Seminar Corporate Governance: Basic Econometrics for Data Analysis

Yuhao Zhu
y.zhu@ese.eur.nl

10 January 2018

Contents

Part I
Introductory

1 Why we are here and how we get there?

Slide 2 **Who am I?**

- Yuhao (Hanan) Zhu.
- Instructor for EBC2.
- Email: y.zhu@ese.eur.nl.
- Questions with title line "[Seminar ACF]".

Slide 3 **Why this topic?**

- We assume that you have sufficient background in Econometrics. However,
- Students are of different academic backgrounds.
- Some lack sufficient knowledge on Econometrics.
- Econometrics is essential skill for this course!

Slide 4 **What to expect?**

- You have better understanding of basic Econometric theories.
- You know how to apply basic Econometrics models.
- You know how to interpret the results.
- You can probably solve the free-rider problem.

2 What to learn today?

Slide 5 **What to learn today?**

- Data Generating Process (DGP).
- Ordinary Least Square (OLS) estimator.
- Hypothesis testing.
- Interpretation.
- Dummy variables.
- Interactive variables.
- Instrumental variable (IV) and Two-stage Least Square (2SLS).
- Panel data set and fixed effects (FE) model.

3 Learn from animation

Slide 6 **Animation**

- Sometimes we lose intuition to econometrics.
- What do these equations actually mean?
- Animation gives you better intuition.

Slide 7 **Common mistakes**

- We do not dive deep into advanced knowleges.
- We focus on common mistakes.
- Be really careful!

Slide 8 **Note**

- There are two kinds of questions: questions and good questions. So feel free to ask.
- I will also ask questions during the lecture.

Part II
Basic Knowledge

4 Data Generating Process (DGP)

Slide 9 **Data Generating Process**

- Data Generating Process is the (imagined) process through which the data is generated.
- A certain x_i.
- An random error term ϵ_i drew from a certain distribution.
- $y_i = \underbrace{f(x_i)}_{deterministic} + \underbrace{\epsilon_i}_{random}$.

Slide 10 **Example: Data Generating Process**

- A typical univariate example:
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- A typical multivariate example:
- $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$.

Slide 11 **Animation: Data Generating Process**

- Deterministic DGP:
- `econometrics.dgp()`
- Interesting for research?
- Random DGP:
- `econometrics.dgp(random=True)`
- Interesting for research?

Slide 12 **Our task**

- Ideal outcome: To find the DGP using observed data set.
- Difficulty: The disturbance of the error terms.
- New task: To estimate the parameters of DGP with highest precision.
- How: Use the data set on x and y, as well as some assumptions on ϵ.
- Opening the black box. Reverse-engineering.

5 Ordinary Least Square (OLS) Estimator

Slide 13 **Estimate and estimator**

- If the true DGP is:
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- We want to estimate (β_0, β_1).
- Estimate: b. Our predicted $\hat{\beta}$, contrary to the true β (Greek letter).
- Estimator: A rule of calculating the estimate given observed data set (a function).

Slide 14 **Choice of estimator**

- Virtually infinite rules of calculating the estimate.
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- For example, the sample mean of the independent variable: $\hat{\beta}_0 = \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i}{n}$.
- Of course it is a bad estimator!
- Which one is the best?

Slide 15 **Criteria**

- Criteria of a good estimator.
- Unbiasedness: If we draw sample T times, and we have calculated T estimates. The average of the estimates will converge to the true value if T goes to infinite.
- Consistency: If we draw a sample of N observation, and we calculate the estimate. The estimate will converge to the true value if N goes to infinite.
- Efficiency: Among all unbiased estimator, we want the one with the lowest variance.
- Questions: Empirically speaking, which criteria is more important?

Slide 16 **Animation: Criteria**

- DGP 1000 times.
- `econometrics.dgp(times=1000, random=True)`
- We love big data set.

Slide 17 **The calculation**

- Given linear DGP (and several other assumptions).
- The Ordinary Least Square (OLS) estimator is unbiased, consistent and efficient.
- We normally assume linear DGP.
- OLS is often used.

