FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

IAPRI-MSU Technical Training

Intro to Applied Econometrics:
Basic theory and Stata examples

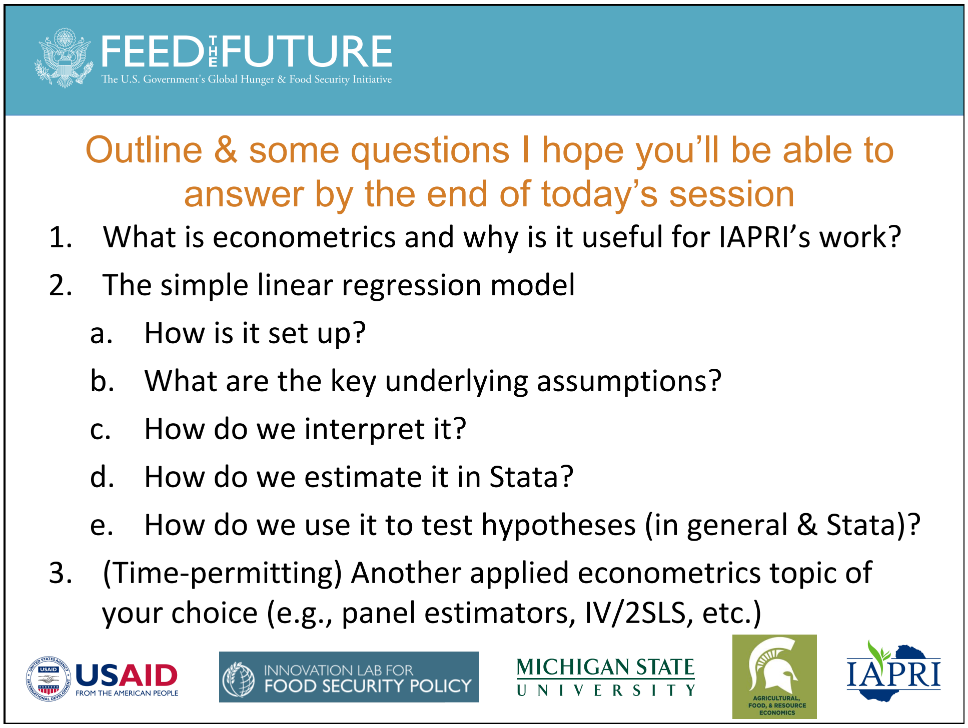
Training materials developed and session facilitated by
Nicole M. Mason
Assistant Professor, Dept. of Agricultural, Food, & Resource Economics
Michigan State University

10:00 AM – 1:00 PM, 25 June 2018
Indaba Agricultural Policy Research Institute
Lusaka, Zambia

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

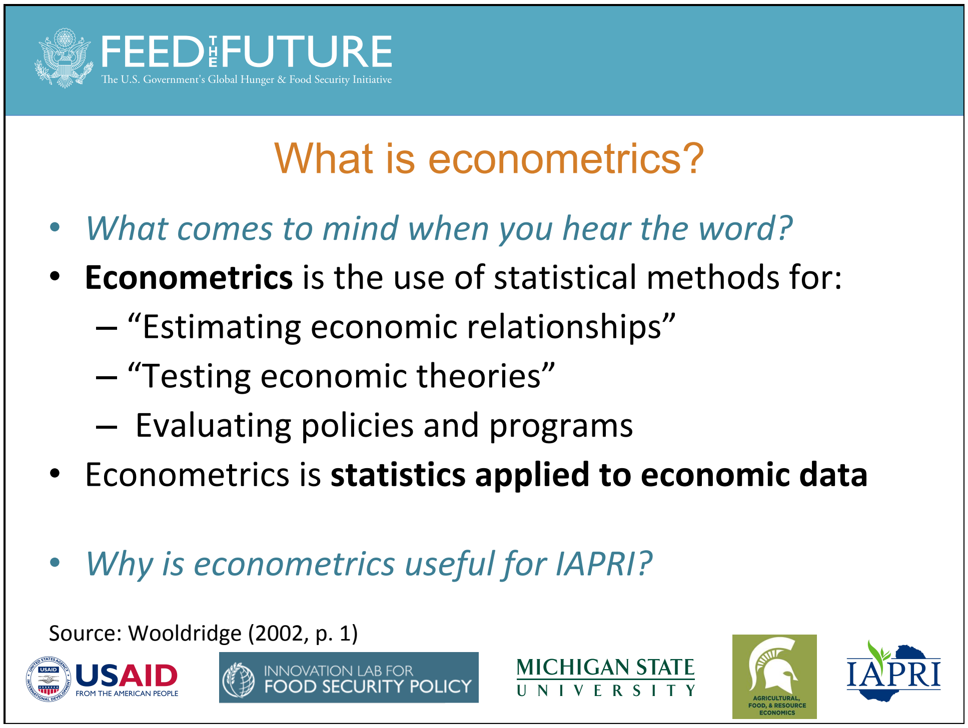
Why this training?

- Requested by IAPRI
- Systematic introduction for non-economist and recently hired economist team members
- Refresher for interested veteran economist team members
  - It has been a while since we last covered these topics (2012-2013)!



Outline & some questions I hope you'll be able to answer by the end of today's session

1. What is econometrics and why is it useful for IAPRI's work?
2. The simple linear regression model
    a. How is it set up?
    b. What are the key underlying assumptions?
    c. How do we interpret it?
    d. How do we estimate it in Stata?
    e. How do we use it to test hypotheses (in general & Stata)?
3. (Time-permitting) Another applied econometrics topic of your choice (e.g., panel estimators, IV/2SLS, etc.)



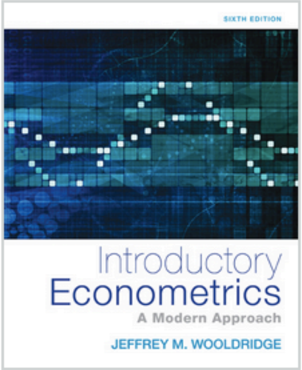
What is econometrics?

- *What comes to mind when you hear the word?*
- **Econometrics** is the use of statistical methods for:
    - "Estimating economic relationships"
    - "Testing economic theories"
    - Evaluating policies and programs
- Econometrics is **statistics applied to economic data**

- *Why is econometrics useful for IAPRI?*

Source: Wooldridge (2002, p. 1)

A great resource (this or an earlier version)



Introductory Econometrics: A Modern Approach, 6th Edition

Jeffrey M. Wooldridge

Published: © 2016
Print ISBN: 9781305270107
Pages: 912
Available

Steps in econometric analysis

For those of you that have done a paper that uses econometrics, what are some steps you went through in going from your research question(s)/ hypothesis(es) through to the analysis and inference?

