More Details on the Simple Linear Regression Model

Tyler Ransom

Univ of Oklahoma

Jan 24, 2019

Today's plan

1. Review reading topics
   1.1 Units of Measurement
   1.2 Functional Form
   1.3 Conditions for Unbiasedness/Computation of standard errors
2. In-class activity: More practice running regressions and interpreting estimates

Units of measurement

Background

- The three challenges of statistical inference are:[1]

  1. Generalizing from sample to population (statistical inference)
  2. Generalizing from control to treatment group (causal inference)
  3. Generalizing from observed measurements to underlying constructs of interest (measurement)

[1] Taken from Andrew Gelman's **blog**.

Units of measurement

- very important to know how y and x are measured in order to interpret regression functions

- example: CEO salary and the company's return on equity (*roe*).

$$\widehat{salary} = 963.191 + 18.501\, roe$$
$$N = 209,\ R^2 = .0132$$

- If *salary* is in thousands and *roe* is in percent, what is interpretation of $\hat{\beta}_1 = 18.501$?

- What is the interpretation of $\hat{\beta}_0 = 963.191$?

Changing the units of measurement

- What if now we decide to measure *roe* as a decimal instead of a percent?

Changing the units of measurement

- What if now we decide to measure *roe* as a decimal instead of a percent?

$$\widehat{salary} = 963.191 + 1,850.1\ roedec$$
$$N = 209,\ \ R^2 = .0132$$

where $roedec = \frac{roe}{100}$

Changing the units of measurement

- What if now we decide to measure *roe* as a decimal instead of a percent?

$$\widehat{salary} = 963.191 + 1,850.1\ roedec$$
$$N = 209,\ R^2 = .0132$$

where $roedec = \frac{roe}{100}$

- And what if *salary* is in dollars instead of thousands of dollars?

Changing the units of measurement

- What if now we decide to measure *roe* as a decimal instead of a percent?

$$\widehat{salary} = 963.191 + 1,850.1\ roedec$$
$$N = 209,\ R^2 = .0132$$

where $roedec = \frac{roe}{100}$

- And what if *salary* is in dollars instead of thousands of dollars?

$$\widehat{salary} = 963,191 + 18,501\ roe$$
$$N = 209,\ R^2 = .0132$$

Units, interpretation, and model performance

- Notice how the R^2 didn't change at all when we changed the units!
- **Changing the units only changes the interpretation, not the performance of the model**
- Typically should choose units that correspond to plausible changes
- e.g. typical $\Delta roe = 1\%$, not 100%

Functional Form

Functional Form

- Sometimes a linear function isn't very realistic
- e.g. a simple wage-education equation

$$\widehat{wage} = -5.12 + 1.43\, educ$$
$$N = 759, \; R^2 = .133$$

  where *wage* is the hourly wage earned, and *educ* is years of education
- What's weird about this?

Functional Form

- Sometimes a linear function isn't very realistic
- e.g. a simple wage-education equation

$$\widehat{wage} = -5.12 + 1.43\, educ$$
$$N = 759,\ R^2 = .133$$

  where *wage* is the hourly wage earned, and *educ* is years of education

- What's weird about this?

  1. $educ = 0 \overset{?}{\Rightarrow} wage = -5.12$

Functional Form

- Sometimes a linear function isn't very realistic
- e.g. a simple wage-education equation

$$\widehat{wage} = -5.12 + 1.43\, educ$$
$$N = 759, \;\; R^2 = .133$$

  where *wage* is the hourly wage earned, and *educ* is years of education

- What's weird about this?

  1. $educ = 0 \overset{?}{\Rightarrow} wage = -5.12$
  2. Constant return to education. Should be increasing!

The *log* transformation

- Instead, consider using $log(wage)$:

$$\widehat{log(wage)} = 1.142 + 0.099\ educ$$
$$N = 759,\ R^2 = .165$$

where $log(\cdot)$ is the natural logarithm

✓ Now we don't have negative wage when $educ = 0$

✓ Model allows for increasing returns to *educ* (but constant *percentage* effect)

  - Interpretation:

The *log* transformation

- Instead, consider using $log(wage)$:

$$\widehat{log(wage)} = 1.142 + 0.099\ educ$$
$$N = 759,\ \ R^2 = .165$$

where $log(\cdot)$ is the natural logarithm

✓ Now we don't have negative wage when $educ = 0$

✓ Model allows for increasing returns to *educ* (but constant *percentage* effect)

  - Interpretation: one-unit $\uparrow$ *educ* corresponds to 9.9% $\uparrow$ *wage*

Other uses of log

- Can also put the log on the x variable (or both), See Table 2.3:

Model	Dep. Var.	Indep. Var.	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\,\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\,\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

- Note: putting in a log changes the R^2 completely

- Use log to allow y and x to vary nonlinearly, but still be **linear in parameters**

Unbiasedness, standard errors

Gauss-Markov Assumptions

1. Linear in parameters
2. Random sampling
3. $Var(x) > 0$
4. $E(u|x) = 0$
5. $Var(u|x) = \sigma^2$ (**homoskedasticity**)

With (1)-(4) satisfied: OLS estimates are **unbiased** and

With (5) satisfied: can **easily compute standard errors**

Are these crazy assumptions?

On a scale of “not at all” to “absolutely”:

Linear in parameters Not too crazy

Random sampling Not crazy if cross-sectional data

$Var(x) > 0$ Not at all crazy

$E(u|x) = 0$ Absolutely crazy if observational data!

$Var(u|x) = \sigma^2$ Can be crazy, especially if time series / panel data

Why do we need to make these assumptions?

You might wonder why we bother to make these assumptions

- We do econometrics to learn something about a population of interest

- We can't learn much if we don't make any assumptions!

- Bothered by these assumptions?

- Think: "tell how to conduct **statistical inference** on **experimental data**"

Variance of OLS estimators

- Last time, we introduced the formulas for OLS estimators

- Also interested in their **variance**

- So we know how far away $\hat{\beta}$ is expected to be from β

- A big component of these estimators is $\sigma^2 = Var(u)$

$$\begin{aligned}\hat{\sigma}^2 &= \frac{SSR}{N-2} \\ &= \frac{1}{N-2}\sum_{i=1}^{N}\hat{u}_i^2\end{aligned}$$

Variance of OLS estimators

- Once we have $\hat{\sigma}^2$, we can obtain the SE of the β's

$$Var\left(\hat{\beta}_0\right) = \frac{\sigma^2 \sum_{i=1}^{N} x_i^2}{N \sum_{i=1} \left(x_i - \overline{x}\right)^2}$$

$$Var\left(\hat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1} \left(x_i - \overline{x}\right)^2} = \frac{\sigma^2}{SST_x}$$

- Don't worry about memorizing these formulas
- Key takeaway: we can write them down in a fairly compact form
- We can do that because of the assumptions we made