holding other
↗ factor unchanged!

3.2 How could the multiple regression analysis enable ceteris paribus analysis?

- Consider a multiple regression function of *wage*

to know impact
only 1 variable

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + u \tag{4.4}$$

Here,

- β_0 is the intercept.

  β_1 measures the change in *wage* with respect to *educ*, holding other factors fixed.

  β_2 measures the change in *wage* with respect to *inc*, holding other factors fixed.

- What if the function of *wage* is, instead written as

$$wage = \beta_0 + \beta_1 educ + \beta_2 inc + \beta_3 educ^2 + u$$

Then,

β_0 is the intercept.

The change in *wage* with respect to *educ* (holding other factors fixed) is measured by:

$$\frac{\partial\, wage}{\partial\, education} = \beta_1 + 2\beta_3\, educ$$

The change in *wage* with respect to *inc* (holding other factors fixed) is measured by:

$$\frac{\partial\, wage}{\partial\, income} = \beta_2$$

TBC :-)

A story to Ceteris Paribus analysis 20/02/2020

- A student in the EE489 class wants to find the impact of parking space on Condo's price.

  So, he proposes price = $\beta_0 + \beta_1$ % parking + u

  ⇒ What do you think should be the sign of β_1? (+) or (-) Why?

  (+) because better quality condominiums usually allocate more space for parking

- But then, the student said, this may not be the case because condos that are closer to BTS & MRT usually have lower % parking / room but their prices are higher!

  ⇒ This student said he expects the sign of β_1 to be (−) !!

  ⇒ Ajarn said, definitely. If you estimate a simple regression (1) having % parking as the only x variable, your analysis will not be able to analyze the impact of % parking, holding other factors constant

  distance in MRT, BTS

- so, to make β_1 measure only the impact of % parking, you should include distance to the nearest BTS, MRT in the regression.

We can expect β_1 to be (+) in this equation

$$\text{Price}_i = \beta_0 + \beta_1 \% \text{ parking} + \beta_2 \text{ dist\_BTS} + \beta_3 \text{ dis\_MRT}$$

By including BTS, MRT into the equation, we can conduct partial effects analysis of % parking, holding BTS, MRT constant.

4 Expected Value of the OLS Estimators in multiple regression

- Under assumptions MLR 1 to 4 (see Wooldridge), $\hat{\beta}_{OLS}$ are unbiased.

- 2 issues should be considered regarding the biasedness of $\hat{\beta}_{OLS}$

ex) distance from condo to Japan to explain condo price

4.1 *Issue #1: Including Irrelevant Variable (Overspecifying the Model)* Too many variable than necessary

- Suppose we specify the model

assume that X_2 is irrelevant

ถ้า include irrelevant จะทำให้ equation biased?

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \tag{4.5}$$

and this model satisfies the multiple regression assumptions 1 to 4

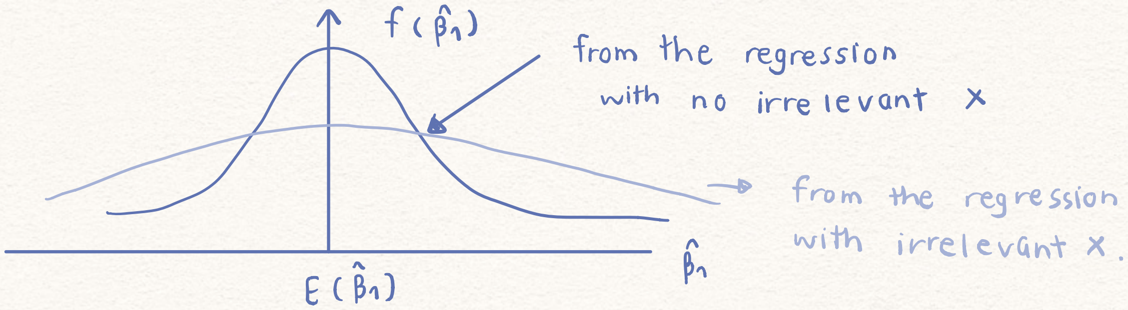
- If the true value of β_2 is "0" $\Rightarrow$ no need to put X_2 in the multiple regression (Because X_2 does not explain Y)

- If we estimate a model with X_2 anyway,

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

$\Rightarrow$ The estimated OLS parameters $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ will still be unbiased because we do not violate assumption MLR 1-4

- However, including irrelevant variable (X_2 in this case) would make the variance of $\hat{\beta}_0$, $\hat{\beta}_1$ become unnecessarily LARGE. So, $\hat{\beta}_0$, $\hat{\beta}_1$ not be the most efficient.



