MICHIGAN STATE UNIVERSITY

AFRE 802
Statistical Methods for Agricultural, Food, & Resource Economists

Linear models & estimation by least squares – Part 1 of 3
(WMS Ch. 11.1-11.3, Wooldridge pp. 22-37)

November 28, 2017

Nicole Mason
Michigan State University
Fall 2017

MICHIGAN STATE UNIVERSITY

GAME PLAN

- Collect Ch. 10 HW
- Hand out graded exercise (due Thursday)

- Linear models & estimation by least squares – Part 1 of 3
  - The simple linear regression model
    - Examples
    - Terminology
    - Assumptions about the error term (and concept of endogeneity)
    - Deriving estimates of the simple linear regression parameters: ordinary least squares (OLS)
    - Compute OLS estimates by hand & in Stata

1

MICHIGAN STATE UNIVERSITY

Simple Linear Regression

2

MICHIGAN STATE UNIVERSITY

The simple linear regression model: motivation

- Suppose *y* and *x* are two variables that represent some population
- **How does *y* change when *x* changes? What is the causal effect (*ceteris paribus* effect) of *x* on *y*?**
- *Examples?*

y	*x*
Corn yield	Fertilizer
Beef demand	Beef price
Wheat acreage	Wheat price
Hourly wage	Years of education
Community crime rate	# of police officers

3

The 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
	Covariate

4

To use data to get unbiased estimates of β_0 and β_1, need to restrict 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 → $E(u|x) = E(u) = 0$ (zero conditional mean)

- If this holds, **x is "exogenous"**; but **if x is correlated with u, x is "endogenous"** (next class)

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*)

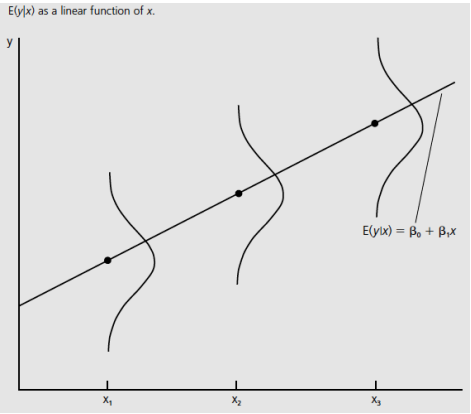
5

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

Hint: Apply the rules for conditional expectations.

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

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



Source: Wooldridge (2003)

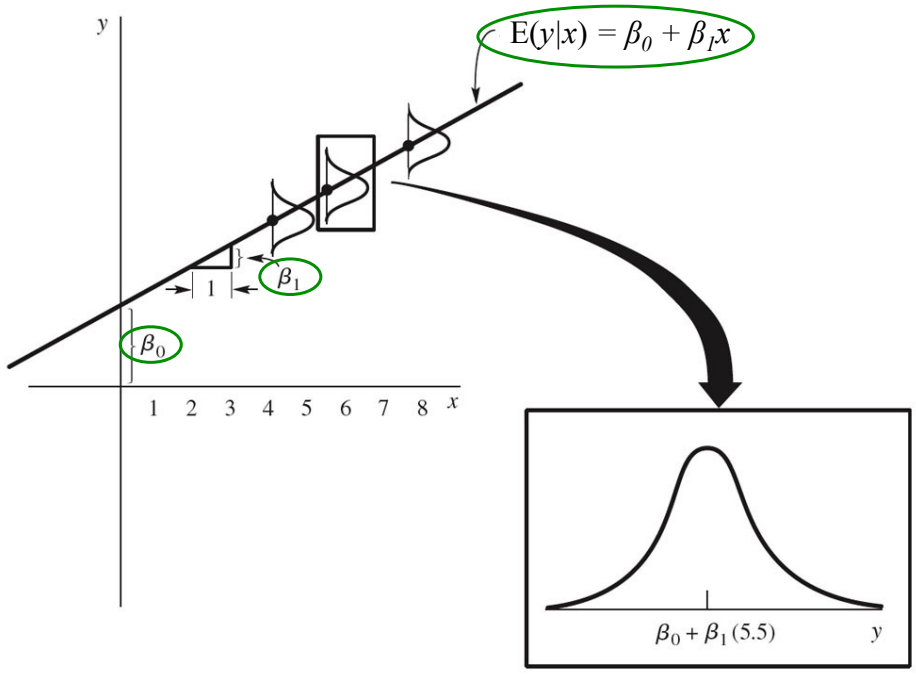
What is $\frac{\partial E(y|x)}{\partial x}$ *and how do we interpret this result?*

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

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)



MICHIGAN STATE UNIVERSITY

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 nonlinear in parameters, e.g.:

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

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

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

8

Estimating β_0 *and* β_1

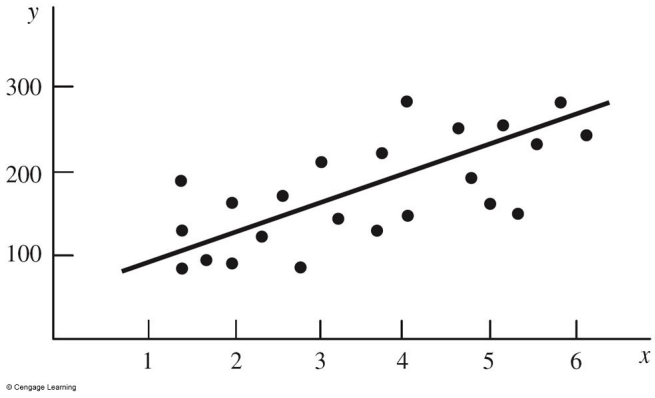
$$y = \beta_0 + \beta_1 x + u$$

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

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

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

How could we use the data in our sample to estimate β_0 and β_1?



MICHIGAN STATE UNIVERSITY

Recall 3 common methods of estimation

1. Method of moments
2. Maximum likelihood
3. Least squares

- All 3 of these approaches lead to the same estimators for β_0 and β_1 (under certain assumptions)
- We'll focus on the least squares approach
- See Wooldridge (2003: 27-29) for method of moments discussion and his panel data book for MLE discussion

10

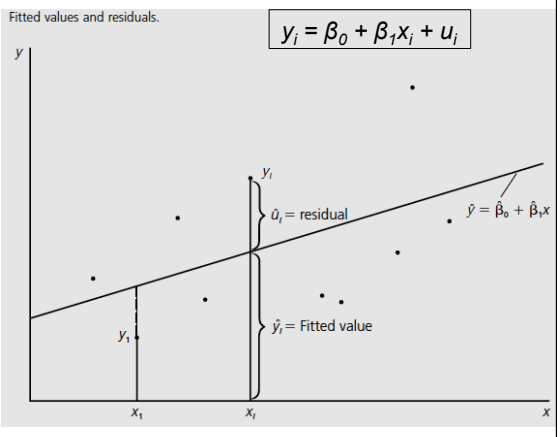
(Ordinary) least squares (OLS) approach

- *What was the gist of this approach?*
  - Choose estimator to minimize the sum of squared deviations b/w observed & estimated values
- "Fitted" values of y and residuals:

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

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

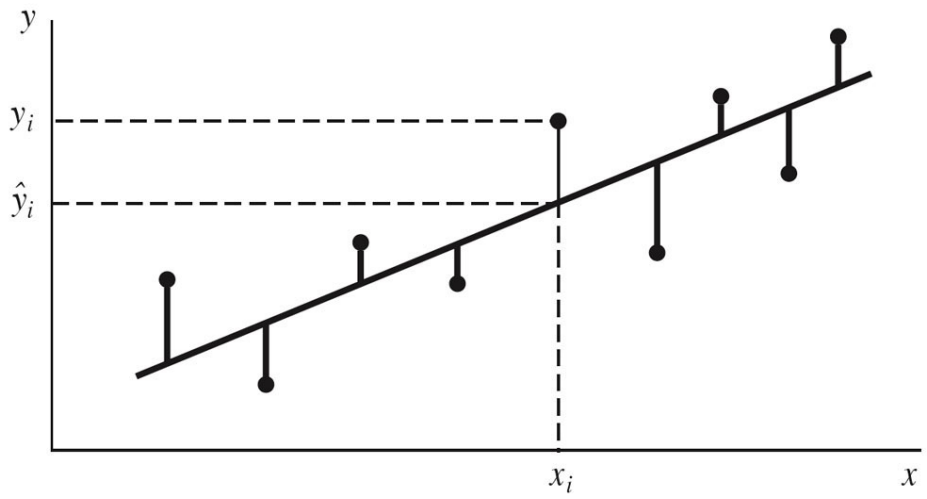
OLS:
Choose $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize:
$\sum_{i=1}^{N} \hat{u}_i^2 = \sum_{i=1}^{N} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$



Source: Wooldridge (2003)

MICHIGAN STATE UNIVERSITY

Where are the fitted values and residuals in this figure?



MICHIGAN STATE UNIVERSITY

The OLS estimators for β_0 and β_1

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

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

Work through proof of formulas above. **This is a proof you should know.**

Another useful expression for $\hat{\beta}_1$:

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

Proof will be posted to D2L.

MICHIGAN STATE UNIVERSITY

Obtaining OLS estimates – example (by hand)

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

EXAMPLE **11.1**

Use the method of least squares to fit a straight line to the $n = 5$ data points given in Table 11.1.

