EEP/IAS 118 - Introductory Applied Econometrics, Lecture 4

Gregory Lane

June 2017

This Lecture

Topics

- Review: Estimators
- The estimator $\hat{\beta}$
- Simple Regression
- Multiple Regression

Assignments

- First Quiz tomorrow, beginning of class
- Problem Set 2 Posted - Due on Wednesday July 5th

Correction from Last Class

I forgot to give you the formula for these two sample properties that form $\hat{\beta}$:

- **Sample Variance:**

$$s_x^2 = \frac{\sum_i (X_i - E(X))^2}{N-1}$$

- **Sample Covariance:**

$$s_{xy} \frac{\sum_i (X_i - E(X))(Y_i - E(Y))}{N-1}$$

We divide by $N - 1$ to make these unbiased estimators of the population variance and covariance

Random Samples and Estimators

Definition: If $X_1, X_2 \cdots, X_n$ are independent random variables with a common probability density function, then $\{X_1, \cdots X_n\}$ is said to be a **random sample** from the population represented by that same PDF.

The random nature of $X_1, X_2 \cdots, X_n$ in the definition of random sampling illustrates that many different outcomes are possible before the sampling is actually carried out.

Example: Obtaining data on family income from a sample of $n = 100$ families in the US: the incomes you observe will usually differ for each different sample of 100 families.

Population Parameters

If X is a random variable, the expected value (or expectation) of X, is the weighted average of all possible values of X.

$$E(X) = \mu = \sum_{j=1}^{k} x_j f(x_j)$$

If X is a random variable, the variance tells us the expected distance from X to its mean:

$$Var(X) = \sigma^2 = E[(X - E(X))^2]$$

Both of these are **population parameters**.

Sample Estimates

We never actually have the entire population of data to work with. We do however have the ability to collect information from a representative sample of the population.

We can proceed to calculate the average and variance in a sample, and say this is the best *estimate* for the average and variance in the population.

Sample Estimator

Recall our population has mean μ and variance σ_X^2. Then

- An **estimator** of μ is the sample mean $\bar{X} = \frac{1}{n} \sum_i X_i$
- An **estimator** of σ_X^2 is $s_X^2 = \frac{1}{n-1} \sum_i (X_i - \bar{X})^2$

When we collect a specific sample from this population, we can get a particular **estimate** for $\bar{X}$ and s_X^2

Note: I will sometimes write $\hat{\sigma}_X^2$ or s_X^2, but they mean the same thing. (A "hat" indicates that something is an estimator)

Properties of Estimators

Remember that estimators themselves are **random variables** because they depend on a random sample: as we obtain different random samples from the population, the values of $\bar{X}$ can change. Hence they have a certain probability distribution, with a certain mean and a certain variance/ standard deviation.

- We can see this if we take several samples from the same population and calculate $\bar{X}$ for each one

Properties of Estimators

$$E[\bar{X}] = E[\frac{1}{n}\sum_i X_i] = \frac{1}{n}E[\sum_i X_i] = \frac{1}{n}nE[X_i] = \frac{1}{n}n(\mu) = \mu$$

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_i X_i\right] = \frac{1}{n^2}Var\left[\sum_i X_i\right] = \frac{1}{n^2}nVar[X_i] = \frac{\sigma_X^2}{n}$$

$$Sd[\bar{X}] = \sqrt{(Var[\bar{X}])} = \frac{\sigma_X}{\sqrt{n}}$$

BUT we don't know σ_X because this is a *population* parameter!
So how can get the standard deviation of our estimator?

