ECON4150 - Introductory Econometrics

Lecture 13: Internal and external validity

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

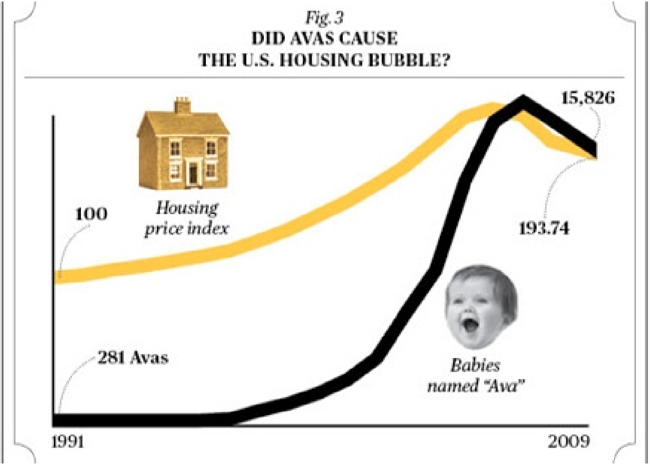
Stock and Watson Chapter 9

Lecture outline

- Definitions of internal and external validity
- Threats to internal validity
  - Omitted variables
  - Functional form misspecification
  - Measurement error
  - Sample selection
  - Simultaneous causality
  - Heteroskedasticity and/or correlated error terms
- Threats to external validity
  - Differences in populations
  - Differences in settings
- Internal and external validity when regression analysis is used for forecasting

Correlation does not imply causation!!

Correlation does not imply causation!!



Correlation does not imply causation!!

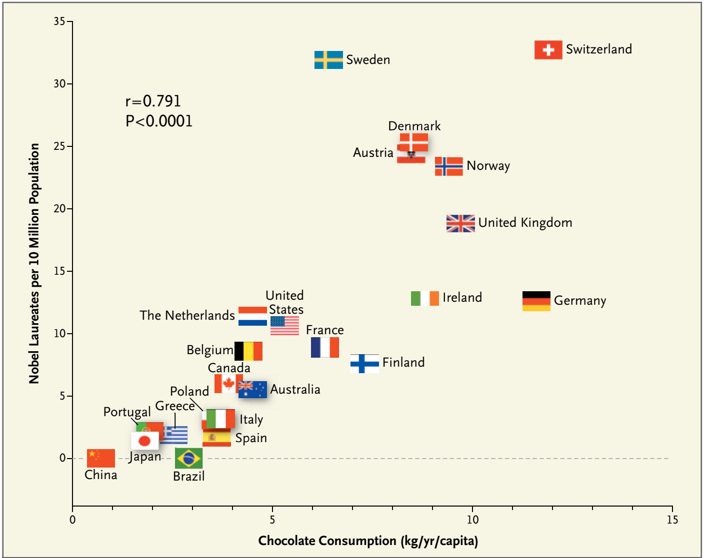


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

Definitions of internal and external validity

Internal validity: the statistical inferences about causal effects are valid for the population and setting being studied.

External validity: the statistical inferences can be generalized from the population and setting studied to other populations and settings

Internal validity in an OLS regression model

Suppose we are interested in the causal effect of X_1 on Y and we estimate the following regression model

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

Internal validity has two components:

1. The OLS estimator of β_1 is unbiased and consistent
   1. $E\left[\widehat{\beta}_1\right] = \beta_1$
   2. $\underset{n \longrightarrow \infty}{plim}\left(\widehat{\beta}_1\right) = \beta_1$
2. Hypothesis tests should have the desired significance level and confidence intervals should have the desired confidence level.

