4.11   0.000     .1732516    .4892261
       _cons |   16.62244   .1494693   111.21   0.000     16.32936    16.91551
------------------------------------------------------------------------------

. test nearc2=nearc4=0

 ( 1)  nearc2 - nearc4 = 0
 ( 2)  nearc2 = 0

       F(  2,  3002) =    9.72
            Prob > F =    0.0001
```

Example: IV and STATA (reduced form)

```
. reg lwage black south smsa exper expersq nearc2 nearc4, robust

Linear regression                               Number of obs =     3010
                                                F(  7,   3002) =   106.80
                                                Prob > F      =   0.0000
                                                R-squared     =   0.1891
                                                Root MSE      =   .40012

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.2665015   .0184674   -14.43   0.000    -.3027115   -.2302915
       south |  -.1385934   .0166737    -8.31   0.000    -.1712864   -.1059005
        smsa |   .1789726   .0168173    10.64   0.000      .145998    .2119472
       exper |   .0534389   .0068004     7.86   0.000     .0401051    .0667728
     expersq |  -.0022011   .0003267    -6.74   0.000    -.0028417   -.0015606
      nearc2 |   .0408917   .0151046     2.71   0.007     .0112753    .0705082
      nearc4 |   .0423137   .0163187     2.59   0.010     .0103168    .0743106
       _cons |    5.94265   .0357661   166.15   0.000     5.872521    6.012779
------------------------------------------------------------------------------

. test nearc2=nearc4=0

 ( 1)  nearc2 - nearc4 = 0
 ( 2)  nearc2 = 0

       F(  2,  3002) =    7.27
            Prob > F =    0.0007
```

Example: IV and STATA (first- and second-stage)

```
. ivreg lwage black south smsa exper expersq (educ = nearc2 nearc4), robust first

First-stage regressions
-----------------------

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F(  7,  3002) =  387.78
       Model |  10238.7117     7  1462.6731           Prob > F      =  0.0000
    Residual |  11323.3684  3002  3.7719415           R-squared     =  0.4748
-------------+------------------------------           Adj R-squared =  0.4736
       Total |  21562.0801  3009  7.16586243           Root MSE      =  1.9421

------------------------------------------------------------------------------
        educ |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |   -1.01298   .0897474   -11.29   0.000    -1.188953   -.8370075
       south |  -.2786568   .0796824    -3.50   0.000    -.4348945   -.1224192
        smsa |   .3886608   .0854936     4.55   0.000     .2210289    .5562928
       exper |   -.409533   .0336888   -12.16   0.000    -.4755885   -.3434776
     expersq |   .0006956   .0016498     0.42   0.673    -.0025393    .0039305
      nearc2 |   .1076585   .0728953     1.48   0.140    -.0352713    .2505882
      nearc4 |   .3312388    .082587     4.01   0.000     .1693061    .4931715
       _cons |   16.62244   .1780999    93.33   0.000     16.27323    16.97165
------------------------------------------------------------------------------

Instrumental variables (2SLS) regression               Number of obs =    3010
                                                       F(  6,  3003) =  119.78
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1455
                                                       Root MSE      =  .41065