Slide 18 **Ordinary Least Square**

- Example: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.
- Literal meaning of "Ordinary Least Square": minimizing the sum of squared residuals.
- $\min_{\hat{\beta}} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$
- The OLS estimator:
- $\hat{\beta}_0 = \frac{(\sum x^2)(\sum y) - (\sum x)(\sum xy)}{n \sum x^2 - (\sum y)^2}$.
- $\hat{\beta}_1 = \frac{n \sum xy - (\sum x)(\sum xy)}{n \sum x^2 - (\sum y)^2}$.

Slide 19 **Consistency of OLS**

- The Ordinary Least Square (OLS) estimator is consistent.
- See animation.
- `econometrics.dgp(times=1000, random=True, fit=True)`

Slide 20 **Standard error of estimates**

- A estimate is a random variable.
- Why?
- $\hat{\beta}$ is the function of x and y.
- And y is the function of x and a random variable ϵ.
- So $\hat{\beta}$ is the function of the random variable ϵ.
- ϵ has mean and variance.
- So do $\hat{\beta}$!

Slide 21 **Estimates**

- Estimate $\hat{\beta}$ is also a random variable.
- $\hat{\beta}$ has mean and standard error.
- The smaller the standard error is, the more accurate the estimate is.

6 Hypothesis Testing

Slide 22 **Hypothesis**

- Sometimes, we not only consider the magnitude of the effect, we also care about whether the effect really exists.
- Recall that the estimate $\hat{\beta}$ is also a random variable.
- We might wrongly estimate.
- So we need hypothesis testing.
- For example: The true effect is actually zero!

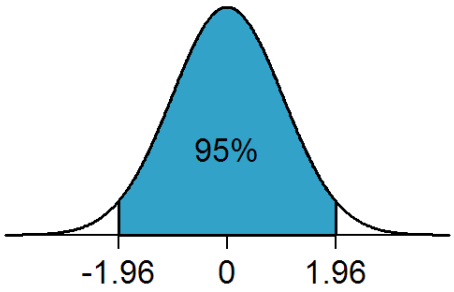
Slide 23 **Type-1 error**

- We can allow certain level of error.
- We define the type-1 error: the hypothesis is correct but we reject it.
- For example, we hypothesize that $\beta = 0$.
- The estimated $\hat{\beta}$ is asymptotically normally distributed $\mathcal{N}(0, 1)$.
- Then $\hat{\beta} \geq 3$ rarely happens (at a probability smaller than 1%).
- If our estimate is 3, then we say that a rare case happen!
- We reject the null hypothesis at the 1% level.
- The probability that "$\beta = 0$ is right but we reject it" is smaller than 1%.

Slide 24 **t-statistics and p value**

- b/s follows t-distribution.
- $t = \frac{b}{S.E.}$
- We can look at the t-statistics to judge whether our estimate is significantly different from zero, and at which level.
- p value is more obvious. It shows at which level the estimate is significantly different from zero.
- We normally require p to be at least smaller than 10%.

Slide 25 ***t-statistics* and *p-value***



- Reject the null hypothesis at the 95% level if t-statistics is located in the white region.
- To know about significant level, simply count stars.

Slide 26 **Estimates as output**

- $\hat{\beta}$ follows t-distribution.
- Statistical softwares report the mean and the standard error of $\hat{\beta}$.
- Also, the calculated t-statistics.
- We have p value of course.

Slide 27 **A typical table**

Dep. Var.	ln(workers' wage)						
	model 1	model 2	model 3	model 4	model 5	model 6	model 7
ln(CEO total)	-0.001	-0.007	-0.002	-0.002	-0.02	-0.023*	
	-0.27	-1.32	-0.2	-0.21	-1.5	-1.65	
ln(CEO total) (t-1)							-0.023
							-1.49
ROA		0.033	0.079*	0.079*	-0.092	0.026	0.032
		0.93	1.73	1.74	-0.97	0.27	0.28
Leverage ratio		0.066***	0.117***	0.117***	0.053	0.093	0.111
		2.64	3.32	3.29	0.44	0.76	0.78
Market-to-book ratio			-0.005**	-0.005**	-0.004	-0.001	-0.013**
			-2.44	-2.33	-1	-0.32	-2.18
ln(total sales)			-0.007	-0.007	-0.013	-0.007	-0.016
			-1.6	-1.61	-0.52	-0.26	-0.57
After 2006	No	No	Yes	Yes	Yes	Yes	Yes
Robust std error	No	No	No	Yes	Yes	Yes	Yes
Firm fixed effects	No	No	No	No	Yes	Yes	Yes
Year fixed effects	No	No	No	No	No	Yes	Yes
State fixed effects	No	No	No	No	No	Yes	Yes
Constant	9.897***	9.940***	9.888***	9.888***	10.067***	10.106***	10.340***
	132.36	129.94	95.01	97.69	45.74	43.85	22.38
Adj. R square	0	0	0.001	0.001	0.012	0.018	0.019
Obs.	16578	16439	13224	13224	13224	13224	9631