Standard Errors of Estimators

So, instead we use our estimator for σ_X,
$s_X = \sqrt{\frac{1}{n-1} \sum_i^n (x_i - \bar{x})^2}$.
We call this term the **standard error** - essentially the standard deviation of our estimator once we replaced the population σ_X with the sample estimator s_X

$$Se[\bar{X}] = \frac{s_X}{\sqrt{n}}$$

Summary: X as continuous variable

We have a random sample, $X_1 \cdots X_n$

	Symbol	Formula
Population parameters	μ	$\sum_{j=1}^{k} x_j f(x_j)$
	σ_X^2	$E[(X - E(X))^2]$
	σ_X	$\sqrt{E[(X - E(X))^2]}$
Sample estimators	$\bar{X}$	$\frac{1}{n} \sum_i X_i$
	s_X^2	$\frac{1}{n-1} \sum_i (X_i - \bar{X})^2$
	s_X	$\sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}$
Estimator properties	$E(\bar{X})$	μ
	$Var(\bar{X})$	$\frac{\sigma_X^2}{n}$
	$Sd(\bar{X})$	$\frac{\sigma_X}{\sqrt{n}}$
SE of estimator	$Se(\bar{X})$	$\frac{s_X}{\sqrt{n}}$

Binary Random Variable

If the random variable X can only take on one of two values $\{0, 1\}$, we call this a binary random variable. The calculation of the *mean* of a binary random variable is the same, but we denote its value as p standing for **proportion**

- p must be between zero and one, and we can interpret it as the probability that X takes on the value 1

The primary difference to keep in mind with a binary random variable is that the *variance* is completely defined by p

$$\sigma_X^2 = p(1 - p)$$

That means, that if we know p, then we know both the mean *and* the variance / standard deviation of X (contrast with continuous X where we have both μ and σ)

Summary: X as binary variable

We have a random sample, $X_1 \cdots X_n$, $X_j \in \{0,1\}$

	Symbol	Formula
Population parameters	p	$\sum_{j=1}^{k} x_j f(x_j)$
	σ_X^2	$p(1-p)$
	σ_X	$\sqrt{p(1-p)}$
Sample estimators	$\hat{p}$	$\frac{1}{n}\sum_i X_i$
	s_X^2	$\hat{p}(1-\hat{p})$
	s_X	$\sqrt{\hat{p}(1-\hat{p})}$
Estimator properties	$E(\bar{X})$	p
	$Var(\bar{X})$	$p(1-p)$
	$Sd(\bar{X})$	$\sqrt{p(1-p)}$
SE of estimator	$Se(\bar{X})$	$\frac{s_X}{\sqrt{n}}$

The estimator $\hat{\beta}$

Transitioning back to the population model we discussed previously:

$$y = \beta_0 + \beta_1 x + u$$

$\hat{\beta}_0$ and $\hat{\beta}_1$ are **estimators** for the parameters β_0 and β_1. Indeed we derived a formula for our β's, this was a rule that assigns each possible outcome of the sample a value of β. Then, for the given sample of data we work with, we obtain particular intercept and slope **estimates**, β_0 and β_1.

Recall that because $\hat{\beta}$ is an estimator based of a random sample, it has a **standard error** of its own.

The estimator $\hat{\beta}$

$\hat{\beta}$ is an estimator. Therefore, we want to know it's properties, in particular:

- What is $E(\hat{\beta})$? - an important property will be that $E(\hat{\beta}) = \beta$. Most of econometrics is finding the conditions under which this is true
- What is $Var(\hat{\beta})$? - will inform us about how far away $\hat{\beta}$ could be from the true population β

To answer either of these questions, we first need to make some assumptions about the true population model...

Assumptions of Linear Regression

We make these assumption about the "true data generating process"

Model	Simple
SLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$
SLR.2	$\{(x_i, y_i), \quad i = 1 \cdots N\}$ is a random sample from the population
SLR.3	The observed explanatory variable (x) is not constant: $Var(x) \neq 0$
SLR.4	No matter what we observe x to be, we expect the unobserved u to be zero: $E[u\|x] = 0$
SLR.5	The "error term" has the same variance for any value of x : $Var(u\|x) = \sigma^2$

Assumption 1

The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + u$

- Rules out models that have things like: β^2 or $\beta_1 \times \beta_2$
- Seems restrictive, but remember we can still include things like: X^2, $log(X)$, $\sqrt{X}$, etc. We can still accommodate most functional forms