1. Research question(s)
2. Economic model or other conceptual/theoretical framework
3. Operationalize #2 → econometric model
4. Specify hypotheses to be tested in #3
5. Collect and clean data; create variables
6. Inspect & summarize data
7. Estimate econometric model
8. Interpret results; hypothesis testing & statistical inference

Source: Wooldridge (2002, p. 1)

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Economic model vs. econometric model

- *What's the difference?*
- **Economic model** = "a relationship derived from economic theory or less formal economic reasoning"
  - *Examples?*
- **Econometric model** = "an equation relating the dependent variable to a set of explanatory variables and unobserved disturbances, where unknown population parameters determine the ceteris paribus effect of each explanatory variable"
  - *Examples?*

Source: Wooldridge (2002, p. 794)

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | AGRICULTURAL, FOOD, & RESOURCE ECONOMICS | IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Economic model vs. econometric model (cont'd)

EX) Modeling demand for beef

What does economic theory tell us are likely to be some critical factors affecting an individual's demand for beef?

- Let:
  - *qbeef* = beef quantity demanded
  - *pbeef* = beef price
  - ***pother*** = a vector of other prices (complements, substitutes, etc.)
  - *income* = income
  - ***Z*** = tastes & preferences (proxies in case of econometric model)
- *How could we write down a general function (**economic model**) relating beef quantity demanded and the factors likely to affect it?*
  - $qbeef = f(pbeef, \mathbf{pother}, income, \mathbf{Z})$
- *What would this look like it we were writing it down as an **econometric model** (e.g., a multiple linear regression model)?*
  - $qbeef = \beta_0 + \beta_1 pbeef + \mathbf{pother}\boldsymbol{\beta}_2 + \beta_3 income + \mathbf{Z}\boldsymbol{\beta}_4 + u$

We'll go through notation/ interpretation in a few minutes

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | AGRICULTURAL, FOOD, & RESOURCE ECONOMICS | IAPRI

The simple linear regression model: Motivation

- Let *y* and *x* are two variables that represent some population
- We want to know:
  - **How does *y* change when *x* changes?**
  - **What is the causal effect (*ceteris paribus* effect) of *x* on *y*?**
- *Examples of y and x (in general or in your research)?*

y	***x***
Beef demand	Beef price
Maize yield	Qty of fertilizer used
Child nutrition	HH receives social cash transfer
Crop diversification	HH receives FISP e-voucher
Forest conservation	Community forest mgmt. program

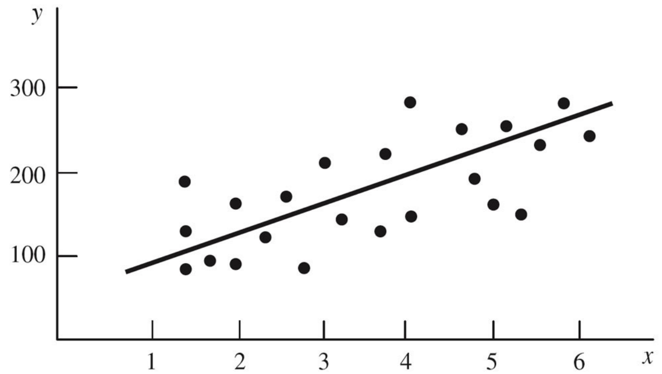
Anatomy of a simple linear regression model

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

β_0 **(intercept)** and β_1 **(slope)** are the population **parameters to be estimated**

- *u* is the **error term** or **disturbance**
  - *u* for "**unobserved**"
  - Represents all factors other than *x* that affect *y*
  - Some use *ε* instead of *u*
- *Terminology for y and x?*

y	***x***
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor
Outcome variable	Covariate



© Cengage Learning

To use data to get unbiased estimates of β_0 and β_1, we have to make some assumptions about the relationship b/w x and u

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

1. $E(u) = 0$ (not restrictive if have an intercept, β_0)
2. *** $E(u|x) = E(u)$ (i.e. the average value of u does not depend on the value of x)

#1 & #2 → $\boxed{E(u|x) = E(u) = 0 \quad \text{(zero conditional mean)}}$

- If this holds, **x is "exogenous"**; but **if *x* is correlated with *u*, *x* is "endogenous"** (we'll come back to this later)

What does this assumption imply below?

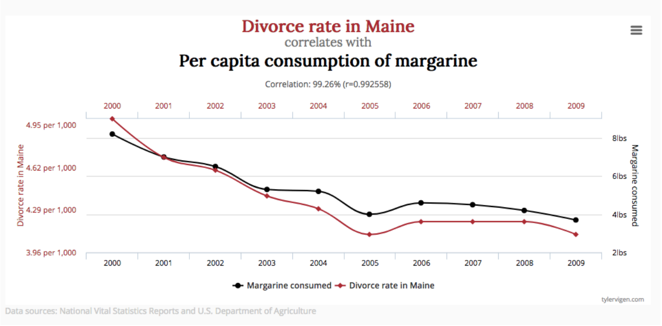
- *yield* = $\beta_0 + \beta_1$*fertilizer* + u, where u is unobserved land quality (*inter alia*)
- *wage* = $\beta_0 + \beta_1$*educ* + u, where u is unobserved ability (*inter alia*)

When is this assumption reasonable?

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Spurious correlations

http://www.tylervigen.com/spurious-correlations



Data sources: National Vital Statistics Reports and U.S. Department of Agriculture

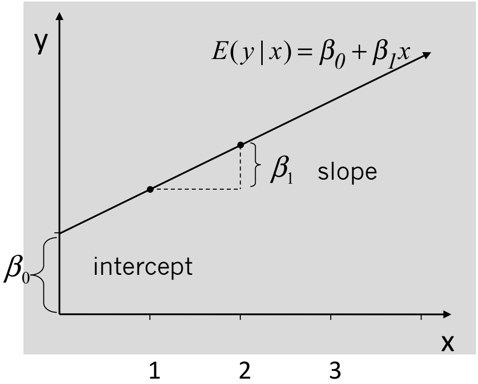
USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY
MICHIGAN STATE UNIVERSITY
AGRICULTURAL, FOOD, & RESOURCE ECONOMICS
IAPRI

What is E(y|x) if we assume E(u|x)=0?

Hint: Apply the rules for conditional expectations.