7 Interpretation

Slide 28 **Interpretation is essential**

- Interpretation is essential.
- Because we are economists.
- $y = \beta_0 + \beta_1 x + \epsilon$.
- Interpreted as slope and intercept.

Slide 29 **Two different types of interpretation: level-level**

- $y = \beta_0 + \beta_1 x + \epsilon$.
- Interpreted as linear (slope and intercept).
- β_1: When x increases by 1 unit, y increases by β_1.
- β_0: When x is zero, y is β_0.

Slide 30 **Two different types of interpretation: log-log**

- $\ln y = \beta_0 + \beta_1 \ln x + \epsilon$.
- Interpreted as pecentage change.
- β_1: When x increases by 1 percent, y increases by β_1 percent.
- β_0: When x is one, y is $\exp\{\beta_0\}$.

Slide 31 **Two different types of interpretation: log-level**

- What about a comibination of $\ln x$ and y?
- Common mistakes!
- $y = \beta_0 + \beta_1 \ln x + \epsilon$.
- β_1: When x increases by 1 percent, y increases by (...)?
- $\ln y = \beta_0 + \beta_1 x + \epsilon$.
- β_1: When x increases by 1 unit, y increases by (...) percent?

Slide 32 **Significance**

- Two different kinds of significance.
- Statistical significance and economic significance.
- Common mistakes!
- Put too much emphasize on statistical significance and too little emphasize on economic significance.

Slide 33 **R-square**

- What is R-square and Adjusted R-square?
- When do we need adjusted R-square and when do we not?
- Which one is larger, *ceteris paribus*?
- How do they change when you add in more variables, *ceteris paribus*?
- Common mistakes!
- If there are miskakes concerning those questions in your work, we may think that you are faking your results.

Part III
Intermediate Knowledge

8 Dummy Variables

Slide 34 **Dummy variables**

- Sometimes, variables are not in numerical terms.
- Yes/No, True/False, High/Low, Female/Male...
- We need dummy (indicator) variable.
- Dummy variable only takes 0 or 1.

- $y = \beta_0 + \beta_1 D + \epsilon$.
- How to interpret the coefficient for the dummy variable?

9 Interactive Variables

Slide 35 **Interactive effect**

- Variables may enhance the effect of each other.
- Example:
- $Wage = \beta_0 + \beta_1 Education + \beta_2 Tenure + \epsilon$.

Dep. Var.: Wage	Low Education	High Education
Short Tenure	2000	3000
Long Tenure	2500	4000

-
- Education and Tenure enhance the effect of each other on Wage. Calculate!

Slide 36 **Interactive effect: Calculation**

Dep. Var.: Wage	Low Education	High Education
Short Tenure	2000	3000
Long Tenure	2500	4000

-
- Effect of Education: 1000.
- Effect of Tenure: 500.
- Joint effect of Education and Tenure: 500.
- $Wage = \beta_0 + \beta_1 Education + \beta_2 Tenure + \beta_3 Education \times Tenure + \epsilon$.

Slide 37 **Interactive effect: Interpretation**

- $Wage = \beta_0 + \beta_1 Education + \beta_2 Tenure + \beta_3 Education \times Tenure + \epsilon$.
- Interpretation 1: When Tenure is high (one unit higher), the effect of Education on Wage is increased by β_3.
- Interpretation 2: The synergy effect (joint effect) of Education and Tenure on Wage is β_3.

10 Instrumental Variable (IV) And Two-stage Least Square (2SLS)

Slide 38 **Endogeneity problem**

- 100% of those who drink water die.
- Conclusion: Drinking water is fatal.
- Common mistakes!
- Neglect endogeneity problem and interpret correlation as causation.

