Day 1

Introduction

A typical empirical economic thesis is composed of:

Introduction --- Establishing the importance of your work

Literature Review --- Building a theory background

Data and Methodology --- *Empirical strategies*

Results --- Empirical results and interpretations (Robustness/Mechanism)

Conclusion --- Implications of your results

What is causal Inference?

"Causal inference is the leveraging of theory and deep knowledge of institutional details to estimate the impact of events and choices on a given outcome of interest."

Here we do it through *empirical analysis*.

What is empirical analysis?

---Use of data to test a theory or to estimate a relationship between variables.

"Correlation does not mean causality."

--- Even if we have significant results, we still need theory to justify it. Prior knowledge is *required* in order to justify any claim of a causal finding.

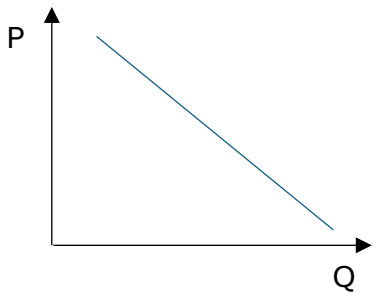
"NO correlation does not mean NO causality."

--- The relationship might be cancelled by other factors (especially when we do not have significant results); therefore, we need to design our *empirical strategy*.

An example with Price elasticity.

Research question: Identify the price elasticity.

We begin with our theory, the demand curve.



From the *demand curve* we have:

$$\epsilon = \frac{\partial \log Q}{\partial \log P}$$

Transforming into one possible empirical model:

$$\log Q_d = \alpha + \delta \log P + \gamma X + u$$

We need two things to estimate δ:

First, we need enough amount of data on price and quantity.

Second, we need the variation of P is independent of u, i.e. exogenous.

Revisiting Probability Theory --- "Regression is conditional mean."

Random Variable: A *function* that "translates" real-world observations into values.

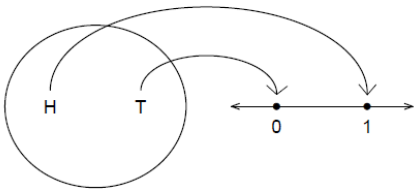


Figure 2.1: A Random Variable is a Function

A continuous random variable can be described by a probability function and follows a certain distribution.

How to describe your Random Variables with values?

Mean or *Expected Value*: Center of mass

Variance or *Standard Deviation*: Spread about mean

Skewness: Symmetry about mean

Kurtosis: Tail thickness

Distributions commonly used: Bernoulli distribution, *Uniform distribution*, *Normal distribution*

The relationship between two random variables:

Covariance: $Cov(X,Y) = \sigma_{XY} = E[(X - EX)(Y - EY)]$

Correlation coefficient: $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$

Uncorrelatedness: $Cov(X,Y) = 0$

Independence: $E(XY) = EX \cdot EY$

Independence indicates uncorrelatedness, BUT NOT the other way around.

Conditional Probability: $\Pr(B|A) = \frac{\Pr(A,B)}{\Pr(A)}$ or $f(x|y) = \frac{f(x,y)}{f(x)}$

Conditional Mean/Expectation: $E(Y|X)$

Law of Iterated Expectation: $EX = E[E(X|Y)]$

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

Suppose our X only takes values of 0 and 1, u is the variation in Y not explained by X.

We have: $\beta_1 = E(y|x = 1) - E(y|x = 0)$, thus, we are calculating the difference in conditional mean.

Causal Inference --- When does regression represent causality?

$$Y = \beta_0 + \beta_1 X + u$$

Other things equal: Factors affecting Y other than X should be accounted for.

Selection problem: X should not be decided in anticipation of Y.

One way to address the above problem: Random experiment, X is randomly selected or implemented.

Simultaneity: pay attention to *when* each event happens.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

Gauss-Markov Assumptions:

1. Linear Parameters

2. Random Sampling: relaxed with *Law of Large Numbers*

3. No *Perfect* Collinearity: High correlation among *X*'s does NOT violate this assumption but leads to *multicollinearity problem*.

4. Zero Conditional Mean: $E(u|x_1, x_2 \ldots x_k) = 0$, error term is not correlated with our explanatory variables.

5. Homoskedasticity: The error u has the same variance given any value of the explanatory variable.

Under these assumptions, the Ordinary Least Squares (OLS) estimator is the best linear unbiased estimator.

For economic studies, to test the significance (*too rare to happen*), we need one more assumption:

6. Normality of error terms

Statistically Significant: Reject null hypothesis (H_0) and β does NOT equal to zero.

R^2: indicates how much of the variance in Y is explained by your model.

Endogeneity problem: X is affected by unobserved factors, *Assumption 4 is violated*.

Prepare for Empirical Analysis --- Data

Data source: check published papers' Data section.

University Database:
https://www2.econ.tohoku.ac.jp/~econlib/material/limit_db.html

DataStream strongly recommended.

Data type: Time series data, Cross-sectional data, *Panel data*

When should we take logarithm function?

Model	Dependent Variable	Independent Variable	Interpretation
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\ln x$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\ln y$	x	$\%\Delta y = 100\beta_1 \Delta x$
Log-log	$\ln y$	$\ln x$	$\%\Delta y = \beta_1 \%\Delta x$

The log-log form gives *constant elasticity*.

---If you are not sure, check others

What unit of measurement should I use?

β will be changed when we use a different unit of measurement. However, if we use the log form, changing in the unit of measurement does NOT affect the slope coefficient except the intercept.

How do I translate qualitive information into regression?

Dummy variables: take the value of 0 and 1.

Dummy variable trap: When using dummy variables, one category must be omitted.

Example:

Suppose we have students from three different classes: A, B and C, we want to introduce dummy variables to indicate which class they belong to.

We will use two dummy variables:

	Class A	Class B	Class C
D_A	1	0	0
D_B	0	1	0

Dependent variable, independent variable, control variables: These variables are decided based on your research context, review your theory.

Empirical Strategies Design --- What model should I use?

Two-way fixed effect:

$$Y = \alpha + \beta X + \delta \boldsymbol{Z} + \boldsymbol{A} + \boldsymbol{T} + u$$

X is our independent variable, $\boldsymbol{Z}$ is our control variables, $\boldsymbol{A}$ is individual fixed effect, and $\boldsymbol{T}$ is time-fixed effect.

--- A set of dummy variables take the value of 1 for current individual/time, and 0 otherwise.

Difference-in-Differences:

$$Y = \alpha + \beta D + \delta \boldsymbol{Z} + \boldsymbol{A} + \boldsymbol{T} + u$$

The independent variable is replaced by a dummy variable indicating whether the event has taken place.

--- Useful when analyzing the impact of an event/policy.

Linear probability model: when Y is a dummy variable

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + u$$

$$\Pr(Y = 1|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Endogeneity Problem --- How do I convince others?

1. Collect more control variables so the endogenous explanatory variable becomes exogenous.

2. Find one or more instrument variables for endogenous explanatory variable.

$$Y = \beta_0 + \beta_1 X + u$$

Instrument Variable:

A good instrument variable Z satisfies:

1. Z is exogenous to the equation, $Cov(Z, u) = 0$

2. Z is relevant for explaining X: $Cov(Z, X) \neq 0$

$$\beta_1 = \frac{Cov(Z, Y)}{Cov(Z, X)}$$

Lagged regression: We bring X one time period earlier so it is by theory not affected by u (*Not recommended*).

Argue with your theory.