Internal validity in an OLS regression model

```
. regress ln_earnings education

      Source |       SS       df       MS              Number of obs =     602
-------------+------------------------------           F(  1,   600) =  111.85
       Model |  30.9485912     1  30.9485912           Prob > F      =  0.0000
    Residual |  166.015196   600  .276691993           R-squared     =  0.1571
-------------+------------------------------           Adj R-squared =  0.1557
       Total |  196.963787   601  .327726767           Root MSE      =  .52602

------------------------------------------------------------------------------
 ln_earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .0932827   .0088202    10.58   0.000     .0759605     .110605
       _cons |   1.622094   .1243055    13.05   0.000     1.377968    1.866221
------------------------------------------------------------------------------
```

- Is this regression internally valid?
- Is the causal effect of an additional year of education on average hourly earnings equal to 9.3%?
- If we increase the education of a random sample of individuals in the U.S. by one year does this increase their average hourly earnings by 9.3%?

Threats to internal validity

The 3 assumptions of an OLS regression model:

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

Threats to internal validity:

- Omitted variables
- Functional form misspecification
- Measurement error
- Sample selection
- Simultaneous causality
- Heteroskedasticity and/or correlated error terms

The first 5 are violations of assumption (1) the last one is a violation of assumption (2).

Omitted variables

- Suppose we want to estimate the causal effect of X_{1i} on Y_i.
- The *true* population regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \underbrace{\beta_2 X_{2i} + w_i}_{u_i} \quad \textit{with} \quad E\left[w_i | X_{1i}, X_{2i}\right] = 0$$

- But we estimate the following model

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

- We have that

$$\begin{aligned}
\underset{n \longrightarrow \infty}{plim} \left(\widehat{\beta}_1\right) &= \frac{Cov(X_{1i}, Y_i)}{Var(X_{1i})} = \beta_1 + \frac{Cov(X_{1i}, u_i)}{Var(X_{1i})} \\
&= \beta_1 + \frac{Cov(X_{1i}, \beta_2 X_{2i} + w_i)}{Var(X_{1i})} \\
&= \beta_1 + \frac{Cov(X_{1i}, \beta_2 X_{2i}) + Cov(X_{1i}, w_i)}{Var(X_{1i})} \\
&= \beta_1 + \beta_2 \frac{Cov(X_{1i}, X_{2i})}{Var(X_{1i})}
\end{aligned}$$

Omitted variables

$$\underset{n \longrightarrow \infty}{plim}\left(\widehat{\beta}_1\right) = \beta_1 + \beta_2 \frac{Cov\left(X_{1i}, X_{2i}\right)}{Var\left(X_{1i}\right)}$$

- An omitted variable X_{2i} leads to an inconsistent OLS estimate of the causal effect of X_{1i} if

1. The omitted variable X_{2i} is a determinant of the dependent variable Y_i
   - $\beta_2 \neq 0$
2. The omitted variable X_{2i} is correlated with the regressor of interest X_{1i}
   - $Cov\left(X_{1i}, X_{2i}\right) \neq 0$

- Only if there exists 1 or more variables that satisfy both conditions
  - the OLS regression is not internally valid
  - The OLS estimator does not provide a unbiased an consistent estimate of the causal effect of X_{1i}

Omitted variables

- Are there important omitted variables in the returns to education regression in slide 7?

- Important and often discussed omitted variable is ability

  1. Ability is likely a determinant of earnings
  2. Ability is likely correlated with education

- Since we expect $\beta_2 > 0$ and $Cov\left(X_{1i}, X_{2i}\right) > 0$

$$\underset{n \longrightarrow \infty}{plim}\left(\widehat{\beta}_1\right) = \beta_1 + \beta_2 \frac{Cov\left(X_{1i}, X_{2i}\right)}{Var\left(X_{1i}\right)} > \beta_1$$

- Omitting ability from the regression will lead OLS to overestimate the effect of educaion on earnings!

- But can we include ability as independent variable in the regression?