Assumption 2

$\{(x_i, y_i), \;\; i = 1 \cdots N\}$ is a random sample from the population

- Relatively straight forward - the data we observe is a true random sample drawn from the population we care about
- Processes that would **NOT** be random:
  - Calling the first 100 people in the phone book
  - Surveying the first 10 people to arrive in class
  - Asking for volunteers to to fill out a survey
- Even if we don't have a *true* random sample, sometimes we are okay with that, as this might be the relevant population to study (e.g. people who apply for a scholarship)

Assumption 3

The observed explanatory variable (x) is not constant: $Var(x) \neq 0$

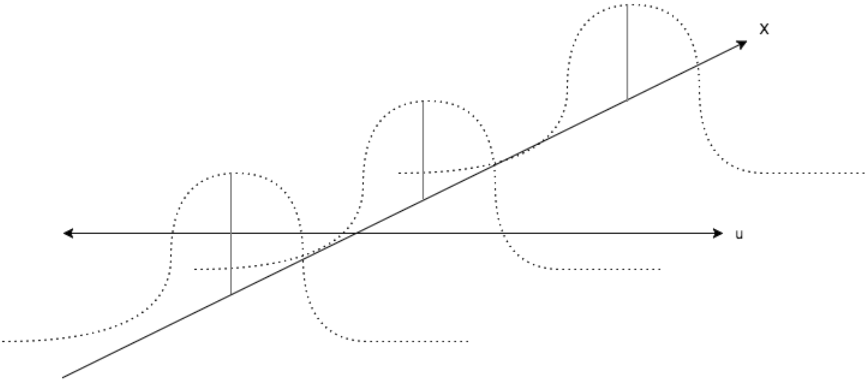
- We need some variation in x in order to even calculate any value for $\hat{\beta}$
- When we only have one x, this assumption is trivial - if we only observe people with 12 years of education, we won't be able to say anything about the effect of education on income

Assumption 4

The "mean independence" assumption on the error term $E[u|x] = 0$ is probably the most critical assumption we make in regression.

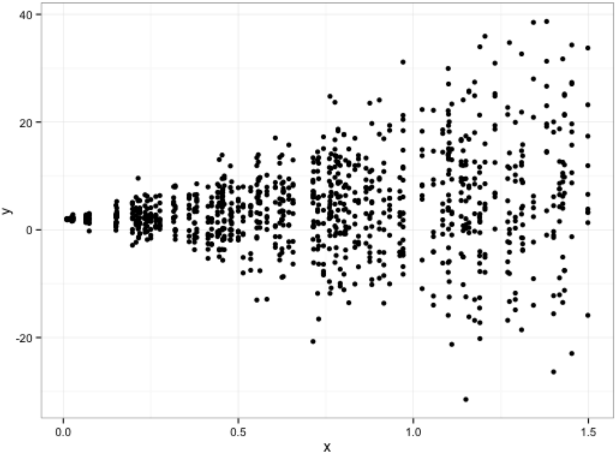
- This assumption allows us to think about β in causal terms - i.e. "the causal effect of one more unit of X's on expected value Y"
- Classic example of violating this assumption is regression of income on education
  - *IF* we could control for all variables that affect income then we could recover the true effect of education on income
  - But we can never observe everything. E.g. we don't observe ability which is correlated with education and income which biases our estimate of educations effect on earnings
- Omitted Variable Bias (OVB) is an example of violating this assumption.

Assumption 4



Assumption 5

The assumption that $Var(u|x) = \sigma^2$ is called the homoskedasticity assumption. A **violation** of this assumption would look like this (heteroskedasticity):



What do we get from these assumptions?