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

$$\boxed{E(y|x) = \beta_0 + \beta_1 x}$$



What is $\frac{\partial E(y|x)}{\partial x}$ and how do we interpret this result?

$$\frac{\partial E(y|x)}{\partial x} = \beta_1$$

Interpretation: β_1 is the expected change in y given a one unit increase in x, *ceteris paribus* (slope)

What is the interpretation of β_0?

Interpretation: β_0 is the expected value of y when $x = 0$ (intercept)

Why is it called linear regression?

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

- **Linear in parameters**, β_0 and β_1
- Does NOT limit us to linear relationships between x and y
- But rules out models that are **non-linear in parameters**, e.g.:

$$y = \frac{1}{\beta_0 + \beta_1 x} + u$$

$$y = \Phi\left(\beta_0 + \beta_1 x\right) + u$$

$$y = \frac{\beta_0}{\beta_1} x + u$$

Estimating β_0 and β_1

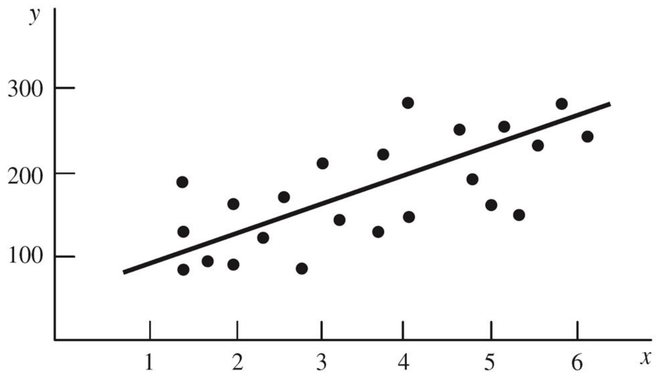
$$y = \beta_0 + \beta_1 x + u$$

- Suppose we have a random sample of size N from the population of interest. Then can write:

$y_i = \beta_0 + \beta_1 x_i + u_i$, $i = 1, 2, 3, \ldots, N$

We don't know β_0 and β_1 but want to estimate them.

How does linear regression use the data in our sample to estimate β_0 and β_1?



© Cengage Learning

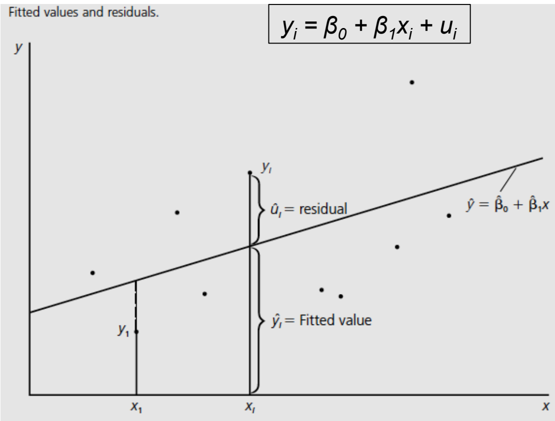
(Ordinary) least squares (OLS) approach

- The estimated values of β$_0$ and β$_1$ are the values that **minimize the sum of squared residuals**
- "Fitted" values of y and residuals:

Fitted (estimated, predicted) values of y: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

Residuals:
$$\hat{u}_i = y_i - \hat{y}_i$$
$$= y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

OLS:
Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize:
$$\sum_{i=1}^{N} \hat{u}_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$



Source: Wooldridge (2002)

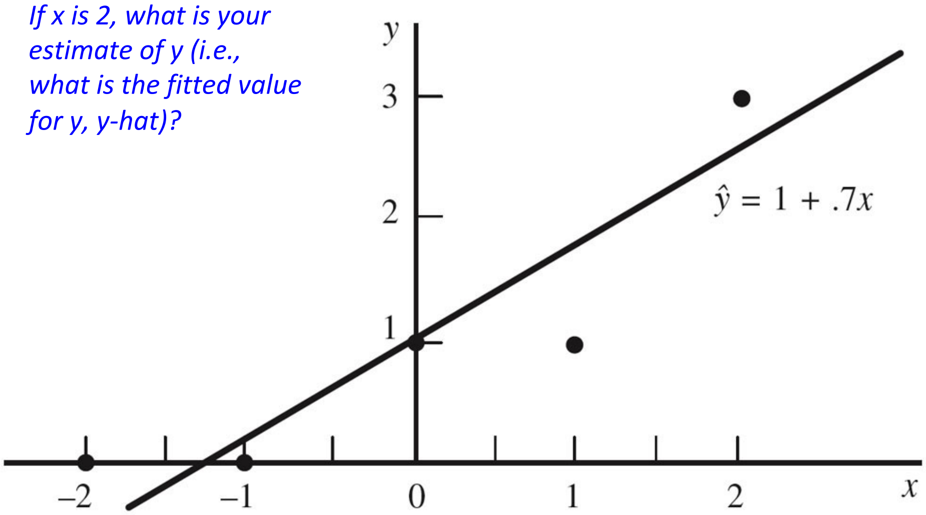
The OLS estimators for β$_0$ and β$_1$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{N} (x_i - \overline{x})^2}$$

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

Fitted (estimated, predicted) values of y: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

If x is 2, what is your estimate of y (i.e., what is the fitted value for y, y-hat)?



© Cengage Learning

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Basic Stata commands

- **regress** *y x* Linear regression of *y* on *x*
  - EX) regress wage educ
- **predict** ***newvar1*****, xb** Compute fitted values
  - EX) predict wagehat, xb (I just made up the name wagehat)
- **predict** ***newvar2*****, resid** Compute residuals
  - EX) predict uhat, resid (I just made up the name uhat)

Obtaining OLS estimates – example (Stata)

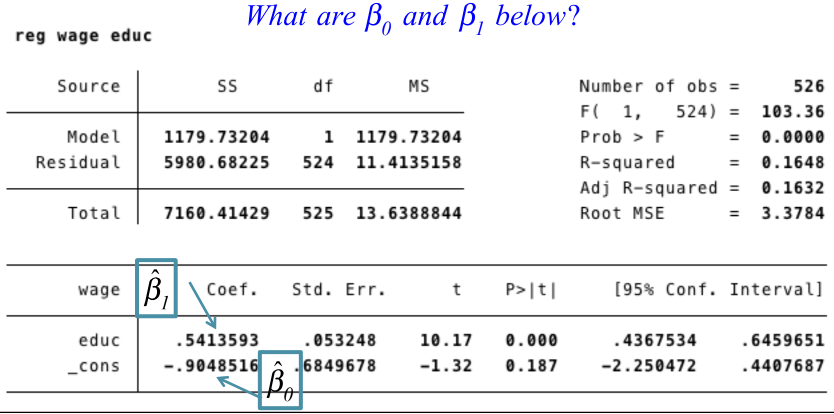
Wooldridge (2002) Example 2.4: Wage and education