Functional form misspecification

- Suppose that the *true* population regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \underbrace{\beta_2 X_{1i}^2 + w_i}_{u_i} \quad with \quad E\left[w_i | X_{1i}\right] = 0$$

- But we estimate the following model

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

- We have that

$$\begin{aligned} \underset{n \longrightarrow \infty}{plim} \left(\widehat{\beta}_1\right) &= \beta_1 + \frac{Cov(X_{1i}, u_i)}{Var(X_{1i})} \\ &= \beta_1 + \frac{Cov\left(X_{1i}, \beta_2 X_{1i}^2 + w_i\right)}{Var(X_{1i})} \\ &= \beta_1 + \beta_2 \frac{Cov\left(X_{1i}, X_{1i}^2\right)}{Var(X_{1i})} \end{aligned}$$

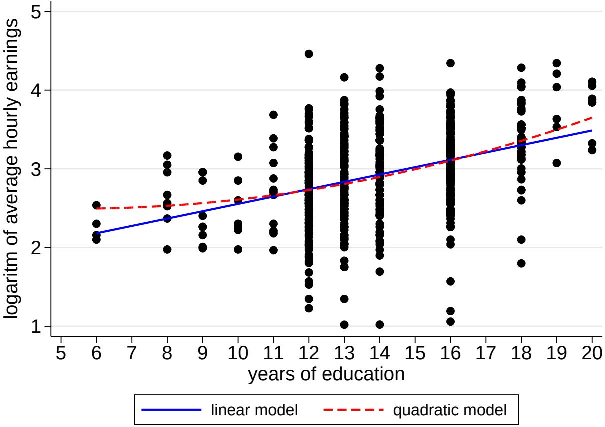
- if $\beta_2 \neq 0$, the simple linear regression model is not internally valid
    - $Cov\left(X_{1i}, X_{1i}^2\right) \neq 0$ by definition.

Functional form misspecification

Should we include education squared in the regression model?

```
. regress ln_earnings education

      Source |       SS       df       MS              Number of obs =     602
-------------+------------------------------           F(  1,   600) =  111.85
       Model |  30.9485912     1  30.9485912           Prob > F      =  0.0000
    Residual |  166.015196   600  .276691993           R-squared     =  0.1571
-------------+------------------------------           Adj R-squared =  0.1557
       Total |  196.963787   601  .327726767           Root MSE      =  .52602

------------------------------------------------------------------------------
 ln_earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |   .0932827   .0088202    10.58   0.000     .0759605     .110605
       _cons |   1.622094   .1243055    13.05   0.000     1.377968    1.866221
------------------------------------------------------------------------------

. regress ln_earnings education education2

      Source |       SS       df       MS              Number of obs =     602
-------------+------------------------------           F(  2,   599) =   58.77
       Model |  32.3114037     2  16.1557019           Prob > F      =  0.0000
    Residual |  164.652383   599   .27487877           R-squared     =  0.1640
-------------+------------------------------           Adj R-squared =  0.1613
       Total |  196.963787   601  .327726767           Root MSE      =  .52429

------------------------------------------------------------------------------
 ln_earnings |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   education |  -.0583157   .0686496    -0.85   0.396    -.1931388    .0765074
  education2 |   .0054138   .0024314     2.23   0.026     .0006387    .0101889
       _cons |   2.651301   .4785439     5.54   0.000     1.711473    3.591129
------------------------------------------------------------------------------
```

Functional form misspecification



- For major part of the support, linear and quadratic models are very similar.

Measurement error

There are different types of measurement error

1. Measurement error in the independent variable X
   - Classical measurement error
   - Measurement error correlated with X
   - Both types of measurement error in X are a violation of internal validity
2. Measurement error in the dependent variable Y
   - Less problematic than measurement error in X
   - Usually not a violation of internal validity
   - Leads to less precise estimates

Measurement error in X: classical measurement error

- Suppose we have the following population regression model

$$Y_i = \beta_0 + \beta_1 X_{1i} + u_i \quad with \quad E[u_i|X_{1i}] = 0$$

- Suppose that we do not observe X_{1i} but we observe $\widetilde{X}_{1i}$ a noisy measure of X_{1i}

$$\widetilde{X}_{1i} = X_{1i} + \omega_i$$

- Adding and subtracting $\beta_1 \widetilde{X}_{1i}$ gives

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_{1i} + \beta_1 (X_{1i} - \widetilde{X}_{1i}) + u_i \\ &= \beta_0 + \beta_1 \widetilde{X}_{1i} - \beta_1 \omega_i + u_i \end{aligned}$$