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1608487   .0485705     3.31   0.001     .0656139    .2560835
       black |  -.1019726   .0520797    -1.96   0.050    -.2040881    .0001429
       south |  -.0951187   .0234332    -4.06   0.000    -.1410654    -.049172
        smsa |   .1165736   .0302929     3.85   0.000     .0571767    .1759705
       exper |   .1192112   .0213279     5.59   0.000     .0773923      .16103
     expersq |  -.0023052   .0003691    -6.25   0.000    -.0030289   -.0015816
       _cons |   3.272102   .8178286     4.00   0.000     1.668541    4.875663
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   black south smsa exper expersq nearc2 nearc4
------------------------------------------------------------------------------
```

Checking Assumption 1: Instrument Exogeneity
Testing Overidentifying Restrictions

- Instrument exogeneity: all the instruments are uncorrelated with the error term: $Cov(U, Z_j \mid Z_1, ..., Z_{j-1}, Z_{j+1}, ..., Z_m, W_1, ..., W_r) = 0$ for all $j \in \{1, ...m\}$ conditional on *all* the other exogenous variables
- Instruments that are correlated with the error term
- $\rightarrow$ the first stage of 2SLS doesn't successfully isolate the component of X that is uncorrelated with the error term
- $\rightarrow \widehat{X}$ is correlated with U and 2SLS is inconsistent
- How do we test that $Cov(U, Z) = 0$?
- As U is unobservable, it is essentially impossible to test this assumption
    - Need to make an economic/econometric case

Checking Assumption 1: Instrument Exogeneity
Testing Overidentifying Restrictions

- If there are more instruments than endogenous regressors, it is possible to test for instrument exogeneity $\Rightarrow$ test of overidentifying restrictions:
  - If we have one endogenous variable and two instruments, then it is possible to compute two different IV estimates
  - If both instruments are exogenous, then the two estimates will be close to each other
  - If on the other hand the two instruments generate very different results, then we have a problem with one or the other of the two instruments, or both
  - An "overidentification" test:
    - maintaining the validity of one instrument, we can evaluate the exogeneity of another
    - if we have two valid instruments, using either should give the same answer (asymptotically)

Checking Assumption 1: Instrument Exogeneity
Testing Overidentifying Restrictions

- **Step 1:** Retrieve 2SLS residuals from the second stage equation:

$$\widehat{U} = Y - X\widehat{\beta}_{IV}$$

- **Step 2:** Regress the residuals on ALL instruments and the exogenous variables (that are included in the first stage equation)
- **Step 3:** Compute the F-test for the null hypothesis that all instruments are jointly equal to zero
  - The overidentifying restrictions test statistic is $J = mF$
  - Under "H_0 : all instruments are exogenous", J is χ^2_q (chi-squared) with q degrees of freedom equal to the degree of overidentification; i.e. the number of instruments minus the number of endogenous variables

Checking Assumption 1: Instrument Exogeneity
Testing Overidentifying Restrictions

- If the null hypothesis is rejected, it suggests that at least one of the instruments is invalid
- If we have that some instruments are exogenous and others are endogenous, the J statistic will be large
  - the null hypothesis that all instruments are exogenous will be rejected

Example: STATA and overidentification test

```
. ivregress 2sls lwage black south smsa exper expersq (educ = nearc2 nearc4),
robust

Instrumental variables (2SLS) regression               Number of obs =     3010
                                                       Wald chi2(6)  =   720.37
                                                       Prob > chi2   =   0.0000
                                                       R-squared     =   0.1455
                                                       Root MSE      =   .41017

------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1608487    .048514     3.32   0.001     .0657631    .2559344
       black |  -.1019726   .0520191    -1.96   0.050    -.2039282   -.000017
       south |  -.0951187   .0234059    -4.06   0.000    -.1409935   -.0492439
        smsa |   .1165736   .0302576     3.85   0.000     .0572697    .1758775
       exper |   .1192112   .0213031     5.60   0.000     .0774578    .1609645
     expersq |  -.0023052   .0003686    -6.25   0.000    -.0030277   -.0015827
       _cons |   3.272102   .8168771     4.01   0.000     1.671052    4.873152
------------------------------------------------------------------------------
Instrumented:  educ
Instruments:   black south smsa exper expersq nearc2 nearc4

. estat firststage

  First-stage regression summary statistics
  --------------------------------------------------------------------------
               |            Adjusted      Partial       Robust
      Variable |   R-sq.       R-sq.        R-sq.      F(2,3002)   Prob > F
  -------------+------------------------------------------------------------
          educ |  0.4748      0.4736       0.0063       9.71677    0.0001
  --------------------------------------------------------------------------

. estat overid

  Test of overidentifying restrictions:

  Score chi2(1)            =  2.65321  (p = 0.1033)
```

Example: STATA and overidentification test