Using only assumptions 1 - 4, we can prove that:

1. $E(\hat{\beta}_1) = \beta_1$
2. $E(\hat{\beta}_0) = \beta_0$

This means that the mean of our estimators $\hat{\beta}_1$ and $\hat{\beta}_0$ are our true population parameters β_1 and β_1

- This is good! If we don't have this, we lose the ability to assign *causality* to our $\hat{\beta}$ estimates
- The proofs for these results are in the notes, but you don't need to know them

What do we get from these assumptions?

If we add assumption 5, we can also show that:

3. $Var(\hat{\beta}_1) = \sigma_u^2 / SST_x = \sigma_u^2 / (n-1)s_x^2$
4. $Var(\hat{\beta}_0) = \frac{\sigma_u^2}{SST_x} \frac{\sum_i x_i}{n}$

NOTE: As before, we don't know σ_u^2 (or SST_x) as this is a population parameters.

- So to calculate this we use an estimator for σ_u^2 in our formula:

$$\hat{\sigma}_u^2 = \frac{\sum_i \hat{u}_i^2}{n-2}$$

The primary driver of the variance of $\hat{\beta}$ is the size of our residuals $\hat{u}$. Should make intuitive sense: implies the data points are not tightly packed around the regression line $\Rightarrow$ the variation in $\hat{\beta}$ will be large as well

What do we get from these assumptions?

3. $Var(\hat{\beta}_1) = \sigma_u^2 / SST_x = \sigma_u^2 / (n-1)s_x^2$
4. $Var(\hat{\beta}_0) = \frac{\sigma_u^2}{SST_x} \frac{\sum_i x_i}{n}$

Ideally we want variance of $\hat{\beta}$ to be low - what can we do?

- Increase sample size (n is in the denominator)
- Large variance in x - may seem counter-intuitive, but true
- Reduce the size of $\hat{\sigma}$ - we can do this by controlling for many variables

Note: The standard error of $\hat{\beta}$ is:

$$\sqrt{Var(\hat{\beta})} = \frac{\hat{\sigma}_u}{\sqrt{(n-1)s_x^2}}$$

Example: Regression $n = 400$

```
. reg wage educ

      Source |       SS           df       MS      Number of obs   =       400
-------------+----------------------------------   F(1, 398)       =     69.64
       Model |  7947.97607         1  7947.97607   Prob > F        =    0.0000
    Residual |  45425.1083       398  114.133438   R-squared       =    0.1489
-------------+----------------------------------   Adj R-squared   =    0.1468
       Total |  53373.0843       399  133.767129   Root MSE        =    10.683

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   2.166719   .2596455     8.34   0.000     1.656271    2.677168
       _cons |  -11.09661   3.579697    -3.10   0.002    -18.13409   -4.059135
------------------------------------------------------------------------------
```

Example: Regression $n = 2000$

```
. reg wage educ

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  1,  1998) =  333.22
       Model |  41122.3613     1  41122.3613           Prob > F      =  0.0000
    Residual |  246572.252  1998  123.409535           R-squared     =  0.1429
-------------+------------------------------           Adj R-squared =  0.1425
       Total |  287694.613  1999  143.919266           Root MSE      =  11.109

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   2.203853   .1207308    18.25   0.000     1.967082    2.440625
       _cons |  -11.93304   1.661577    -7.18   0.000    -15.19164   -8.674433
------------------------------------------------------------------------------
```

Notice how $se(\hat{\beta})$ has dropped

Example: Regression $n = 4000$

```
. reg wage educ

      Source |       SS       df       MS              Number of obs =    4000
-------------+------------------------------           F(  1,  3998) =  728.91
       Model |  85546.3393     1  85546.3393           Prob > F      =  0.0000
    Residual |  469213.909  3998  117.362158           R-squared     =  0.1542
-------------+------------------------------           Adj R-squared =  0.1540
       Total |  554760.248  3999  138.724743           Root MSE      =  10.833

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   2.221041   .0822659    27.00   0.000     2.059754    2.382328
       _cons |  -12.02081   1.133032   -10.61   0.000    -14.24218   -9.799433
------------------------------------------------------------------------------
```

