Notes on Econometrics

Victor Li

Autumn Sememster, 2023

Contents

1 Preparation and foundational knowledge **3**

1.1 Core of econometrics . . . 3

1.2 Data structure . . . 3

1.3 Understanding the Shape of Your Data: Moments . . . 4

2 Linear Regression Model **5**

2.1 The Population and the Sample . . . 5

2.2 OLS Estimator . . . 6

2.3 The Gauss-Markov Theorem and BLUE . . . 6

2.4 Measures of Fit . . . 7

3 Hypothesis and Test **9**

3.1 The t-test . . . 10

3.2 Confidence Interval . . . 11

4 multi-variate linear regression **12**

5 other regression stuff **13**

6 Panel data **14**

7 Causal effect **15**

7.1 IV . . . 16

7.2 DID . . . 17

7.3 RDD . . . 17

8 Time series data **17**

9 Structural Equations **18**

1 Preparation and foundational knowledge

Reading this note book, you should understand all the basics of higher math.

1.1 Core of econometrics

A student studying econometrics should be able to differentiate that

- causal relationship
- correlative relationship

are two different things. The latter may be intuitive in live but is essentially misguiding in true meaning. [1]

The core pursuit of econometrics is to move beyond simple observation to rigorously estimate causal relationships.

We often observe that two things move together—a correlation—but the goal is to determine if a change in one causes a change in the other. For example, an online retailer might see that on days with high advertising spending, they also have high sales. This is a correlation. But did the advertising cause the increase in sales, or did both rise because of an external factor, like a holiday weekend? Econometrics provides the theoretical framework and practical tools to answer such questions with data.

Only with this kind of understanding, you can bear in mind that the core mission of econometrics is to use statistical methods and mathematical models to give empirical content to economic theory. In simpler terms, it's about **turning broad economic ideas into testable, quantifiable statements**. This is what sets econometrics apart from other causal studies.

1.2 Data structure

The data we use dictates the methods we can apply. Econometric data is typically organized in one of three ways:

- **Cross-Sectional Data**: A snapshot of many different entities at a single point in time. Example: Data on daily sales and advertising expenditure for 500 different online stores on December 1st, 2023. Each row is a different store.
- **Time Series Data**: Observations of a single entity over multiple time periods. Example: Data on the daily sales and advertising expenditure for one specific online store from January 1st to December 31st, 2023. Each row is a different day.
- **Panel Data (Longitudinal Data)**: A combination of the two, observing multiple entities over multiple time periods. Example: Data on the daily sales and advertising expenditure for 500 different online stores, tracked each day for the entire year of 2023. This is incredibly powerful as it allows us to control for factors unique to each store that don't change over time.

[1] Modern day science has brought us intuitive idea. The simple and commonly accepted idea of science is that agnosticism and determinism are required if you want to be rational and truth-seeking. Which, is mostly true.

1.3 Understanding the Shape of Your Data: Moments

Before modeling, we must understand the fundamental characteristics of our variables. The shape of a variable's distribution can be summarized by its statistical moments.

Skewness Skewness measures the asymmetry of a distribution. A perfectly symmetric distribution has zero skewness.

Definition: Skewness (Standardized 3rd Central Moment)

$$\tilde{\mu}_3 = \frac{E[(Y - \bar{Y})^3]}{\sigma_Y^3} \tag{1}$$

The skewness of a random variable Y is the average of its cubed standardized deviations. Cubing the deviations preserves their sign.

- $\approx 0 \iff$ Symmetric, almost like normal distribution.
- $> 0 \iff$ Right Skew, right side lower, meaning more outliers ar right side.
- $< 0 \iff$ Left Skew, left side lower, meaning more outliers ar left side.

Kurtosis Kurtosis measures the "tailedness" of a distribution. It tells us how much of the data's variance is driven by infrequent, extreme events (fat tails) versus frequent, modest deviations.

Definition: Kurtosis (Standardized 4th Central Moment)

$$Kurt = \frac{E[(Y - \bar{Y})^4]}{\sigma_Y^4} \tag{2}$$

The kurtosis of Y is the average of its standardized deviations raised to the fourth power. The fourth power makes extreme values dominate the calculation.

For a normal distribution, the kurtosis is 3. Based on this standard,

- $\approx 3 \iff$ (Mesokurtic): The distribution has tails similar to a normal distribution.
- $> 3 \iff$ (Leptokurtic): "Fat tails." The distribution has more mass in its tails than a normal distribution. In finance, this implies that extreme market movements (crashes or booms) are more likely than a normal model would predict.
- $< 3 \iff$ (Platykurtic): "Thin tails." Extreme events are less likely than in a normal distribution.

2 Linear Regression Model

The linear regression model is the workhorse of econometrics. It provides a simple yet powerful way to model how a dependent variable, Y changes in response to an independent (or explanatory) variable, X.

2.1 The Population and the Sample

It is crucial to distinguish between the unobservable reality we wish to understand and the limited data we have to work with. The reality is the population of data, the part of reality that we are able to direct observe is the sample of the population.

The Population Regression Function (PRF) This is the true, underlying relationship that governs how $Y$is determined. It is a theoretical ideal that we can never observe directly.