- Classical measurement error:

$$Cov(X_{1i}, \omega_i) = 0, \quad Cov(\omega_i, u_i) = 0, \quad E[\omega_i] = 0, \quad Var(\omega_i) = \sigma_\omega^2$$

- For example: measurement error due to someone making random mistakes when imputing data in a database.

Measurement error in X: classical measurement error

- Suppose we estimate the following regression model

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_{1i} + e_i \quad with \quad e_i = -\beta_1 \omega_i + u_i$$

- With classical measurement error the OLS estimate of β_1 is inconsistent.

$$\underset{n \longrightarrow \infty}{plim} \left(\widehat{\beta}_1 \right) \quad = \quad \beta_1 + \frac{Cov\left(\widetilde{X}_{1i}, e_i\right)}{Var\left(\widetilde{X}_{1i}\right)}$$

- Substituting $\widetilde{X}_{1i} = X_{1i} + \omega_i$ and $e_i = -\beta_1 \omega_i + u_i$ gives

$$\underset{n \longrightarrow \infty}{plim} \left(\widehat{\beta}_1 \right) \quad = \quad \beta_1 + \frac{Cov(X_{1i}+\omega_i, -\beta_1 \omega_i + u_i)}{Var(X_{1i}+\omega_i)}$$

Measurement error in X: classical measurement error

- From the previous slide we have:

$$\underset{n\longrightarrow\infty}{plim}\left(\widehat{\beta}_1\right) = \beta_1 + \frac{Cov(X_{1i}+\omega_i, -\beta_1\omega_i+u_i)}{Var(X_{1i}+\omega_i)}$$

- Using that $Cov\left(X_{1i},\ \omega_i\right) = Cov\left(X_{1i},\ u_i\right) = Cov\left(\omega_i, u_i\right) = 0$

$$\begin{aligned}
\underset{n\longrightarrow\infty}{plim}\left(\widehat{\beta}_1\right) &= \beta_1 - \frac{\beta_1 Cov(\omega_i,\omega_i)}{Var(X_{1i})+Var(\omega_i)} \\
&= \beta_1\left(1 - \frac{Var(\omega_i)}{Var(X_{1i})+Var(\omega_i)}\right) \\
&= \beta_1\left(\frac{Var(X_{1i})+Var(\omega_i)}{Var(X_{1i})+Var(\omega_i)} - \frac{Var(\omega_i)}{Var(X_{1i})+Var(\omega_i)}\right) \\
&= \beta_1\left(\frac{Var(X_{1i})}{Var(X_{1i})+\sigma^2_\omega}\right)
\end{aligned}$$

- With classical measurement error $\widehat{\beta}_1$ is biased towards 0!