Use Stata to run the simple linear regression of wage (*y*) on educ (*x*).

$$\boxed{wage_i = \beta_0 + \beta_1 educ_i + u_i}$$

Command: regress wage educ (or: reg wage educ)

What are $\hat{\beta}_0$ and $\hat{\beta}_1$ below?



```
reg wage educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Practice time!

1. Open the dataset "WAGE1.DTA" in Stata
   a. Type the command "describe" to see what variables are in the dataset
   b. Estimate the model on the previous slide (reg wage educ)
   c. Find and interpret (put in a sentence!) the estimates of β_0 and β_1 in the regression output
2. Open the dataset "RALS1215_training.dta" in Stata
   a. Use "describe" to see what variables are in the dataset
   b. Regress the variable for gross value of crop production on the variable for landholding size
   c. Find and interpret (put in a sentence!) the estimates of β_0 and β_1 in the regression output

USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY
MICHIGAN STATE UNIVERSITY
AGRICULTURAL, FOOD, & RESOURCE ECONOMICS
IAPRI

Total, explained, & residual sum of squares, R^2

Total sum of squares: $$SST \equiv \sum_{i=1}^{N} (y_i - \bar{y})^2$$

Explained sum of squares: $$SSE \equiv \sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2$$

Residual sum of squares: $$SSR \equiv \sum_{i=1}^{N} \hat{u}_i^2$$

$$\boxed{SST = SSE + SSR}$$

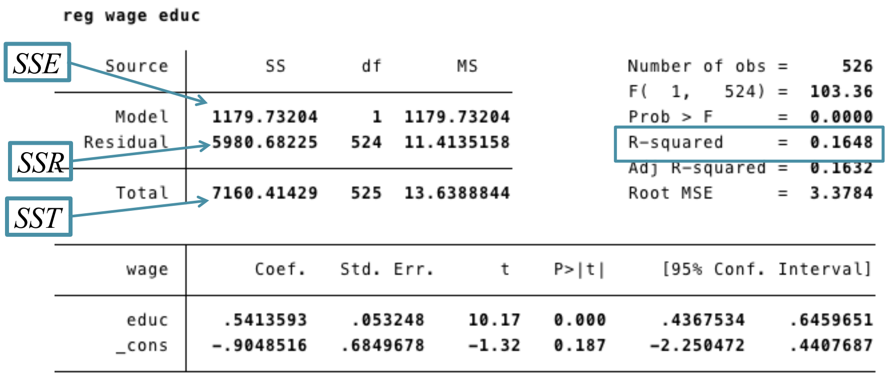
Proof is on p. 39 of Wooldridge (2002)

Coefficient of determination or *R*²: *Interpretation?*

$$\boxed{R^2 = SSE / SST = 1 - (SSR / SST)}$$

The proportion of the sample variation in y that is explained by x

SST (total SS), SSE (explained SS), SSR (residual SS), and *R*² in Stata



```
reg wage educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

$$\boxed{SST = SSE + SSR} \qquad \boxed{R^2 = SSE / SST = 1 - (SSR / SST)}$$

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Practice time!

Look at the output from your RALS-based regression:

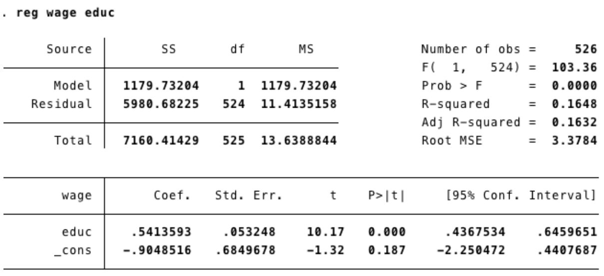
1. What is the SSE?
2. What is the SSR?
3. What is the SST?
4. What is the R-squared? (Find the actual number and then check that it equals SSE/SST)
5. Interpret the R-squared (put it in a sentence)

USAID FROM THE AMERICAN PEOPLE · INNOVATION LAB FOR FOOD SECURITY POLICY · MICHIGAN STATE UNIVERSITY · AGRICULTURAL, FOOD, & RESOURCE ECONOMICS · IAPRI

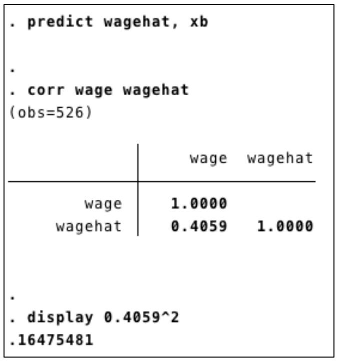
Why is it called R^2?

- Letter R sometimes used to refer to correlation coefficient (others use *ρ*, which is "rho")

R^2 is the squared sample correlation coefficient between y_i and $\hat{y}_i$



```
. reg wage educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```



```
. predict wagehat, xb


.
. corr wage wagehat
(obs=526)

             |     wage  wagehat
-------------+------------------
        wage |   1.0000
     wagehat |   0.4059   1.0000



