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

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?

A great resource (this or an earlier version)


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

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?


---

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:
  - \( q_{beef} \) = beef quantity demanded
  - \( p_{beef} \) = beef price
  - \( p_{other} \) = 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?
  - \( q_{beef} = f(p_{beef}, p_{other}, income, Z) \)

- What would this look like if we were writing it down as an **econometric model** (e.g., a multiple linear regression model)?
  - \( q_{beef} = \beta_0 + \beta_1p_{beef} + \beta_2p_{other} + \beta_3income + Z\beta + u \)

We’ll go through notation/interpretation in a few minutes
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)?

<table>
<thead>
<tr>
<th>$y$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beef demand</td>
<td>Beef price</td>
</tr>
<tr>
<td>Maize yield</td>
<td>Qty of fertilizer used</td>
</tr>
<tr>
<td>Child nutrition</td>
<td>HH receives social cash transfer</td>
</tr>
<tr>
<td>Crop diversification</td>
<td>HH receives FISP e-voucher</td>
</tr>
<tr>
<td>Forest conservation</td>
<td>Community forest mgmt. program</td>
</tr>
</tbody>
</table>

Anatomy of a simple linear regression model
$y = \beta_0 + \beta_1 x + u$
• $u$ is the error term or disturbance
  -- $u$ for “unobserved”
  -- Represents all factors other than $x$ that affect $y$
  -- Some use $\epsilon$ instead of $u$
• Terminology for $y$ and $x$?

<table>
<thead>
<tr>
<th>$y$</th>
<th>$x$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent variable</td>
<td>Independent variable</td>
</tr>
<tr>
<td>Explained variable</td>
<td>Explanatory variable</td>
</tr>
<tr>
<td>Response variable</td>
<td>Control variable</td>
</tr>
<tr>
<td>Predicted variable</td>
<td>Predictor variable</td>
</tr>
<tr>
<td>Regressand</td>
<td>Regressor</td>
</tr>
<tr>
<td>Outcome variable</td>
<td>Covariate</td>
</tr>
</tbody>
</table>
To use data to get unbiased estimates of $\beta_0$ and $\beta_1$, 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, $\beta_0$)
2. *** $E(u|x) = E(u)$ (i.e. the average value of $u$ does not depend on the value of $x$)

#1 & #2 $\Rightarrow E(u|x) = E(u) = 0$ (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?
What is $E(y|x)$ if we assume $E(u|x)=0$?

Hint: Apply the rules for conditional expectations.

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

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

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: $\beta_1$ is the expected change in $y$ given a one unit increase in $x$, ceteris paribus (slope).

What is the interpretation of $\beta_0$?

Interpretation: $\beta_0$ 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**, \( \beta_0 \) and \( \beta_1 \)
- Does NOT limit us to linear relationships between \( x \) and \( y \)
- But rules out models that are non-linear in parameters, e.g.:

\[
\begin{align*}
  y &= \frac{1}{\beta_0 + \beta_1 x} + u \\
  y &= \Phi(\beta_0 + \beta_1 x) + u \\
  y &= \frac{\beta_0}{\beta_1} x + u 
\end{align*}
\]

Estimating \( \beta_0 \) and \( \beta_1 \)

\[ y = \beta_0 + \beta_1 x + u \]

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

\[ y_i = \beta_0 + \beta_1 x_i + u_i, \quad i = 1, 2, 3, \ldots, N \]

We don’t know \( \beta_0 \) and \( \beta_1 \) but want to estimate them.

How does linear regression use the data in our sample to estimate \( \beta_0 \) and \( \beta_1 \)?
(Ordinary) least squares (OLS) approach
• The estimated values of $\beta_0$ and $\beta_1$ are the values that minimize the sum of squared residuals
• “Fitted” values of $y$ and residuals:

\[
\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i
\]

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

Residuals:
\[
\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i
\]

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


The OLS estimators for $\beta_0$ and $\beta_1$

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

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]
If \( x \) is 2, what is your estimate of \( y \) (i.e., what is the fitted value for \( y \), \( y \)-hat)?

\[ \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \]

**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`)

---

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

\[ \hat{y} = 1 + .7x \]
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 \( (x) \).

\[
\text{wage}_i = \beta_0 + \beta_1 \text{educ}_i + u_i
\]