Practice: Calculate $se(\hat{\beta})$

```
. sum wage educ

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        educ |      2000      13.633      2.0877          9         18
        wage |      2000    18.34701    11.49495         .7   82.42857

. reg wage educ

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  1,  1998) =  376.94
       Model |  41922.0349     1  41922.0349           Prob > F      =  0.0000
    Residual |  222213.443  1998  111.217939           R-squared     =  0.1587
-------------+------------------------------           Adj R-squared =  0.1583
       Total |  264135.478  1999  132.133806           Root MSE      =  10.546

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   2.193546    .112983    19.41   0.000     1.971969    2.415123
       _cons |   -11.5576   1.558244    -7.42   0.000    -14.61355   -8.501649
------------------------------------------------------------------------------
```

How could we calculate $se(\hat{\beta})$ if we didn't see it's value here (you know, like on an exam...)?

Practice: Calculate $se(\hat{\beta})$

$$se(\hat{\beta}) = \frac{\hat{\sigma}_u}{\sqrt{SST_x}} = \frac{\hat{\sigma}_u}{\sqrt{(n-1)s_x^2}}$$

- $\hat{\sigma}_u^2 = \frac{\sum_i \hat{u}_i^2}{n-2} = \frac{SSR}{n-2} = \frac{22213.44}{1998} = 111.22$
- $SST_x = (2.088)^2 * 1999 = 8715.13$
- $var(\hat{\beta}) = \frac{\hat{\sigma}_u^2}{SST_x} = \frac{111.22}{8715.13} = 0.01276$
- $se(\hat{\beta}) = \sqrt{(var(\hat{\beta})} = \sqrt{0.01276} = 0.1130$

Summary: Regression

We have a random sample, $X_1 \cdots X_n$, and A1-A5 are satisfied:

	Symbol	Formula
Population estimators	β_0	
	β_1	
Sample estimators	$\hat{\beta}_0$	$\bar{y} - \hat{\beta}_1 \bar{x}$
	$\hat{\beta}_1$	$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
Estimator properties	$E(\hat{\beta}_0)$	β_0
	$E(\hat{\beta}_1)$	β_1
	$Var(\hat{\beta}_1)$	$\frac{\sigma_u^2}{SST_x}$
	$Sd(\hat{\beta}_1)$	$\frac{\sigma_u}{\sqrt{SST_x}}$
SE of estimator	$Se(\hat{\beta}_1)$	$\frac{\hat{\sigma}_u}{\sqrt{SST_x}}$

*I don't show $Var(\hat{\beta}_0)$, $Sd(\hat{\beta}_0)$, or $Se(\hat{\beta}_0)$ because we rarely care

Multiple Linear Regression: Intro

Up to now, we have dealt with regressions with only one explanatory variable. In practice, we almost always include many more explanatory variables. E.g.:

$$wage = \beta_0 + \beta_1 educ + \beta_2 experience + u$$

Why add additional x?

1. Interested in effect of x_2 on y
2. We want to remove unobservables from u - remember everything that affects y that is not specified in our regression is hidden in u
   - Can increase precision of $\hat{\beta}_1$ and reduce bias (more on this in the future)
3. Need to account for non-linear relationship (x_1 and x_1^2)

Multiple Linear Regression: Interpretation

How do we think about β_j now that there are multiple x?

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

If we assume that $E(u|x_1, ..., x_k) = 0$ then we can write:

$$E(y|x_1, ..., x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

- Now, β_1 measures the partial effect of increase x_1 on $E(y)$, **holding $x_2, ..., x_k$ constant**
- I.e, we are "controlling" for $x_2, .., x_k$

Multiple Linear Regression: Interpretation

How do we think about $\hat{\beta}_j$ now that there are multiple x?

- $\hat{\beta}_1$ measured the effect on the predicted $\hat{y}$ of a change in x_1 by 1 unit, holding $x_2, x_3, \dots$ fixed
- Ex: "Holding experience and gender fixed, a one year increase in education leads to a 11.7%

Multiple Linear Regression: Derivation

How do we go about finding values for $\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_k$?