```
. qui ivreg lwage black south smsa exper expersq (educ = nearc2 nearc4), robust

. predict uhat_iv, resid

. reg uhat_iv black south smsa exper expersq nearc2 nearc4

      Source |       SS       df       MS              Number of obs =    3010
-------------+------------------------------           F(  7,  3002) =    0.38
       Model |  .445974796     7  .063710685           Prob > F      =  0.9156
    Residual |  505.958899  3002  .168540606           R-squared     =  0.0009
-------------+------------------------------           Adj R-squared = -0.0014
       Total |  506.404874  3009  .168296734           Root MSE      =  .41054

------------------------------------------------------------------------------
     uhat_iv |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       black |  -.0015923   .0189711    -0.08   0.933      -.03879    .0356053
       south |   .0013469   .0168435     0.08   0.936    -.0316791    .0343728
        smsa |  -.0001166   .0180719    -0.01   0.995    -.0355511    .0353179
       exper |   .0001006   .0071212     0.01   0.989    -.0138624    .0140636
     expersq |  -7.78e-06   .0003487    -0.02   0.982    -.0006916     .000676
      nearc2 |    .023575   .0154088     1.53   0.126    -.0066379    .0537879
      nearc4 |  -.0109657   .0174575    -0.63   0.530    -.0451955    .0232641
       _cons |  -.0031499   .0376473    -0.08   0.933    -.0769669    .0706672
------------------------------------------------------------------------------

.
. scalar nobs=e(N)

. scalar r2=e(r2)

. scalar x2=nobs*r2

. scalar pval=1-chi2(1,x2)

. scalar list x2 pval
        x2 =   2.650812
      pval =  .10349701
```

Example: STATA and overidentification test

- P-value=0.103
- insignificant at the 5% level
- (marginally) cannot reject H_0 at the 10% level in the `overid` test
  - one of the two instruments may be invalid (given the validity of the other)
  - when looking at the regression: see that $nearc2$ estimate is larger – suggest that the two-year instrument is the one giving the biased estimates
  - However, technically the test is not saying that

Testing OLS vs. IV (Hausman test)

- We should use OLS (which is BLUE) unless there is a compelling reason not to
- If we have a plausible IV estimator, and a suspicious OLS estimator, we can use a Hausman test to test the consistency of OLS
- The Hausman test is a general test for endogeneity
  - Two estimators are compared
  - One is consistent both under H_0 and under H_1
  - The other is only consistent (and efficient) under H_0
  - This means that if H_0 is rejected, the assumption of exogenous regressors, uncorrelated with the error terms is rejected

$$
\begin{aligned}
H_0 &: \widehat{\beta}_{IV} = \widehat{\beta}_{OLS} \text{ or } Cov(X, U) = 0 \\
H_1 &: \widehat{\beta}_{IV} \neq \widehat{\beta}_{OLS} \text{ or } Cov(X, U) \neq 0
\end{aligned}
$$

Testing OLS vs. IV (Hausman test)

- Under H_0 : $\widehat{\beta}_{IV}$ and $\widehat{\beta}_{OLS}$ are consistent but only $\widehat{\beta}_{OLS}$ is efficient
- Under H_1 : $\widehat{\beta}_{IV}$ is consistent and $\widehat{\beta}_{OLS}$ is inconsistent
- The Hausman test checks if the difference $q = \widehat{\beta}_{IV} - \widehat{\beta}_{OLS}$ is statistically different from zero
- The formula for the Hausman test is:

$$H = \frac{(\widehat{\beta}_{IV} - \widehat{\beta}_{OLS})^2}{\widehat{Var}(\widehat{\beta}_{IV}) - \widehat{Var}(\widehat{\beta}_{OLS})}$$

- where $\widehat{Var}(\widehat{\beta}_{IV})$ and $\widehat{Var}(\widehat{\beta}_{OLS})$ are the estimated Variances of the IV and OLS estimates
  - Under H_0, H follows a chi-squared distribution with k (number of instrumented regressors) degrees of freedom, $H \sim \chi^2_k$
  - If we cannot reject the null $\rightarrow$ use the OLS estimator (consistent *and* efficient)
  - If we reject $\rightarrow$ use the IV estimator (consistent)

A regression based Hausman test

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 W_1 + \beta_3 W_2 + U$$

and the reduced form (first-stage) regression:

$$X_1 = \pi_0 + \pi_1 W_1 + \pi_2 W_2 + \pi_3 Z_1 + \pi_4 Z_2 + \upsilon$$

- The endogeneity of X_1 arises from $Cov(U, \upsilon) \neq 0$. To test this estimate:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 W_1 + \beta_3 W_2 + \upsilon + U$$

- υ is unobserved. Replace it by the estimate retrieved from the first stage:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 W_1 + \beta_3 W_2 + \delta\widehat{\upsilon} + U$$

- do a t-test of $H_0 : \delta = 0$

Example: STATA and Hausman test