Measurement error in X: classical measurement error

```
1 . program simulate1, rclass
    1. quietly {
    2.         drop _all
    3.         set obs 10000
    4.         gen x1 = rnormal()
    5.         gen x1_observed=x1+rnormal()
    6.         gen y=5+10*x1+rnormal()
    7.
2 .         regress y x1
    8.         return scalar c1 = _b[x1]
    9.
3 .         reg y x1_observed
   10.         return scalar c2 = _b[x1]
   11. }
   12. end

4 .
5 . simulate bhat_NoError=r(c1) bhat_Error=r(c2), reps(100): simulate1

        command:  simulate1
    bhat_NoError:  r(c1)
      bhat_Error:  r(c2)

  Simulations ( 100)
  ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
  ..................................................    50
  ..................................................   100

6 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bhat_NoError	**100**	**9.999062**	**.0106733**	**9.97547**	**10.02077**
bhat_Error	**100**	**4.991671**	**.0507424**	**4.884142**	**5.179945**

Measurement error in X: correlated with X

- Measurement error can also be related to X_i
- For example if X_i is taxable income and individuals systematically underreport by 10%

$$\widetilde{X}_{1i} = 0.9X_{1i}$$

- Suppose we estimate

$$Y_i = \beta_0 + \beta_1 \widetilde{X}_{1i} + e_i \quad \textit{with} \quad e_i = \beta_1 \left(X_i - \widetilde{X}_i \right) + u_i = 0.1\beta_1 X_i + u_i$$

- This will give an OLS estimate of β_1 which is too high!

$$\begin{aligned}
\underset{n \longrightarrow \infty}{plim} \left(\widehat{\beta}_1 \right) &= \beta_1 + \frac{Cov\left(\widetilde{X}_{1i}, e_i\right)}{Var\left(\widetilde{X}_{1i}\right)} \\
&= \beta_1 + \frac{Cov(0.9X_i,\, 0.1\beta_1 X_i + u_i)}{Var(0.9X_i)} \\
&= \beta_1 + \frac{0.9 \cdot 0.1 \cdot \beta_1 Var(X_i)}{0.9^2 Var(X_i)} \\
&= \beta_1 \cdot (1 + \tfrac{1}{9})
\end{aligned}$$

Measurement error in the dependent variable Y

- Measurement error in Y is generally less problematic than measurement error in X
- Suppose Y is measured with classical error

$$\widetilde{Y}_i = Y_i + \omega_i$$

and we estimate

$$\widetilde{Y}_i = \beta_0 + \beta_1 X_i + \underbrace{u_i + \omega_i}_{e_i}$$

- The OLS estimate $\widehat{\beta}_1$ will be unbiased and consistent because $E\left[e_i | X_i\right] = 0$
- The OLS estimate will be less precise because $Var\left(e_i\right) > Var\left(u_i\right)$

Measurement error in the dependent variable Y

```
1 . program simulate2, rclass
    1. quietly {
    2.          drop _all
    3.          set obs 10000
    4.          gen x1 = rnormal()
    5.          gen y=5+10*x1+rnormal()
    6.          gen y_observed=y+rnormal()
    7.
2 .         regress y x1
    8.          return scalar c1 = _b[x1]
    9.
3 .         reg y_observed x1
   10.          return scalar c2 = _b[x1]
   11. }
   12. end

4 .
5 . simulate bhat_NoError=r(c1) bhat_Error=r(c2), reps(100): simulate2

        command:  simulate2
    bhat_NoError:  r(c1)
      bhat_Error:  r(c2)

  Simulations ( 100)
  ----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
  ..................................................    50
  ..................................................   100