- Again we minimize the sum of the squared errors:

$$min \sum_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} \cdots \hat{\beta}_k x_{ik})^2$$

  No easy formula for $\hat{\beta}$, but fortunately we have computers that can solve this*

- Once we do solve, this gives us:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} + \hat{\mu}_{ik}$$

*This is why we use matrix notation in advanced courses

Practice Interpretation

Data on urbanization % (scale 1 to 100), logGDP per capita, and agriculture productivity (average yield) were used to run this regression

$$\widehat{urban} = -25.13 + 10.43 logGDP + 0.41 agprod$$

1. Interpret the coefficients on $logGDP$ and $agprod$

Practice interpretation

$$\widehat{urban} = -25.13 + 10.43 logGDP + 0.41 agprod$$

1. Interpret the coefficients on $logGDP$ and $agprod$
   logGDP:
   - **Sign:** There is a positive sign, this makes sense - as a country gets richer more people move to the city
   - **Significance:** We'll get here (but let's assume it is)
   - **Size:** A 1% increase in GDP per capita will cause an increase in predicted urbanization by 0.1043 percentage points holding agricultural productivity constant

Assumptions for Multiple Linear Regression

How do the necessary assumptions change when we have multiple X ?

Model	Multiple
MLR.1	The population model is linear in parameters $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \mu$
MLR.2	$\{(x_{i1}, \cdots, x_{ik}, y_i), \quad i = 1 \cdots N\}$ is a random sample from the population
MLR.3	No perfect colinearity among observed variables and $Var(x_j) \neq 0$, $j = 1 \cdots k$
MLR.4	No matter what we observe $(x_{i1}, \cdots, x_{ik})$ to be, we expect the unobserved u to be zero $E[u \mid x_1, \cdots, x_k] = 0$
MLR.5	The “error term” has the same variance for any value of$(x_1, \cdots x_k)$: $Var(u \mid x_1, \cdots x_k) = \sigma^2$

What do we get from these assumptions?

Using only assumptions 1 - 4, we can prove that:

1. $E(\hat{\beta}_j) = \beta_j$

This means that the mean of our estimators $\hat{\beta}_j$ are our true population parameters β_j

If we add assumption 5, we can also show that:

2. $Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1-R_j^2)}$

where $SST = \sum_j (x_{ij} - \bar{x})^2$ is the total sample variation in x_j, and R_j^2 is the R squared from regressing x_j on all other independent variables, and (as before)

$$\hat{\sigma}_u^2 = \frac{\sum_i \hat{u}_i^2}{n-2}$$

MLR.3 - Multicolinearity

- **Definition**: Two variables are said to be perfectly multi-collinear if one variable is a linear combination of the other variable ($x_2 = ax_1 + b$)
- **Intuition**: think about including two variables in your regression (male and female), and remember in the MLR framework we want to "hold all else constant"
- **Note**: some correlation between X variables is normal - we only have a problem when there is a *perfect* or near perfect (very high) correlation between X variables
  - Problem with *near* multicollinearity is that the variance of our estimator $\hat{\beta}$ increases greatly.

MLR.3 - Multi-colinearity

If we have *perfect* multi-colineartiy, our OLS algorithm can't work

- Stata will automatically remove one of the variables for you

If we have *near perfect* multicolineartiy, we have a harder problem

- $Var(\hat{\beta})$ will be very high
- We can see this in the variance formula: $\frac{\sigma_u^2}{SST_j(1-R_j^2)}$
- If another x variable are very closely related to x_j, then R_j^2 will be close to 1. (note, if we had perfect multi-collinearity, then $R_j^2 = 1$, which breaks the formula)
- Implies that the denominator will be very close to zero $\Rightarrow$ high variance

MLR.3 - Multi-collinearity

Common examples of multi-collinearity:

1. "Dummy variable trap": can't include all categories for indicator variables. Ex:
   - Include both a $female$ and $male$ indicator variable
   - Include all education categories ($highschool$, $somecollege$, $college$, $graduate$)
2. Two variables are different measures of the same variable: e.g. GDP measured using two different sources

What do we do?