```
. qui reg lwage black south smsa exper expersq educ

. est store ols_est

. qui ivreg lwage black south smsa exper expersq (educ = nearc2 nearc4)

. est store iv_est

. hausman iv_est ols_est

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     iv_est      ols_est       Difference          S.E.
-------------+----------------------------------------------------------------
        educ |    .1608487     .074009        .0868397        .0485026
       black |   -.1019726   -.1896315         .087659        .0495785
       south |   -.0951187   -.1248615        .0297428         .017955
        smsa |    .1165736     .161423       -.0448494        .0260073
       exper |    .1192112    .0835958        .0356153        .0201074
     expersq |   -.0023052   -.0022409       -.0000644        .0001481
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from ivreg
           B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        3.21
                Prob>chi2 =      0.7826
```

Example: STATA and Hausman test

```
. qui reg lwage black south smsa exper expersq educ

. est store ols_est

. qui ivreg lwage black south smsa exper expersq (educ = nearc4)

. est store iv_est

. hausman iv_est ols_est

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |     iv_est      ols_est        Difference          S.E.
-------------+----------------------------------------------------------------
        educ |    .1322888      .074009        .0582798        .0491083
       black |   -.1308019    -.1896315        .0588296        .0498476
       south |   -.1049005    -.1248615         .019961        .0174301
        smsa |    .1313237      .161423       -.0300993         .025793
       exper |     .107498     .0835958        .0239021        .0202367
     expersq |   -.0022841    -.0022409       -.0000432        .0001031
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from ivreg
            B = inconsistent under Ha, efficient under Ho; obtained from regress