Slide 39 **Smoking and health**

- How endogeneity problem occurs? Example:
- True DGP:
- $Health = \beta_0 + \beta_1 Smoking + \beta_2 Income + \epsilon$ with $\beta_1 < 0$ and $\beta_2 > 0$.
- Regression model (omitted variable problem):
- $Health = \alpha_0 + \alpha_1 Smoking + \eta$.
- If $cov(Smoking, Income) > 0$, $\hat{\alpha_1}$ can be larger than 0!
- The coefficient of Smoking actually involves the effect of income! It is incorrect estimate.

Slide 40 **Endogeneity problem**

- The error term ϵ is correlated with the independent variable x: hidden variables!
- Common mistakes!
- Be careful in interpretating your coefficients!

Slide 41 **A solution**

- Participation in a war increases one's further salary.
- $Salary = \beta_0 + \beta_1 War + \epsilon$.
- Maybe those who earn less participate the war? Missing variable: Income before war.
- The variable War is correlated with the error term. Endogeneity!
- Solution: Instrumental variable.

Slide 42 **Instrumental variable**

- IV is only correlated to the independent variable, but not the error term.
- For example: The government randomly call people to participate in the war.
- Call is positively corrected to War, but is not correlated to other variables.
- Call is an IV.

Slide 43 **Two stage least square**

- First stage: regress the endogenous variable on IV:
- $War = \alpha_0 + \alpha_1 Call + \eta$.
- Use the estimated $\hat{\alpha_1}$ to predict $\hat{War} = \hat{\alpha_0} + \hat{\alpha_1} Call$. $\hat{War}$ is not correlated to the error term!
- Second stage: Use $\hat{War}$ instead of War and conduct the OLS.
- $Salary = \beta_0 + \beta_1 \hat{War} + \epsilon$.

Slide 44 **Two stage least square: mistakes**

- Common mistakes!
- Do not run 2SLS on your own!
- Do you packages in statistical softwares.
- Why: Different formulas in calculating standard errors.

11 Panel Data Set And Fixed Effects (FE) Model

Slide 45 **Panel data set**

- Observations can be indexed by individuals and time.
- Example: yearly GPA of every student in a school in the past 3 years.
- Why we define panel data structure?
- There are some structural phenomena.
- Assumptions for pooled OLS fail!

Slide 46 **Structural phenomena**

- Time invariant factors within individuals.
- Common trend across time.
- Correlation of the observations within individuals.

Slide 47 **Solution: Fixed Effects model**

- Time invariant factors within individuals.
- Assign each individual an unique dummy variable.
- The dummy variable equals one only when the observation belongs to the corresponding individual.
- If there are N groups (individuals), we need (...) effective dummies.
- How to interpret the coefficients for the dummies?

Slide 48 **Time fixed effects**

- Common trend across time.
- Assign each year an unique dummy variable.
- We need $T-1$ effective dummies.

Slide 49 **Multiple Fixed Effects**

- Multiple fixed effects are possible.
- For example: a branch can belong to different firms, industries, and regions.
- Firm FE, industry FE, Region FE.
- Maybe time FE as well.

Slide 50 **Collinearity problem**

- Common mistakes.
- Collinearity problems when using fixed effects model.
- For example: We want to see the firm performance during recession period (2008-2012).
- We create a dummy variable Recession when year is between 2008 and 2012.
- We also include year fixed effects.
- Is the coefficient for Recession still of economic meaning?

Slide 51 **Collinearity problem**

- Another example: we want to analyze the performance of branches.
- We include branch FE, firm FE, industry FE, Region FE...
- What is the potential problem here?

Slide 52 **Correlation of the observations within individuals**

- We have a sample of 1000 observations.
- The coefficient of a variable is 2, and the t-statistics is 1.
- We copy and paste our sample: the number of observations doubled!
- What happens to the estimates in terms of value and t-statistics?
- When observations within individuals are highly correlated, we need to clsuter the standard errors!
- `regress y x, vce(cluster group_id)`

Part IV
Conclusion

12 Concluding remarks

Slide 53 **What we learnt**

- Why we are here?
- Basic econometrics knowledge.
- Intermediate econometric knowledge.
- Common mistakes.

Slide 54 **Questions**

- Thank you for your attention.
- Please feel free to ask questions.