- Drop one of the variables

Example 1: Multi-collinearity

Dummy variable trap:

```
. reg lwage educ exper female male

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  3,  1996) =  247.50
       Model |  182.35726     3  60.7857535           Prob > F      =  0.0000
    Residual |  490.219607  1996  .245601005           R-squared     =  0.2711
-------------+------------------------------           Adj R-squared =  0.2700
       Total |  672.576867  1999  .336456662           Root MSE      =  .49558

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1167441   .0053157    21.96   0.000     .1063191     .127169
       exper |   .0109089    .000869    12.55   0.000     .0092046    .0126132
      female |  -.2543189   .0222067   -11.45   0.000    -.2978696   -.2107682
        male |  (dropped)
       _cons |   1.055792   .0757381    13.94   0.000     .9072576    1.204326
------------------------------------------------------------------------------
```

Example 2: Multi-collinearity

Near multi-collinearity between age and experience

```
. reg lwage educ exper female age

      Source |       SS       df       MS              Number of obs =    2000
-------------+------------------------------           F(  4,  1995) =  185.69
       Model |  182.468262     4  45.6170655           Prob > F      =  0.0000
    Residual |  490.108605  1995  .245668474           R-squared     =  0.2713
-------------+------------------------------           Adj R-squared =  0.2698
       Total |  672.576867  1999  .336456662           Root MSE      =  .49565

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .1692465   .1115687     1.52   0.129    -.0495568    .3880498
       exper |   .0633711   .1113346     0.57   0.569    -.1549732    .2817154
      female |  -.2545469   .0222135   -11.46   0.000     -.298111   -.2109827
         age |  -.0524796   .1113744    -0.47   0.638     -.270902    .1659428
       _cons |   1.370917   .6728026     2.04   0.042     .0514472    2.690386
------------------------------------------------------------------------------
```

Note: not easy to detect. Why you should look at correlation between x variables (use "corr" command in Stata)

$Var(\hat{\beta})$ part 2

Moving back to $var(\hat{\beta})$, how can we reduce variance with multiple regressors:

$$Var(\hat{\beta}_j) = \frac{\sigma_u^2}{SST_j(1 - R_j^2)}$$

1. Add more explanatory variables that explain variation in y
2. Avoid multi-collinearity
3. Increase sample size
4. Consider x with a larger variance

Q: Identify which part of the variance equation the four options above affects

Choosing what goes into the regression

How do we decide which variables to include?

There are three cases we want to think about:

1. Adding/omitting an irrelevant variable
2. Adding/omitting an important variable that is **NOT** correlated with the other independent variable
3. Adding/omitting an important variable that **IS** correlated with the other independent variable

Choosing what goes into the regression

How do we decide which variables to include?

There are three cases we want to think about:

1. Adding/omitting an irrelevant variable
2. Adding/omitting an important variable that is **NOT** correlated with the other independent variable
3. Adding/omitting an important variable that **IS** correlated with the other independent variable

Irrelevant x variable

We are trying to explain wages using education, experience, and gender:

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,\text{educ} + \underset{(.0009)}{.011}\,\text{exp} - \underset{(.02)}{.25}\,\text{female} \qquad R^2 = .27,\; n = 2000$$

Now we add an "irrelevant" variable - whether someone is non-white:

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,\text{educ} + \underset{(.0009)}{.011}\,\text{exp} - \underset{(.02)}{.25}\,\text{female} - \underset{(.031)}{.037}\,\text{nonwhite} \qquad R^2 = .27,\; n = 2000$$

Irrelevant x variable

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,\text{educ} + \underset{(.0009)}{.011}\,\text{exp} - \underset{(.02)}{.25}\,\text{female} \qquad R^2 = .27,\ n = 2000$$