6 . sum
```

Variable	Obs	Mean	Std. Dev.	Min	Max
bhat_NoError	**100**	**10.00078**	**.0103849**	**9.973895**	**10.02678**
bhat_Error	**100**	**10.00046**	**.0148451**	**9.957288**	**10.0349**

Measurement error in the returns to education example

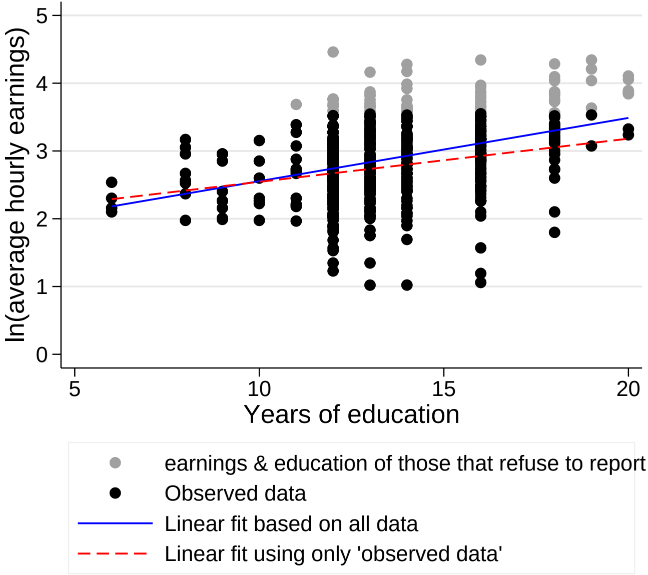
- Is measurement error a threat to internal validity in the regression of earnings on education?
- Data come from the Current Population Survey, a survey among households in the U.S.
- When individuals have to report their earnings and years of education in a survey it is not unlikely that they make mistakes.
- Earnings is the dependent variable so measurement error not so problematic.
- Measurement error in years of education is problematic and will give a biased and inconsistent estimate of the returns to education

Sample selection

- Missing data are a common feature of economic data sets
- We consider 3 types of missing data

1. **Data are missing at random**
   - this will not impose a threat to internal validity
2. **Data are missing based on X**
   - This will not impose a threat to internal validity.
   - For example when we only observe education & earnings for those who completed high school.
   - Can impose a threat to external validity.
3. **Data are missing based on Y**
   - This imposes a threat to internal validity.
   - For example when individuals with high earnings refuse to report how much they earn
   - Resulting bias in OLS estimates is called “sample selection bias”.

Sample selection



Simultaneous causality

- So far we assumed that X affects Y, but what if Y also affects X ?

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad\qquad X_i = \gamma_0 + \gamma_1 Y_i + v_i$$

- Simultaneous causality leads to biased & inconsistent OLS estimate.

- To show this we first solve for $Cov(X_i, u_i)$

$$\begin{aligned} Cov(X_i, u_i) &= Cov(\gamma_0 + \gamma_1 Y_i + v_i,\ u_i) \\ &= Cov(\gamma_1 Y_i,\ u_i) \qquad \text{assuming } Cov(v_i, u_i) = 0 \\ &= Cov(\gamma_1(\beta_0 + \beta_1 X_i + u_i), u_i) \\ &= \gamma_1 \beta_1 Cov(X_i, u_i) + \gamma_1 Var(u_i) \end{aligned}$$

- Solving for $Cov(X_i, u_i)$ gives

$$Cov(X_i, u_i) = \frac{\gamma_1}{1 - \gamma_1 \beta_1} Var(u_i)$$

Simultaneous causality

- Substituting $Cov(X_i, u_i)$ in the formula for the plim of $\widehat{\beta}_1$ gives

$$plim\left(\widehat{\beta}_1\right) = \beta_1 + \frac{Cov(X_{1i}, u_i)}{Var(X_{1i})} = \beta_1 + \frac{\gamma_1 Var(u_i)}{(1-\gamma_1\beta_1) Var(X_{1i})} \neq \beta_1$$

- Simultaneous causality is unlikely a threat to internal validity in returns to education example
  - earnings are generally realized after completing (formal) education
- Simultaneous causality is more likely a threat to internal validity when
  - estimating the effect of class size on average test scores
  - estimating the effect of increasing the price on product demand

Heteroskedasticity and/or correlated error terms

- The threats to internal validity discussed so far
  - Lead to a violation of the first OLS assumption: $E[u_i|X_i] = 0$
  - Lead to biased & inconsistent OLS estimates of the coefficient(s)
- Heteroskedasticity and/or correlated error terms
  - Are a violation of the second OLS assumption: (X_{1i}, Y_i) are iid
  - **Do not** lead to biased & inconsistent OLS estimates of the coefficient(s)
  - But lead to incorrect standard errors
  - Hypothesis tests do not have the desired significance level
  - Confidence intervals do not have the desired confidence level.

Heteroskedasticity and/or correlated error terms

- Heteroskedasticity ($Var\left(u_i\right) \neq \sigma_u^2$) has been discussed during previous lectures
- Solution is to compute heteroskedasticity robust standard errors
- Correlated error terms

  $$Cov(u_i, u_j) \neq 0 \quad for \quad i \neq j$$

  are due to nonrandom sampling
- For example if a dataset contains multiple members from 1 family, because instead of individuals entire families are sampled.
- Solution: Compute cluster-robust se's that are robust to autocorrelation