.
. display 0.4059^2
.16475481
```

Practice time!

Immediately after your RALS regression command, use the previous slide as a guide and:

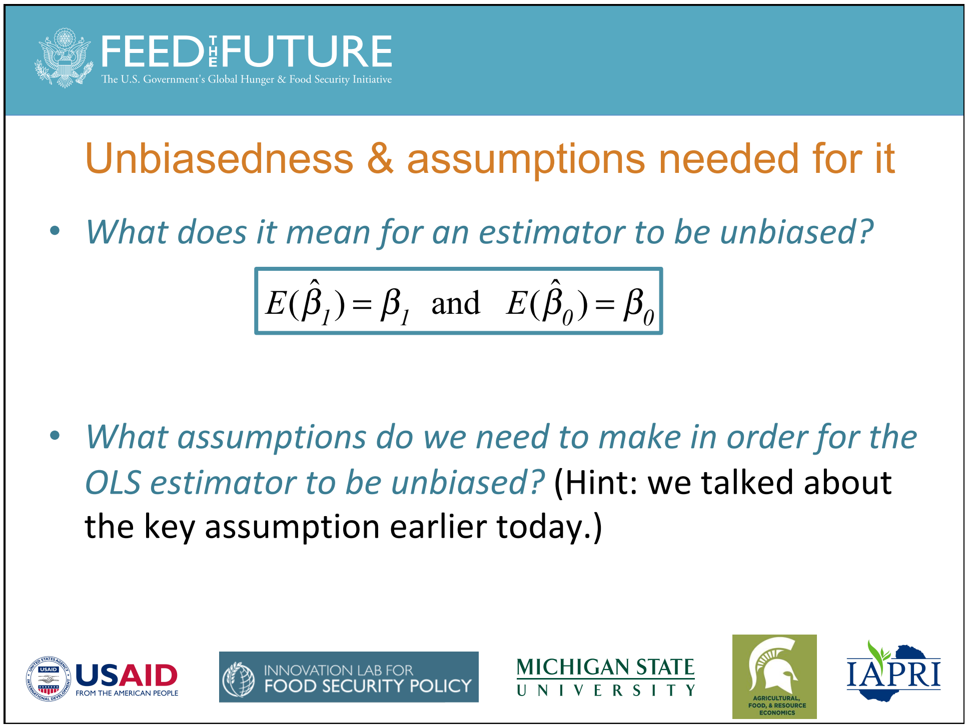
1. Compute the predicted values of gvharv, calling them gvharvhat
2. Compute the correlation coefficient between gvharv and gvharvhat
3. Square this correlation coefficient (using the Stata "display" command)
4. Compare the R-squared you just computed "by hand" to the Stata-generated R-squared in the regression output. Do they match?

"My R-squared it too low!"

Does a low R^2 mean the regression results are useless? Why or why not?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\hat{\beta}_1$ may still be good (unbiased) estimate of *ceteris paribus* (causal) effect of *x* on *y* even if R^2 is low



FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

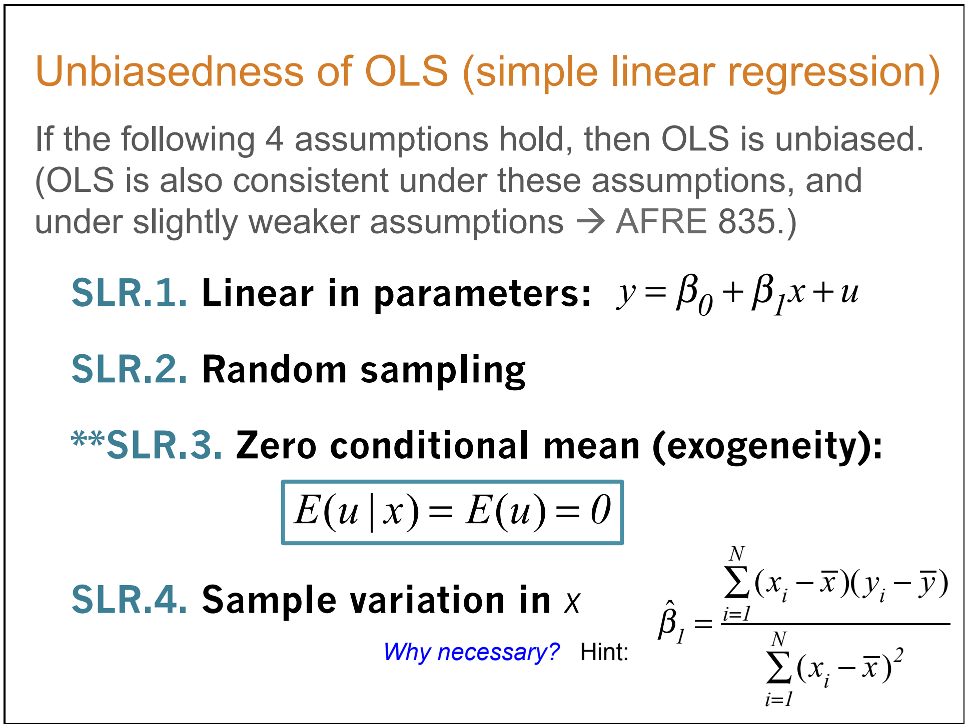
Unbiasedness & assumptions needed for it

- *What does it mean for an estimator to be unbiased?*

$$\boxed{E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0}$$

- *What assumptions do we need to make in order for the OLS estimator to be unbiased?* (Hint: we talked about the key assumption earlier today.)

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | AGRICULTURAL, FOOD, & RESOURCE ECONOMICS | IAPRI



Unbiasedness of OLS (simple linear regression)

If the following 4 assumptions hold, then OLS is unbiased. (OLS is also consistent under these assumptions, and under slightly weaker assumptions → AFRE 835.)

SLR.1. Linear in parameters: $y = \beta_0 + \beta_1 x + u$

SLR.2. Random sampling

****SLR.3. Zero conditional mean (exogeneity):**

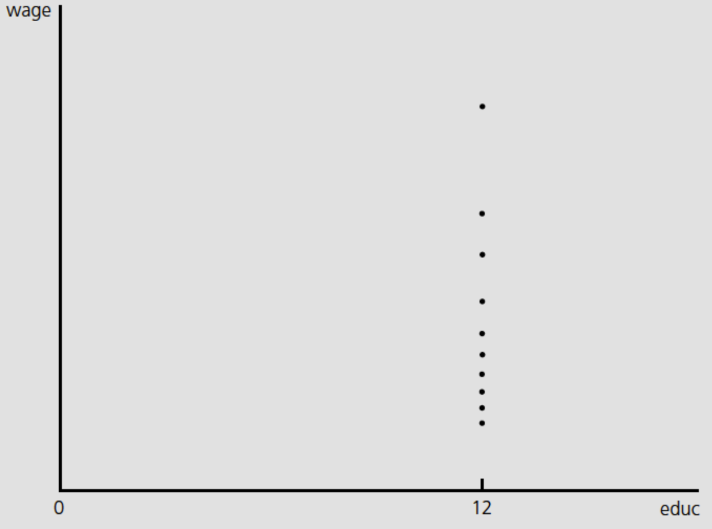
$$\boxed{E(u \mid x) = E(u) = 0}$$

SLR.4. Sample variation in x