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

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
<th>Number of obs = 526</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1179.73284</td>
<td>1</td>
<td>1179.73284</td>
<td>F( 1, 524) = 103.36</td>
</tr>
<tr>
<td>Residual</td>
<td>5900.68225</td>
<td>524</td>
<td>11.4135150</td>
<td>Prob &gt; F = 0.0000</td>
</tr>
<tr>
<td>Total</td>
<td>7160.41429</td>
<td>525</td>
<td>13.6388844</td>
<td>R-squared = 0.1648</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Adj R-squared = 0.1632</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Root MSE = 3.784</td>
</tr>
</tbody>
</table>

|         | Coef.   | Std. Err. | t     | P>|t| | [95% Conf. Interval] |
|---------|---------|-----------|------|------|----------------------|
| wage    | \( \hat{\beta}_0 \) | .5413593 | .053248 | 10.17 | .000 | .4367534 | .6459651 |
| educ    | \( \hat{\beta}_1 \) | .9048510 | .0849670 | -1.32 | .187 | -2.250472 | .4407687 |
| _cons   |         | -.9048510 | .0849670 | -1.32 | .187 | -2.250472 | .4407687 |

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 \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) 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 \( \hat{\beta}_0 \) and \( \hat{\beta}_1 \) in the regression output
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:** \[ SSR \equiv \sum_{i=1}^{N} \epsilon_i^2 \]

\[ SST = SSE + SSR \]

Proof is on p. 39 of Wooldridge (2002)

**Coefficient of determination or $R^2$:**

\[ R^2 = SSE / SST = 1 - (SSR / SST) \]

Interpretation?

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
```

<table>
<thead>
<tr>
<th>Source</th>
<th>SS</th>
<th>df</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1179.73204</td>
<td>1</td>
<td>1179.73204</td>
</tr>
<tr>
<td>Residual</td>
<td>5900.68225</td>
<td>524</td>
<td>11.4135150</td>
</tr>
<tr>
<td>Total</td>
<td>7160.41429</td>
<td>525</td>
<td>13.6388444</td>
</tr>
</tbody>
</table>

Number of obs = 526

F( 1, 524) = 103.36

Prob > F = 0.0000

R-squared = 0.1648

Adj R-squared = 0.1632

Root MSE = 3.3744

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

\[ SST = SSE + SSR \]

\[ R^2 = SSE / SST = 1 - (SSR / SST) \]
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)

Why is it called $R^2$?

- Letter R sometimes used to refer to correlation coefficient (others use $\rho$, which is “rho”)

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

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

1. Compute the predicted values of gvhav, calling them gvhavhat
2. Compute the correlation coefficient between gvhav and gvhavhat
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
Unbiasedness & assumptions needed for it
• What does it mean for an estimator to be unbiased?
  \[ E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0 \]

• 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.)

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 \( \rightarrow \) AFRE 835.)

**SLR.1. Linear in parameters:** \( y = \beta_0 + \beta_1 x + u \)

**SLR.2. Random sampling**

**SLR.3. Zero conditional mean (exogeneity):**
  \[ E(u \mid x) = E(u) = 0 \]

**SLR.4. Sample variation in** \( x \)
  \[ \hat{\beta}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2} \]

Why necessary? Hint:
Can’t estimate slope parameter if no variation in x

![A scatterplot of wage against education when educ = 12 for all i.](image)


OLS estimators for $\beta_0$ and $\beta_1$ are unbiased under SLR.1-SLR.4

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

- 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 $\beta_1$ 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):

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

Which implies:

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

---

**Homoskedasticity**

The simple regression model under homoskedasticity.

**Heteroskedasticity**

Var (independent variable) increasing with $x$.

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 $\rightarrow$ too small.) *Why is this a problem?*

---

**In Stata**

```
reg wage educ

Source | SS    | df | MS
---|-------|----|---
Model  | 1179.73204 | 1  | 1179.73204
Residual| 5900.68225  | 524| 11.4135158
Total  | 7160.41429  | 525| 13.6388844

Number of obs = 526
F( 1, 524) = 103.36
Prob > F = 0.0000
R-squared = 0.1648
Adj R-squared = 0.1632