ถ้าไม่ใส่ β_2, β_3 จะทำให้ β_1 biased

4.2 Issue #2: Excluding Relevant Variable (Underspecifying the Model → omitted variable bias. This is a serious problem!)

- Suppose we the **TRUE** model is actually

so, we omitted the factor that explains y

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u \rightarrow \text{TRUE MODEL}$$

where none of the β is zero **and this model satisfies the multiple regression assumptions 1 to 4.**

- But we omit variable X_2 and estimate the following equation using OLS

$$Y = \beta_0 + \beta_1 X_1 + v \rightarrow (v = \beta_2 X_2 + u)$$

and we estimate the wrong model :

$$\tilde{y} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1$$

let's check if $\tilde{\beta}_1$ will be biased by checking if we have violated assumption MLR 1 – 4

$\Rightarrow E(v) = E(\beta_2 X_2 + u) = \beta_2 E(X_2) + E(u)$

$= \beta_2 E(X_2) \neq 0 \Rightarrow E(v)$ in the wrong model $\neq 0$

MRL 1 - linear parameter

MRL-2 random sampling

MRL-3 multi

MRL-4 zero conditional

$\Rightarrow$ Since X_2 is in the error term of the wrong model (v), $Cov(X_1, X_2)$ has to be "0" for $\tilde{\beta}_1$ to be unbiased (or for $Cov(X_1, v) = 0$)

☺ some proof omitted see textbook ☺

$E[\tilde{\beta}_1] = \beta_1$ (unbiased)

result from doing OLS of the wrong model we get $E[\tilde{\beta}_1] = \beta_1 + \beta_2 E\left(\frac{\sum_i x_{1i} x_{2i}}{\sum x_{1i}^2}\right)$; $x_{1i} = X_{1i} - \bar{X}$, $x_{2i} = X_{2i} - \bar{X}$

$$= \beta_1 + \beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)} \rightarrow \text{Biased}$$

so, $E(\tilde{\beta}_1) \neq \beta_1 \Rightarrow$ it will be biased by the size of $\beta_2 \frac{Cov(X_1, X_2)}{Var(X_1)}$

Direction of Bias

If $\beta_2 > 0$, then $Cov(X_1, X_2) > 0 \Rightarrow \tilde{\beta}_1$ will be upward biased

$Cov(X_1, X_2) < 0 \Rightarrow \tilde{\beta}_1$ will be downward biased

$\beta_2 < 0$, then if $Cov(X, X_2) > 0 \Rightarrow \hat{\beta}_1$ will be downward biased

$Cov(X, X_2) < 0 \Rightarrow$ upward biased

Example

Suppose the real population regression is

$$price = \beta_0 + \beta_1 \%\ parking + \beta_2\ dist\_BTS\_MRT + u$$

% parking spot to # rooms

distance to the nearest BTS or MRT station

But we estimate

$$price = \beta_0 + \beta_1 \%\ parking_i + v_i$$

what we omitted

$$\text{Then}, \quad E(\tilde{\beta}_1) = \underset{(+)}{\beta_1} + \underbrace{\underset{(-)}{\beta_2} \frac{\overset{(+)}{Cov}(\%\ parking, distance\_BTS\_MRT)}{\underset{(+)}{var(\%\ parking)}}}_{(-)}$$

ยิ่งใกล้ยิ่งแพง

we have a downward biased here!

★ $\tilde{\beta}_1$ is going to be smaller than expected if we omitted the dis_MRT_BTS variable!

— more ex in Gujarati —

5 Variance of the OLS Estimators

- The $\hat{\beta}_{OLS}$ would be the most efficient among the linear unbiased estimators if assumption 5 is satistifed
- Multiple Linear Regression (MLR) assumption 5: Homoskedasticity

  The error term u has the same variance given any values of the explanatory variables.

$$Var(u|X_1, X_2, ..., X_k) = \sigma^2$$

← constant regardless of the value of x

- Example:

$salary = \beta_0 + \beta_1 profits + \beta_2 sales + u$

homoskedasticity require that $var(u / profit, sales) = \sigma^2 \Rightarrow$ MLR 5

- If the MLR assumption 5 is true, then → with some prove (in the textbook)

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{\sum_i (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)} = \frac{\sigma^2}{SST_j (1 - R_j^2)}$$

where $j \in 1, 2, \ldots k^{th}$ explanatory variable)

- R_j^2 is the R^2 from regression x_j on all other X_s (except x_j itself)

i.e. R_j^2 is the R^2 from

$$x_j = \alpha_0 + \underbrace{\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k}_{\text{not including } x_j}$$

or $x_2 = \alpha_0 + \alpha_1 x_1 + \alpha_3 x_3 + \ldots + \alpha_k x_k$

6 Estimator of the OLS Variance

- Since we don't know what σ^2 is (population concept), we need to find an estimator of it.