Why necessary? Hint: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$

Can't estimate slope parameter if no variation in x

A scatterplot of wage against education when $educ_i = 12$ for all i.



Source: Wooldridge (2002)

OLS estimators for β_0 and β_1 are unbiased under SLR.1-SLR.4

$$\boxed{E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0}$$

- *What does unbiasedness mean in plain language?*
- The key assumption is $E(u|x) = E(u) = 0$ (zero conditional mean/exogeneity) – SLR.3
  - Under SLR.1-SLR.4, OLS estimate of β_1 is the **causal effect (*ceteris paribus* effect)** of x on y
  - If $E(u|x) \neq E(u)$, then x is **endogenous** to y → OLS estimates biased → need to take other measures to deal with this (IV/2SLS, panel data methods, etc.)

If make one more assumption - homoskedasticity (SLR.5) - then OLS is "BLUE"

Let $V(u) = \sigma^2$

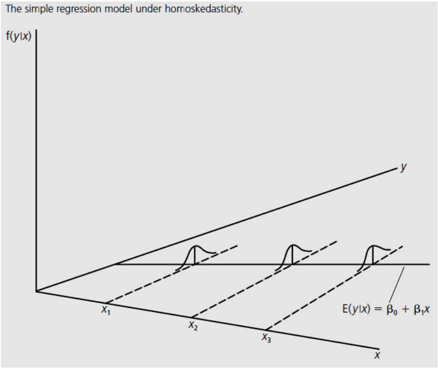
SLR.5. Homoskedasticity (constant variance):

$$\boxed{V(u \mid x) = V(u) = \sigma^2}$$

Which implies:

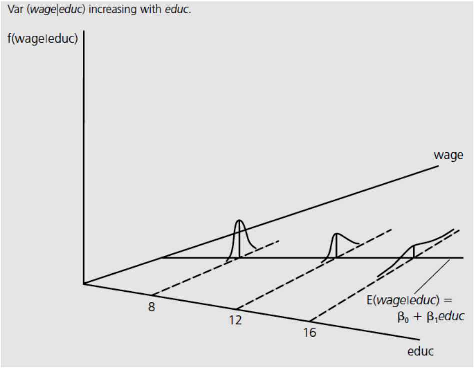
$$\boxed{V(y \mid x) = V(u \mid x) = \sigma^2}$$

Homoskedasticity



The simple regression model under homoskedasticity.

Heteroskedasticity



Var (wage|educ) increasing with educ.

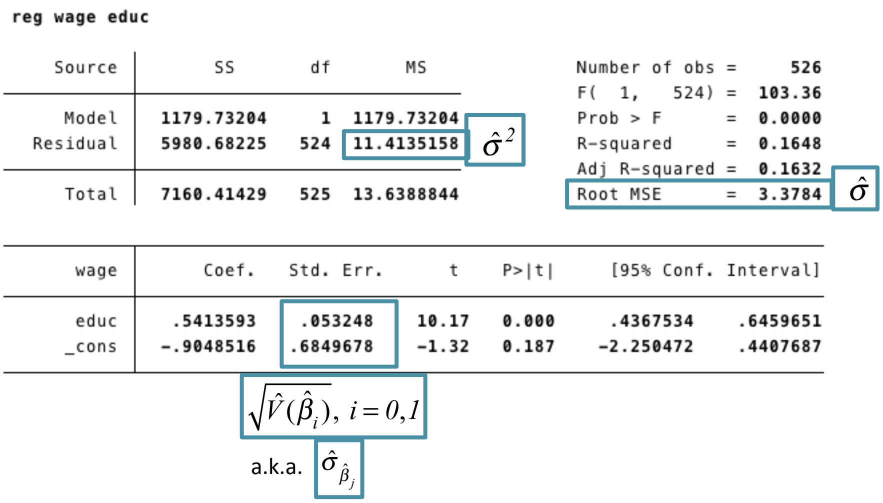
Source: Wooldridge (2002)

If SLR.1 through SLR.5 hold, then OLS is "BLUE"

- **Best** (most efficient, i.e., smallest variance)
- **Linear** (linear function of the y_i)
- **Unbiased**
- **Estimator**

Also, if homoskedastic, then the "regular" variance formulas for OLS estimators are correct (i.e., are unbiased estimators for true variances). (If heteroskedastic, then these formulas and the regular standard errors reported by Stata biased → too small.) *Why is this a problem?*

In Stata