Root MSE = 3.3784
```

```
| wage | Coef. | Std. Err. | t | P>|t| | [95% Conf. Interval] |
|------|-------|-----------|---|-----|----------------------|
| educ | .5413593 | .053248  | 10.17 | 0.000 | .4367534 -.6459651 |
| _cons | -.9004516 | .6849678 | -1.32 | .187 | -2.250472 .4407687 |
```

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

a.k.a. $$\hat{\sigma}_{\hat{\beta}_i}$$
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?*

---

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 $N \to \infty$) *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.:

\[
 u \sim \text{Normal}(0, \sigma^2)
\]
SLR.1-SLR.6 = “classical linear model assumptions”

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

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

T-statistic refresher

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

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

Replace $\beta_j$ with the value under the null hypothesis. e.g., $H_0 : \beta_j = 0, H_1 : \beta_j \neq 0$
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 or fail to reject it?
• What do you conclude if p-value < 0.10?

Type I vs. Type II error refresher

<table>
<thead>
<tr>
<th>REALITY</th>
<th>NULL HYPOTHESIS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>TRUE</td>
</tr>
<tr>
<td>STUDY FINDINGS</td>
<td>TRUE</td>
</tr>
<tr>
<td></td>
<td>FALSE</td>
</tr>
</tbody>
</table>

Type I error: reject $H_0$ when $H_0$ is true
Probability: $\alpha$ (significance level)

Type II error: fail to reject $H_0$ when $H_0$ is false
Probability: $\beta$ (1 - $\beta$ = power of the test) – different $\beta$!
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 $H_0: \beta_{\text{land}} = 0$ vs. $H_1: \beta_{\text{land}} \neq 0$? 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?
### Interpretation of 95% Confidence Interval (CI):

“If random samples were obtained over and over again, with $\beta_j^L$ and $\beta_j^U$ computed each time, then the (unknown) population value $\beta_j$ 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 $\beta_j$ 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 $\beta_j$, but we have no guarantee.” (Wooldridge 2002, p. 134)
95% confidence intervals in Stata output

```
. reg bwght cigs

Source | SS      | df | MS
--------+---------+----+------
Model   | 13060.4194 | 1  | 13060.4194
Residual| 561551.3   | 1386| 405.159668
Total   | 574611.72  | 1387| 414.203864

Number of obs = 1388
F( 1, 1386) = 32.24
Prob > F    = 0.0000
R-squared   = 0.8227
Adj R-squared = 0.8220
Root MSE    = 20.129

bwght   Coef.        Std. Err.     t    P>|t|    [95% Conf. Interval]
--------+-------------------+------------+-------+-------------------+--------------------------
cigs     -.5137721     .0904999   -5.68 0.000   -.6912061   -.3363481
_cens    119.7719    .5723497      209.27 0.000  118.6492     120.8946
```

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

If the value of $\beta_j$ 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 $\beta_j$ 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 $\beta_j$ (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

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)

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)
Putting it all together: simple linear regression

\[ y = \beta_0 + \beta_1 x + u \]

**OLS estimators for \( \beta_0 \) and \( \beta_1 \):**

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

\[
\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\]

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

\[ E(\hat{\beta}_1) = \beta_1 \text{ and } E(\hat{\beta}_0) = \beta_0 \]

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

\[
\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}
\]

\[
\hat{\sigma}^2_{\hat{\beta}_0} = \frac{\hat{\sigma}^2 N^{-1} \sum_{i=1}^{N} x_i^2}{\sum_{i=1}^{N} (x_i - \bar{x})^2}
\]

\[
\hat{\sigma}^2 = \frac{1}{N - 2} \sum_{i=1}^{N} \hat{u}_i^2 = \frac{SSR}{N - 2}
\]

\( \hat{\sigma} \) is the standard error of the regression

---

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

Summary of Functional Forms Involving Logarithms

\[ y = \beta_0 + \beta_1 x + u \]

<table>
<thead>
<tr>
<th>Model</th>
<th>Dependent Variable</th>
<th>Independent Variable</th>
<th>Interpretation of ( \beta_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>level-level</td>
<td>( y )</td>
<td>( x )</td>
<td>( \beta_1 = \frac{\Delta y}{\Delta x} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \Delta y = \beta_1 \Delta x )</td>
</tr>
<tr>
<td>level-log</td>
<td>( y )</td>
<td>( \log(x) )</td>
<td>( \beta_1 = \frac{\Delta y}{100 \Delta x} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( \Delta y = (\beta_1/100)% \Delta x )</td>
</tr>
<tr>
<td>log-level</td>
<td>( \log(y) )</td>
<td>( x )</td>
<td>( 100 \beta_1 = \frac{% \Delta y}{% \Delta x} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( % \Delta y = (100 \beta_1)% \Delta x )</td>
</tr>
<tr>
<td>log-log</td>
<td>( \log(y) )</td>
<td>( \log(x) )</td>
<td>( \beta_1 = \frac{% \Delta y}{% \Delta x} )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>( % \Delta y = \beta_1 % \Delta x )</td>
</tr>
</tbody>
</table>

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.

www.feedthefuture.gov