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

$\hat{\beta}_0 = 1 - 0.7(0) = 1$

$\hat{\beta}_1 = \frac{\sum x_i y_i - \frac{1}{N}\sum x_i \sum y_i}{\sum x_i^2 - \frac{1}{N}(\sum x_i)^2} = \frac{7 - \frac{1}{5}(0)(5)}{10 - \frac{1}{5}(0)} = \frac{7}{10} = 0.7$

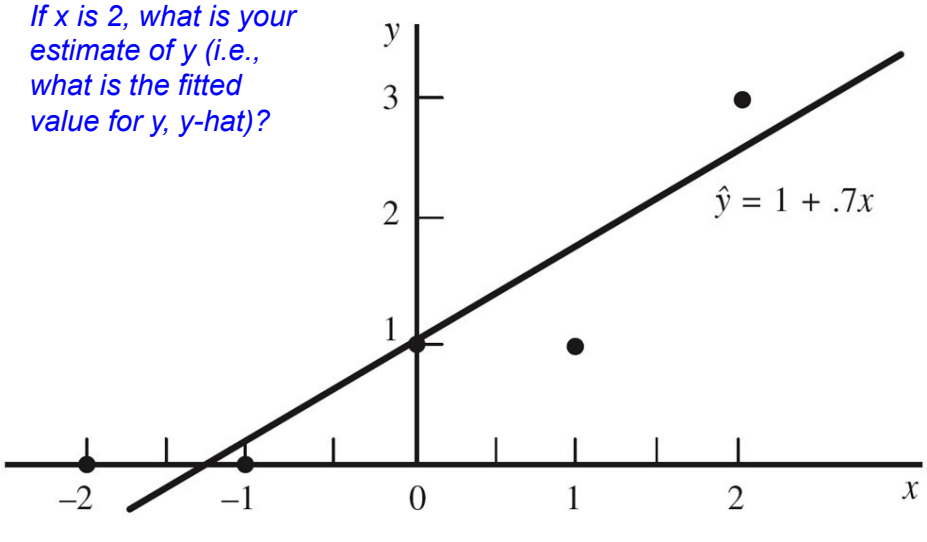
Table 11.1 Data for Example 11.1

x	y	$x_i y_i$	x_i^2
−2	0	0	4
−1	0	0	1
0	1	0	0
1	1	1	1
2	3	6	4
		$\sum x_i y_i = 7$	$\sum x_i^2 = 10$

$\sum x_i = 0$ $\sum y_i = 5$

MICHIGAN STATE UNIVERSITY

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



MICHIGAN STATE UNIVERSITY

Calculations using alternative formula

Table 11.2 Calculations for finding the coefficients

x_i	y_i	$x_i y_i$	x_i^2
−2	0	0	4
−1	0	0	1
0	1	0	0
1	1	1	1
2	3	6	4
$\sum_{i=1}^{n} x_i = 0$	$\sum_{i=1}^{n} y_i = 5$	$\sum_{i=1}^{n} x_i y_i = 7$	$\sum_{i=1}^{n} x_i^2 = 10$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{N}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{N}x_i y_i - \frac{1}{N}\sum_{i=1}^{N}x_i\sum_{i=1}^{N}y_i}{\sum_{i=1}^{N}x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N}x_i\right)^2} = \frac{7-\frac{1}{5}(0)(5)}{10-\frac{1}{5}(0)^2} = 0.7$$

16

MICHIGAN STATE UNIVERSITY

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)

17

Obtaining OLS estimates – example (Stata)

Wooldridge (2003) Example 2.4: Wage and education
Use Stata to run the simple linear regression of wage (y) on educ (x).

$$wage_i = \beta_0 + \beta_1 educ_i + u_i$$

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

What are β_0 and β_1 below?

```
reg wage educ
```

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

wage	$\hat{\beta}_1$ 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

$\hat{\beta}_0$

Homework:

- WMS Ch. 11:
  - 11.1, 11.3 (do calculations by hand)
  - 11.4 and 11.5 (for these two problems, do calculations using Excel and the formulas on slide 17, and in Stata using the "regress" command; the data are on D2L in the Stata folder)
- Try to complete all Ch. 11 HW before last day of class (Dec. 8) so that we can go over it then. (You won't turn in Ch. 11.)

Next class:

- Linear regression part 2 of 3
  - Properties of OLS estimators

Reading for next class:

- WMS Ch. 11: section 11.4
- Wooldridge *Introductory Econometrics* (2003): pp. 38-60, 101-102

19

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)

20

MICHIGAN STATE UNIVERSITY

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, 2003: 417)

21