    Test:  Ho:  difference in coefficients not systematic

                  chi2(6) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        1.41
                Prob>chi2 =      0.9653
```

How to find Instruments?

- Actual random draws (but not controlled experiments)
  Example:
  - Angrist (1990): Vietnam veterans were randomly drafted based on birth day $\rightarrow$ used to estimate the wage impact of military service
- "Natural" randomness
  Examples:
  - Angrist & Krueger (1991): quarter of birth affects school duration $\rightarrow$ used to estimate returns to schooling
  - Paxson (1992): climate shocks $\rightarrow$ used to estimate the impact of income variation on savings and consumption among farm households
  - Angrist & Evans (1998): have same or different sex of children $\rightarrow$ used to estimate the impact of an additional birth on labor supply for women

How to find Instruments?

- Institutional features: institutional rules that have no relation with the endogenous variable
  Examples:
  - Levitt (1997): local elections $\rightarrow$ used to estimate the impact of police on crime
  - Duflo (2001): school building program $\rightarrow$ used to estimate returns to schooling
  - Joensen & Nielsen (2009, 2016): pilot curricula $\rightarrow$ used to estimate returns to math (and science)

Practical Tips for IV

- Use IV when you have an endogeneity problem
- Justify instruments. In practice, the most difficult aspect of IV estimation is to find instruments that are both exogenous and relevant
- Test for significance of excluded (identifying) instruments
- Do overidentification tests (if possible)
- Conduct Hausman test to see whether IV is necessary
- Always show OLS as well!

Practical Tips for IV

- Report the first-stage
  - Does it make sense?
  - Do the coefficients have the "right" magnitude and sign?
- Report the F-statistic on the excluded instrument(s)
  - F-statistics above 10 indicate that you do not have a weak instrument problem (Staiger & Stock, 1997; Stock, Wright & Yogo (2002); not a proof)
- If you have many IVs, pick your best instrument and report the just identified model
- Look at the Reduced Form
  - The reduced form is estimated with OLS and is therefore unbiased
  - If you can't see the causal relationship of interest in the reduced form, it is probably not there

Appendix: Measurement Error

- Sometimes we have variables that are measured with error
- For example, survey questions about:
  - how many hours people worked in the previous week
  - household savings last year
  - household income last three years
- Measurement error in the dependent (Y) and independent (X) variables have different consequences

Appendix: Measurement Error in Dependent Variable

- Let Y^* denote the *true* value of the dependent variable and Y the value we *observe* in the data
- $e_0 = Y - Y^*$ is the measurement error
- Consider a linear regression model of the usual form (and assume the Gauss-Markov assumptions are fulfilled)

$$Y^* = \beta_0 + \beta_1 X_1 + U$$

- Since $Y^* = Y - e_0$, the model we actually can estimate is:

$$Y = \beta_0 + \beta_1 X_1 + U + e_0$$

Appendix: Measurement Error in Dependent Variable II

- The Gauss-Markov assumptions imply that $\mathbb{E}[U] = 0$ and $Cov(X_1, U) = 0$
- We impose the (weak) assumption that $\mathbb{E}[e_0] = 0$ (β_0 biased if $\mathbb{E}[e_0] \neq 0$)
- The important question is whether the measurement error is correlated with X_1
- Typically, it is assumed that $Cov(X_1, e_0) = 0$. This assumption implies that the OLS estimates are unbiased and consistent
- What about standard errors? If e_0 and U are independent, then $Var(U + e_0) = \sigma_u^2 + \sigma_{e_0}^2 > \sigma_u^2$, implying that standard errors increase

Appendix: Measurement Error in an Independent Variable

- Consider the simple regression model

$$Y = \beta_0 + \beta_1 X_1^* + U \tag{1}$$

  and assume that it satisfies the Gauss-Markov assumptions