$$\hat{\sigma}^2 = \frac{\Sigma_i \hat{u}_i^2}{n - \text{number of estimated parameter}}$$

$$= \frac{\Sigma_i \hat{u}_i^2}{n - (k+1)}$$

+1 intercept k slopes

$$\underbrace{\beta_0 + \beta_1 + \ldots + \beta_k}$$

Simple regress

n-2

so, variance of $(\hat{\beta}_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$

If simple regression, $var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SST_x}$

Practice question on moodle in Cengage

- Thus, STATA's calculation of the std.err. of $\hat{\beta}_j$ is

$$\widehat{std.deviation}.\hat{\beta}_j = std.err.\hat{\beta}_j = \sqrt{\frac{\hat{\sigma}^2}{\Sigma_{i=1}^N (X_{ij} - \bar{X}_j)^2 (1 - R_j^2)}}.$$

Comments: * Aj. will come back to this after the midterm exam

Stata Lab 1 – Introduction

1 What is STATA?

- A statistical software package used mostly in economics, sociology, political science and epidemiology.
- Stata can be used to manage database, run regressions, generate graphics, do simulations, etc.
- The user should have their own dataset. The Stata data file is usually saved in the .dta format.
- Data of any other formats (like excel) can be imported and/or converted into .dta format.

1.1 STATA supports

- Stata's own website: http://www.stata.com/support/faqs/
- Stata program's help function: For example, suppose you would like to know more about the "regress" command, then... Open the stata program > in the "command" box > type "help regress" without the " "> press enter.
- Stata's official manual (can be found in the library and embeded in the program)
- Other Stata's user's manual: My favorite one is "An Introduction to Modern Econometrics Using Stata" by Christopher F. Baum.
- or... simply type your question(s) into a search engine.

1.2 Data files and Do-files

- The Stata's data file keeps all the data points. For example, each Y_i and the corresponding X_i, $\forall\ i = 1, 2, 3..., n$.
- The do-file records all the commands that you use to analyze the data.
- Once the data is cleaned, it is best not to keep on re-saving the original data file. If several steps have to be done before analyzing the data (like running a regression), do it on the do-file.

2 Tutorial 1: Exploring the Data and Running a Simple Regression

- **Download Wooldridge datasets**

1. Download Wooldridge's data – go to thomsonedu's website (or go to your BE Moodle: EE325):

   "http://www.thomsonedu.com/aise/economics/wooldridge_2e_datasets/".

- **To open the STATA program**

1. Double click on the STATA icon.
2. Click on the "Do-file" icon on the top panel of the Stata program. "Save As" your Do-file (and name it "EE325") on your computer.

- **Open the data file using Do-file**

1. file -> open -> then, direct the program to the file "CEOSAL2.DTA".
2. On the command window, you will see a command to open this file. If you would like to open the file from your do-file in the future, you can use this command.

- **To explore and understand the data**

1. type: browse
2. type: describe
3. type: summarize
4. type: sum
5. type: codebook
6. type: describe salary
7. type: tabulate college
8. type: tab college
9. to use the "if" command to find conditional mean (average) type: sum if grad == 1
10. type: sum salary if age <= 40
11. type: correlate salary sales profits
12. type: correlate salary sales profits, covariance
13. type: plot salary profits
14. type: twoway scatter salary profits

- **To run a simple (OLS) regression (one explanatory variable)**

1. type: regress salary profits

- **To create the fitted value ($\hat{Y}_i$) and the residual ($\hat{u}_i$)**

1. type: predict y_hat, xb
2. type: predict u_hat, residual

- **To see how well we do at finding a Sample Linear Function**

1. type: twoway scatter salary profits || line y_hat profits
2. type: twoway scatter u_hat profits
3. To check if the OLS estimation makes X_i uncorrelated with $\hat{u}_i$ (by the OLS calculation, they should not correlate), type: correlate sales u_hat

- **To execute mathematical operations**

1. type: generate log_salary = log(salary)
2. type: gen log_profit = log(profit)
3. type: gen profit_2 = 3+5*profit
4. type: regress salary profits profit_2
5. type: gen profit_sq = profit^2
6. type: regress salary profits profit_sq

- **To perform a multiple regression analysis**

1. type: regress salary profits sales
2. type: regress salary profits sales ceoten

- **To exit your Stata**

1. Save your do-file
2. file -> exit -> don't save (never ever modify your master dataset!)

- **To find out what all the above commands mean** – (type in the command box) help summarize, help predict, help twoway, etc etc.