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,\text{educ} + \underset{(.0009)}{.011}\,\text{exp} - \underset{(.02)}{.25}\,\text{female} - \underset{(.031)}{.037}\,\text{nonwhite} \qquad R^2 = .27,\ n = 2000$$

- R^2 stays the same $\Rightarrow$ "nonwhite" does not explain much of the wage variation
- Coefficients on other variables stay the same
- Standard errors on other coefficients may rise (can't see that here because effect is small)

Important x variable, NOT correlated with others

```
. correlate lwage educ exp female profocc nonwhite
(obs=2000)
```

	lwage	educ	exper	female	profocc	nonwhite
lwage	1.0000					
educ	0.4097	1.0000				
exper	0.2358	0.0010	1.0000			
female	-0.1935	0.0489	0.0210	1.0000		
profocc	0.2181	0.4276	-0.0383	0.1077	1.0000	
nonwhite	-0.0379	-0.0051	-0.0200	0.0368	-0.0143	1.0000

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,\text{educ} + \underset{(.0009)}{.011}\,\text{exp} - \underset{(.02)}{.25}\,\text{female} \qquad R^2 = .27, \; n = 2000$$

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.28} + \underset{(.006)}{.117}\,\text{educ} - \underset{(.02)}{.25}\,\text{female} \qquad R^2 = .21, \; n = 2000$$

Important x variable, NOT correlated with others

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.06} + \underset{(.005)}{.117}\,\text{educ} + \underset{(.0009)}{.011}\,\text{exp} - \underset{(.02)}{.25}\,\text{female} \qquad R^2 = .27,\ n = 2000$$

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.28} + \underset{(.006)}{.117}\,\text{educ} - \underset{(.02)}{.25}\,\text{female} \qquad R^2 = .21,\ n = 2000$$

- R^2 drops because experience did explain some of the variation in wages
- Other coefficients stay the same
- Because exp is not strongly correlated with the other explanatory variables

Important x variable, IS correlated with others

```
. correlate lwage educ exp female profocc nonwhite
(obs=2000)

             |    lwage     educ    exper   female  profocc nonwhite
-------------+------------------------------------------------------
       lwage |   1.0000
        educ |   0.4097   1.0000
       exper |   0.2358   0.0010   1.0000
      female |  -0.1935   0.0489   0.0210   1.0000
     profocc |   0.2181   0.4276  -0.0383   0.1077   1.0000
    nonwhite |  -0.0379  -0.0051  -0.0200   0.0368  -0.0143   1.0000
```

$$\widehat{\log(\text{wage})} = \underset{(.08)}{1.17} + \underset{(.005)}{.106\,\text{educ}} \underset{(.0009)}{+.011\,\text{exp}} \underset{(.02)}{-.26\,\text{female}} + \underset{(.03)}{.012\,\text{profocc}} \qquad R^2 = .28, \; n = 2000$$

$$\widehat{\log(\text{wage})} = \underset{(.03)}{2.57} + \underset{(.0009)}{+.011\,\text{exp}} \underset{(.02)}{-.26\,\text{female}} + \underset{(.03)}{.358\,\text{profocc}} \qquad R^2 = .16, \; n = 2000$$

Important x variable, IS correlated with others

$\widehat{\log(\text{wage})} =$	1.17	+ .106 educ	+.011 exp	- .26 female	+ .012 profocc	$R^2 = .28$
	(.08)	(.005)	(.0009)	(.02)	(.03)	n = 2000
$\widehat{\log(\text{wage})} =$	2.57	+	+.011 exp	- .26 female	+ .358 profocc	$R^2 = .16$
	(.03)		(.0009)	(.02)	(.03)	n = 2000

- R^2 drops because education explained a lot of the variation
- Coefficient on professional occupation changes a lot
- Education is strongly correlated with occupation choice
- We have **omitted variable bias**!! We'll cover this next time

Q: What is the intuition here?