- However, unlike the standard case, X_1^* is not observed, but we have an imperfect measure of X_1^* which we denote X_1
- $e_1 = X_1 - X_1^*$ is the measurement error
- We assume that the average measurement error in the population is zero; i.e. $\mathbb{E}[e_1] = 0$
- We also assume that U is uncorrelated with X_1 and X_1^*
  - This assumption implies that we have no omitted variable bias due to using X_1 instead of X_1^*

Appendix: Measurement Error in an Independent Variable II

- What happens if we just replace X_1^* with X_1?
  - Since X_1 is observed this is a regression model we can actually estimate
- The answer is that it depends on how the measurement error, e_1, is correlated with X_1 and X_1^*
- Two different cases:
  1. $Cov(X_1, e_1) = 0$ implies that the **observed value** (X_1) is uncorrelated with the measurement error
     - ⋆ This is analogous to the **Proxy Variable** (**PV**) assumption, or measurement error in the dependent variable
  2. $Cov(X_1^*, e_1) = 0$ implies that the **true value** (X_1^*) is uncorrelated with the measurement error
     - ⋆ Known as the **Classical Errors in Variables** (**CEV**) assumption

Appendix: Measurement Error in an Independent Variable, PV case

- Let's look more closely at the first case, when $Cov(X_1, e_1) = 0$
- Note that $e_1 = X_1 - X_1^* \Rightarrow X_1^* = X_1 - e_1$. Thus regression (1) can be rewritten as:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1^* + U \\ &= \beta_0 + \beta_1 (X_1 - e_1) + U \\ &= \beta_0 + \beta_1 X_1 + (U - \beta_1 e_1) \end{aligned}$$

Appendix: Measurement Error in an Independent Variable, PV case II

- The Gauss-Markov assumptions imply that $\mathbb{E}[U] = 0$ and $Cov(X_1, U) = 0$
- Above we further assumed that $\mathbb{E}[e_1] = 0$ and $Cov(X_1, e_1) = 0$
- These two assumptions combined imply that $\mathbb{E}[U - \beta_1 e_1] = 0$ and $Cov(X_1, U - \beta_1 e_1) = 0$
- For notational convenience, let $V = U - \beta_1 e_1$. Then the model we estimate can be written as:

$$Y = \beta_0 + \beta_1 X_1 + V$$

where $\mathbb{E}[V] = 0$ and $Cov(X_1, V) = 0$

Appendix: Measurement Error in an Independent Variable, PV case III

- Remember the condition for consistency from Theorem 5.1 in Lecture 5. Since,

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1 + \frac{Cov(X_1, V)}{Var(X_1)}$$

  we have $\operatorname{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1$ as $Cov(X_1, V) = 0$

- Consequently, measurement error in the independent variable does not give inconsistent (or biased) estimates if the error is uncorrelated with X_1

- However, standard errors increase as the variance in the error term now also includes the measurement error

Appendix: Measurement Error in an Independent Variable, CEV

- Let's look more closely at the second case, when $Cov(X_1^*, e_1) = 0$
- This case is known as the classical errors-in-variables (CEV) assumption
- This assumption comes from writing the observed measure (X_1) as the sum of the true independent variable (X_1^*) and the measurement error (e_1)

$$X_1 = X_1^* + e_1$$

  and then assuming that X_1^* and e_1 are uncorrelated
- A consequence of assuming $Cov(X_1^*, e_1) = 0$ is that $Cov(X_1, e_1) \neq 0$, because:

$$Cov(X_1, e_1) = \mathbb{E}[X_1 e_1] - \mathbb{E}[X_1]\,\mathbb{E}[e_1] = \mathbb{E}[X_1 e_1]$$

  since $\mathbb{E}[e_1] = 0$ by assumption

Appendix: Measurement Error in an Independent Variable, CEV II

- Since $X_1 = X_1^* + e_1$, we can write:

$$\mathbb{E}[X_1 e_1] = \mathbb{E}[(X_1^* + e_1)e_1] = \mathbb{E}[X_1^* e_1 + e_1^2]$$

- Since X_1^* and e_1 are uncorrelated by assumption, we have that:

$$\mathbb{E}[X_1^* e_1 + e_1^2] = \mathbb{E}[X_1^* e_1] + \mathbb{E}[e_1^2]$$

- Moreover, since $Cov(X_1^*, e_1) = \mathbb{E}[X_1^* e_1] + \mathbb{E}[X_1^*]\,\mathbb{E}[e_1]$ and $\mathbb{E}[e_1] = 0$ by assumption, assuming that $Cov(X_1^*, e_1) = 0$ implies that $\mathbb{E}[X_1^* e_1] = 0$
- We thus get:

$$\mathbb{E}[X_1^* e_1 + e_1^2] = \mathbb{E}[e_1^2] = \sigma_{e_1}^2$$

and thus:

$$Cov(X_1, e_1) = \sigma_{e_1}^2 \neq 0$$

Appendix: Measurement Error in an Independent Variable, CEV III

- In the slide above, we showed that the covariance between X_1 (i.e. what we observe) and e_1 equals the variance of e_1
  - Intuition: If measurement error variance is high, then a lot of the variation in X_1 comes from differences in errors errors (e_1) relative to differences in the true value (X_1^*)
- Now let's have a look at the regression we actually estimate; i.e.

$$Y = \beta_0 + \beta_1 X_1 + (U - \beta_1 e_1)$$

- As before, $Cov(X_1, U) = 0$ by assumption. However,

$$Cov(X_1, -\beta_1 e_1) = -\beta_1 Cov(X_1, e_1) = -\beta_1 \sigma_{e_1}^2$$

- Using the notation from above, where $V = U - \beta_1 e_1$, we have $Cov(X_1, V) = -\beta_1 \sigma_{e_1}^2$

Appendix: Measurement Error in an Independent Variable, CEV IV

- As above, consider the probability limit for $\widehat{\beta}_1$:

$$\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 = \beta_1 + \frac{Cov(X_1, V)}{Var(X_1)} = \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{Var(X_1)}$$

- Using the fact that $Var(X_1) = Var(X_1^*) + Var(e_1)$, since by assumption $Cov(X_1^*, e_1) = 0$, we get:

$$\begin{aligned}
\operatorname*{plim}_{n\to\infty} \widehat{\beta}_1 &= \beta_1 - \frac{\beta_1 \sigma^2_{e_1}}{\sigma^2_{X_1^*} + \sigma^2_{e_1}} = \beta_1 \left(1 - \frac{\sigma^2_{e_1}}{\sigma^2_{X_1^*} + \sigma^2_{e_1}}\right) \\
&= \beta_1 \left(\frac{\sigma^2_{X_1^*} + \sigma^2_{e_1}}{\sigma^2_{X_1^*} + \sigma^2_{e_1}} - \frac{\sigma^2_{e_1}}{\sigma^2_{X_1^*} + \sigma^2_{e_1}}\right) = \beta_1 \left(\frac{\sigma^2_{X_1^*}}{\sigma^2_{X_1^*} + \sigma^2_{e_1}}\right)
\end{aligned}$$

Appendix: Measurement Error in an Independent Variable, CEV V

- Note from above that the ratio of the probability limit of estimated coefficient ($\widehat{\beta}_1$) and the true effect (β_1) is:

$$\frac{\text{plim}\,\widehat{\beta}_1}{\beta_1} = \frac{\sigma^2_{X_1^*}}{\sigma^2_{X_1^*} + \sigma^2_{e_1}} < 1$$

- This ratio is called the *reliability ratio*, or the *signal-to-noise ratio*, and reflects the share of the overall variation of X_1 due to variation in X_1^*
- As $\sigma^2_{e_1}$ goes to infinity, plim $\widehat{\beta}_1$ goes to zero
- The bias due to measurement error is called **attenuation bias** as it biases estimated coefficients toward zero; i.e. the true effect is "attenuated"

Appendix: Measurement Error, Summary

- Suppose we have a data on a cognitive test score (X_1) as a measure of true cognitive ability (X_1^*)
- Think of the measurement error as test takers having a "good day" ($e_1 > 0$) or a "bad day" ($e_1 < 0$)
- (**CEV**) $Cov(X_1^*, e_1) = 0$ implies that test takers **ability** (X_1^*) is uncorrelated with the measurement error, but that the test score is positively correlated with the error
  - Test takers with high scores are more likely to have had a good day, test takers with low scores more likely to have had a bad day
- (**PV**) $Cov(X_1, e_1) = 0$ implies that **test scores** (X_1) are uncorrelated with the error, but that ability is negatively correlated with the error
  - Test takers with high ability are more likely to have a bad day, test takers with low ability more likely to have a good day

Appendix: Measurement Error, Summary II

- In most cases, the (**CEV**) assumption $Cov(X_1^*, e_1) = 0$ is more reasonable
- The "test takers" case is such an example:
  Why would people of high ability systematically have "bad" days?
- It is not the case that $Cov(X_1^*, e_1) = 0$ and $Cov(X_1, e_1) = 0$ are the only two possible cases. In reality, anything could happen...

Appendix: Solution to Attenuation Bias due to CEV

- There are two solutions to attenuation bias:
  1. Instrumental Variables (IV)
  2. If we know (or have a good idea) about the reliability ratio, we can use the `eivreg` command in Stata

Back