```
reg wage educ

      Source |       SS       df       MS              Number of obs =     526
-------------+------------------------------           F(  1,   524) =  103.36
       Model |  1179.73204     1  1179.73204           Prob > F      =  0.0000
    Residual |  5980.68225   524  11.4135158           R-squared     =  0.1648
-------------+------------------------------           Adj R-squared =  0.1632
       Total |  7160.41429   525  13.6388844           Root MSE      =  3.3784

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        educ |   .5413593    .053248    10.17   0.000     .4367534    .6459651
       _cons |  -.9048516   .6849678    -1.32   0.187    -2.250472    .4407687
------------------------------------------------------------------------------
```

$\hat{\sigma}^2$

$\hat{\sigma}$

$\sqrt{\hat{V}(\hat{\beta}_i)},\ i = 0,1$

a.k.a. $\hat{\sigma}_{\hat{\beta}_j}$

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Practice time!

Locate the standard errors for your estimates in the RALS-based regression

Question: Why do we need these standard errors? How will we use them?

USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY
MICHIGAN STATE UNIVERSITY
AGRICULTURAL, FOOD, & RESOURCE ECONOMICS
IAPRI

The sampling distributions of the OLS estimators

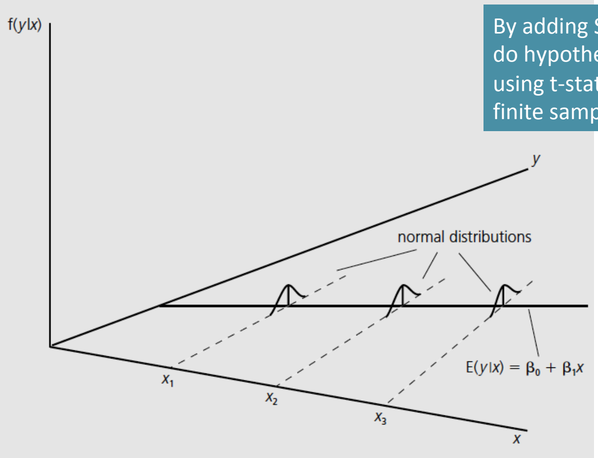
- **By the Central Limit Theorem**, under assumptions SLR.1-SLR.5, the OLS estimators are **asymptotically** (i.e., as *N*→∞) **normally distributed**
- If we **add one more assumption**, then we can obtain the sampling distribution of the OLS estimators in **finite samples**

SLR.6. Normality: The population error, ***u*, is independent of *x*** and is **normally distributed** with $E(u)=0$ and $V(u)=\sigma^2$, i.e.: $\boxed{u \sim Normal(0,\sigma^2)}$

SLR.1-SLR.6 = “classical linear model assumptions”

- CLM = SLR.1-SLR.6
- CLM assumptions imply $\boxed{y\,|\,x \sim Normal(\beta_0 + \beta_1 x,\ \sigma^2)}$

The homoskedastic normal distribution with a single explanatory variable.



By adding SLR.6, we can do hypothesis testing using t-statistics even in finite samples

Source: Wooldridge (2002)

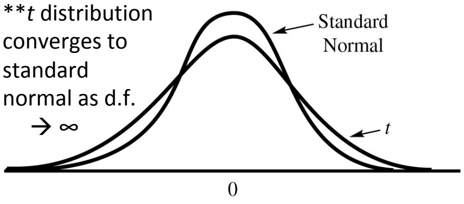
FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

T-statistic refresher

Under SLR.1-SLR.6 (simple linear regression case):

$$\boxed{T = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t \text{ with } N-2 \text{ d.f.}}$$

**t distribution converges to standard normal as d.f. → ∞



Replace β_j with the value under the null hypothesis

e.g., $H_0 : \beta_j = 0, H_1 : \beta_j \neq 0$

© Cengage Learning

USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY
MICHIGAN STATE UNIVERSITY
AGRICULTURAL, FOOD, & RESOURCE ECONOMICS
IAPRI

P-value refresher

What is a p-value?

- "The smallest significance level at which the null hypothesis can be rejected" (Wooldridge 2002, p. 800)
- Significance level = P(Type I error)
  = P(reject the null when the null is actually true)
- Suppose you are conducting your hypothesis test using the 10% significance level as your cut-off for statistical significance
- *What do you conclude if p-value > 0.10? Do you reject the null hypothesis of fail to reject it?*
- *What do you conclude if p-value $\leq$ 0.10?*

Type I vs. Type II error refresher

		REALITY	
		NULL HYPOTHESIS	
		TRUE	FALSE
STUDY FINDINGS	TRUE	☺	Type II error (β) 'False negative'
	FALSE	Type I error (α) 'False positive'	☺

Type I error: reject H_0 when H_0 is true

Probability: α (significance level)

Type II error: fail to reject H_0 when H_0 is false

Probability: β (1- β = power of the test) – different β!

T-statistics and p-values in Stata output

```
. reg bwght cigs

      Source |       SS       df       MS              Number of obs =    1388
-------------+------------------------------           F(  1,  1386) =   32.24
       Model |  13060.4194     1  13060.4194           Prob > F      =  0.0000
    Residual |    561551.3  1386  405.159668           R-squared     =  0.0227
-------------+------------------------------           Adj R-squared =  0.0220
       Total |   574611.72  1387  414.283864           Root MSE      =  20.129

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|
-------------+----------------------------------------------------------------
        cigs |  -.5137721   .0904909    -5.68   0.000
       _cons |   119.7719   .5723407   209.27   0.000
------------------------------------------------------------------------------
```

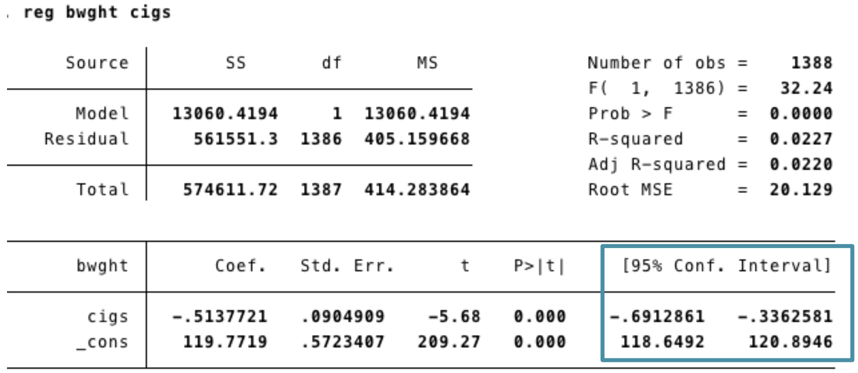
The p-values reported by Stata are for $H_0: \beta_j = 0$ vs. $H_1: \beta_j \neq 0$

Practice time!

Using your RALS-based regression output:

1. What is the value of the t-statistic that you would use to test H_0: β_{land}=0 vs. H_1: β_{land}≠0? Find it in the regression output and also calculate it using the formula a few slides back.
2. Conduct this hypothesis test at the 10% level (using the p-value reported in Stata). What do you conclude?
3. What does this mean in practice?

95% confidence intervals in Stata output

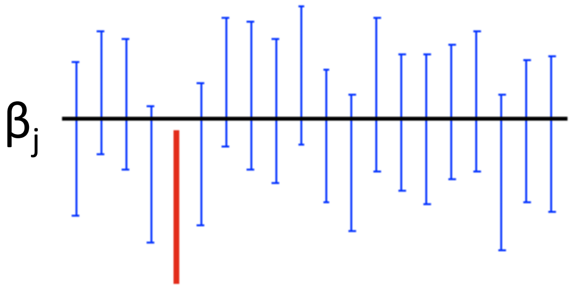


```
. reg bwght cigs

      Source |       SS           df       MS      Number of obs   =      1388
-------------+----------------------------------   F(  1,  1386)   =     32.24
       Model |  13060.4194         1  13060.4194   Prob > F        =    0.0000
    Residual |    561551.3      1386  405.159668   R-squared       =    0.0227
-------------+----------------------------------   Adj R-squared   =    0.0220
       Total |   574611.72      1387  414.283864   Root MSE        =    20.129

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5137721   .0904909    -5.68   0.000    -.6912861   -.3362581
       _cons |   119.7719   .5723407   209.27   0.000     118.6492    120.8946
------------------------------------------------------------------------------
```

Interpretation of 95% confidence interval (CI): "If random samples were obtained over and over again, with β_j^L and β_j^U computed each time, then the (unknown) population value β_j would lie in the interval $[\beta_j^L, \beta_j^U]$ for 95% of the samples. Unfortunately, for the single sample that we use to construct the CI, we do not know whether β_j is actually contained in the interval. We hope we have obtained a sample that is one of the 95% of all samples where the interval estimate contains β_j, but we have no guarantee." (Wooldridge 2002, p. 134)

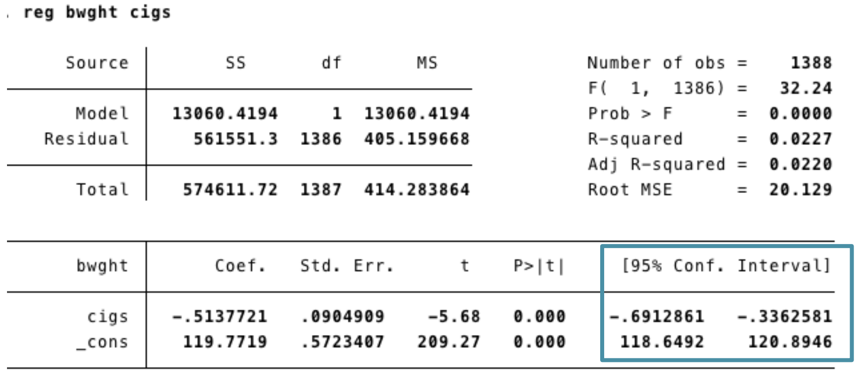
FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Confidence interval interpretation



Twenty (20) 95% CIs for β_j

95% confidence intervals in Stata output