- More about this in lecture on panel data

What to do when you doubt the internal validity?

- Apart from the last one, all discussed threats to internal validity lead to a violation $E[u_i|X_i] = 0$
- This implies OLS can't be used to estimate causal effect of X on Y.

What to do in this case:

- **Omitted variables:**
  - if observed, include them as additional regressors
  - if unobserved: use panel data or instrumental variables
- **Functional form misspecification:** adjust the functional form
- **Measurement error:**
  - develop model of measurement error and adjust estimates
  - Use instrumental variables
- **Sample selection:** use different estimation method (beyond scope of this course)
- **Simultaneous causality:** use instrumental variables

External validity

- Suppose we estimate a regression model that is internally valid.
- Can the statistical inferences be generalized from the population and setting studied to other populations and settings?

Threats to external validity:

1. Differences in populations
    - The population from which the sample is drawn might differ from the population of interest
    - If you estimate the returns to education for men, these results might not be informative if you want to know the returns to education for women

External validity

Threats to external validity (continued):

2. Differences in settings

   - The setting studied might differ from the setting of interest due to differences in laws, institutional environment and physical environment.
   - For example, the estimated returns to education using data from the U.S might not be informative for Norway

     - the educational system is different
     - different labor market laws (minimum wage laws,..)

Internal & external validity when using regression analysis for forecasting

- Up to now we have dicussed the use of regression analysis to estimate causal effects
- Regression models can also be used for forecasting
- When regression models are used for forecasting
  - external validity is very important
  - internal validity less important
  - not very important that the estimated coefficients are unbiased and consistent

Internal & external validity when using regression analysis for forecasting

Consider the following 2 questions:

1. What is the causal effect of an additional year of education on earnings
2. What are the average earnings of a 40 year old man with 14 years of education in the U.S in 2014?

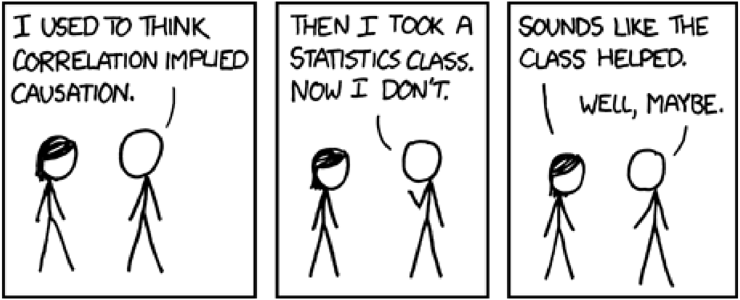
We have these results based on CPS data collected in March 2009 in the U.S.:

```
Linear regression                               Number of obs =        602
                                                   F(  4,   597) =      51.62
                                                   Prob > F      =     0.0000
                                                   R-squared     =     0.2640
                                                   Root MSE      =     10.868
```

earnings	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
education	2.127391	.2012048	10.57	0.000	1.732236	2.522546
female	-6.7412	.9069453	-7.43	0.000	-8.522391	-4.960009
age	1.350918	.2594303	5.21	0.000	.841411	1.860425
age2	-.0144463	.0031923	-4.53	0.000	-.0207159	-.0081767
_cons	-34.79395	5.314045	-6.55	0.000	-45.23044	-24.35745

Internal & external validity when using regression analysis for forecasting

- The regression results can be used to answer the first question
  - if the OLS estimate on education is unbiased and consistent
  - if there are no threats to internal validity
- The regression results can be used to answer the second question
  - if the included explanatory variables 'explain' a lot of the variation in earnings
  - if the regression is externally valid
  - if the population and setting studied are sufficiently close to the population and setting of interest
  - It is not nessesary that the OLS estimate on education is unbiased and consistent.


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.