```
. reg bwght cigs

      Source |       SS           df       MS      Number of obs   =      1388
-------------+----------------------------------   F(1, 1386)      =     32.24
       Model |  13060.4194         1  13060.4194   Prob > F        =    0.0000
    Residual |    561551.3      1386  405.159668   R-squared       =    0.0227
-------------+----------------------------------   Adj R-squared   =    0.0220
       Total |   574611.72      1387  414.283864   Root MSE        =    20.129

------------------------------------------------------------------------------
       bwght |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        cigs |  -.5137721   .0904909    -5.68   0.000    -.6912861   -.3362581
       _cons |   119.7719   .5723407   209.27   0.000     118.6492    120.8946
------------------------------------------------------------------------------
```

95% CIs are **useful for testing hypotheses at the 5% sig. level about values of β_j under the null other than zero** (against the corresponding 2-sided alternative hypothesis)

If the value of β_j under the null falls within the 95% CI, what would you conclude?

- Fail to reject the null in favor of the alternative at the 5% level

If the value of β_j under the null is outside of the 95% CI, what would you conclude?

- Reject the null in favor of the alternative at the 5% level

---

Practice time!

Use the 95% CI in your RALS-based regression output to test a null hypothesis of your choice for β_j (against its corresponding 2-sided alternative hypothesis). Conduct your hypothesis test at the 5% significance level. What do you conclude?

(Time-permitting) What else do you want to cover?

Thank you for your attention & participation!

Nicole Mason (masonn@msu.edu)
Assistant Professor
Department of Agricultural, Food, & Resource Economics (AFRE)
Michigan State University (MSU)

Main reference

- Wooldridge, J. (2002). *Introductory econometrics: A modern approach (2nd edition).* Cincinnati, OH: South-Western College Pub.

EXTRA SLIDES

Aside: NPR "Hidden Brain" example of a natural experiment, and when it might be reasonable to assume $E(u|x)=E(u)$

- Listen for the following:
  - *What is the dependent variable?*
  - *What is the main explanatory variable of interest?*
  - *Why might it be reasonable to assume $E(u|x)=E(u)$ here?*
  - *What is a natural experiment?*
- Dependent variable: cognitive function of elderly
- Main explanatory variable: wealth
- $E(u|x)=E(u)$ might be reasonable – Congress computational mistake – people in one cohort got higher benefits that next cohort (level of benefits shouldn't be correlated with unobservables)

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Aside: Natural experiments

A natural experiment occurs when some exogenous event—often a change in government policy—changes the environment in which individuals, families, firms, or cities operate. A natural experiment always has a control group, which is not affected by the policy change, and a treatment group, which is thought to be affected by the policy change. Unlike with a true experiment, where treatment and control groups are randomly and explicitly chosen, the control and treatment groups in natural experiments arise from the particular policy change. (Wooldridge, 2002: 417)

USAID FROM THE AMERICAN PEOPLE | INNOVATION LAB FOR FOOD SECURITY POLICY | MICHIGAN STATE UNIVERSITY | AGRICULTURAL, FOOD, & RESOURCE ECONOMICS | IAPRI

Putting it all together: simple linear regression

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

OLS estimators for β_0 and β_1:

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \quad \boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Expected values (under SLR.1-SLR.4):

$$\boxed{E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0}$$

Sample variances (under SLR.1-SLR.5):

$$\boxed{\hat{V}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \quad \boxed{\hat{V}(\hat{\beta}_0) = \frac{\hat{\sigma}^2 N^{-1} \sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}} \quad \boxed{\text{where } \hat{\sigma}^2 = \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{SSR}{N-2}}$$

$\hat{\sigma}$ is the **standard error** of the regression

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

A useful cheat-sheet for interpreting models with **logged** variables

Summary of Functional Forms Involving Logarithms $y = \beta_0 + \beta_1 x + u$

Model	Dependent Variable	Independent Variable	Interpretation of β_1
level-level	y	x	$\boxed{\beta_1 = \frac{\Delta y}{\Delta x}}$ $\Delta y = \beta_1 \Delta x$
level-log	y	$\log(x)$	$\boxed{\frac{\beta_1}{100} = \frac{\Delta y}{\%\Delta x}}$ $\Delta y = (\beta_1/100)\%\Delta x$
log-level	$\log(y)$	x	$\boxed{100\beta_1 = \frac{\%\Delta y}{\Delta x}}$ $\%\Delta y = (100\beta_1)\Delta x$
log-log	$\log(y)$	$\log(x)$	$\boxed{\beta_1 = \frac{\%\Delta y}{\%\Delta x}}$ $\%\Delta y = \beta_1 \%\Delta x$

Source: Wooldridge (2002)

USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY
MICHIGAN STATE UNIVERSITY
AGRICULTURAL, FOOD, & RESOURCE ECONOMICS
IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

Acknowledgements

This training was made possible by the generous support of the American People provided to the Feed the Future Innovation Lab for Food Security Policy [grant number AID-OAA-L-13-00001] through the United States Agency for International Development (USAID).

The contents are the responsibility of the training material author and do not necessarily reflect the views of USAID or the United States Government.

USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY
MICHIGAN STATE UNIVERSITY
AGRICULTURAL, FOOD, & RESOURCE ECONOMICS
IAPRI

FEED THE FUTURE
The U.S. Government's Global Hunger & Food Security Initiative

www.feedthefuture.gov

USAID FROM THE AMERICAN PEOPLE
INNOVATION LAB FOR FOOD SECURITY POLICY
MICHIGAN STATE UNIVERSITY
AGRICULTURAL, FOOD, & RESOURCE ECONOMICS
IAPRI