For a simplified version of function format (that is the simple form of linear function), the PRF can be stated as:

$$E(Y|X) = \beta_0 + \beta_1 X = E(Y|X) \text{ (regression equation)} \tag{3}$$

$$Y = \beta_0 + \beta_1 X + \underbrace{\mu}_{\text{disturbance}} = E(Y|X) + \mu \text{ (regression model)} \tag{4}$$

- β_0 (Intercept) and β_1(Slope) are the population parameters. They are fixed, unknown constants. β_1 is typically the object of our interest; it represents the true causal effect on Y of a one-unit change in X.
- μ is the unobservable disturbance or error term. It captures all other factors that affect Yapart from X, as well as any inherent randomness. In our retail example, if $Y$is sales and $X$is ad spend, μincludes competitor actions, news events, website glitches, and customer mood.

The Sample Regression Function (SRF) Since we cannot see the entire population, we use a random sample of data to estimate the PRF. The SRF is the estimated relationship for our specific sample.

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \text{ (regression equation)} \tag{5}$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \underbrace{e}_{\text{error}} = \hat{Y}_i + e_i \text{ (regression model)} \tag{6}$$

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimators (or coefficients). They are our data-driven "best guesses" for the true population parameters β_0 and β_1. The "hat" notation ($\hat{\cdot}$) always denotes an estimate.
- e_i is the residual. It is the sample counterpart of the disturbance μ and represents the difference between the actual value Y_i and the predicted value from our model, $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$. Thus, $e_i = Y_i - \hat{Y}_i$.

2.2 OLS Estimator

The Ordinary Least Squares (OLS) Estimator How do we choose the best estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ to draw a line through our data points? The OLS method provides the answer: we choose the values that minimize the sum of the squared residuals.

The goal is to minimize the deviation of estimation from the real world

$$\min \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{7}$$

$$= \min \sum_{i=1}^{n} e_i^2 \tag{8}$$

$$= \min \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2 \tag{9}$$

The OLS Principle The goal is to find the line that is, on the whole, "closest" to all the data points. We define "closeness" as the vertical distance (e_i). By squaring each residual, we ensure that negative and positive deviations don't cancel out and that larger errors are penalized more heavily.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} e_i^2 = \min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{n} [Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)]^2 \tag{10}$$

So OLS is basically an optimization problem.

$$\text{FOC: } \begin{cases} \frac{\partial \min}{\partial \hat{\beta}_0} = 0 \\ \frac{\partial \min}{\partial \hat{\beta}_1} = 0 \end{cases} \Rightarrow \text{yielding the optimal coefficients } \begin{cases} \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \\ \hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \end{cases}$$

Using OLS, we would have fitted value $\hat{Y}_i$ and residual value $\hat{e}_i$

$$\begin{cases} \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = 1, 2 \dots n \\ \hat{e}_i = Y_i - \hat{Y}_i, i = 1, 2 \dots n \end{cases} \tag{11}$$

2.3 The Gauss-Markov Theorem and BLUE

Why should we prefer the OLS method over any other way of fitting a line? The Gauss-Markov theorem provides the theoretical justification. It states that if a set of assumptions holds, then the OLS estimator is the Best Linear Unbiased Estimator (BLUE).

The Gauss-Markov Assumptions (Classical Linear Regression Model - CLRM) is required by:

1. Linearity in Parameters: The model is linear in β_0 and β_1.
2. Random Sampling: The data is a random sample from the population.

3. Variation in X: The sample outcomes for X are not all the same value.

4. Zero Conditional Mean ($E(\mu|X) = 0$): This is the most critical assumption. It states that the unobserved factors in μ are, on average, unrelated to the value of X. In our example, it means that a competitor's promotion (part of μ) is not systematically launched on days when we happen to increase our ad spend (X). A violation of this assumption leads to biased estimates.

5. Homoskedasticity ($var(\mu|X) = \sigma^2$): The variance of the unobserved factors is constant for all values of X. This means the "unpredictability" of sales is the same on high-spend ad days as it is on low-spend ad days.

Note 1 (What is BLUE?)**.** *If the five Gauss-Markov assumptions hold, the OLS estimator has the following desirable properties:*

- **Best:** It has the minimum variance among all linear unbiased estimators. This means OLS is the most precise or efficient.
- **Linear:** The estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of the dependent variable Y.
- **Unbiased:** On average, the estimator will equal the true population parameter. Formally, $E(\hat{\beta}) = \beta$. Your estimate from one sample may be high or low, but if you could repeat the sampling process infinitely, the average of your estimates would be the true value.
- **Estimator:** It is a rule that tells us how to use data to compute an estimate of a population parameter.

In essence, the theorem gives us confidence that, under ideal conditions, OLS is the optimal choice. Much of advanced econometrics is concerned with what to do when one or more of these assumptions are violated.

2.4 Measures of Fit

Once we have estimated a regression model using OLS, a natural question arises: how well does our model actually fit the data? We need metrics to quantify the model's explanatory power.

Decomposing Variance The foundation of the most common goodness-of-fit measure is the decomposition of the total variation in the dependent variable, Y. The total variation is the sum of the squared deviations of each Y_i from its mean $\bar{Y}$. This is called the **Total Sum of Squares (TSS)**.

This total variation can be broken into two parts: the portion that is explained by our model, called the **Explained Sum of Squares (ESS)**, and the portion that is left unexplained, which is captured by the residuals and is called the **Sum of Squared Residuals (SSR)**.

$$\underbrace{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}_{\text{TSS}} = \underbrace{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}_{\text{ESS}} + \underbrace{\sum_{i=1}^{n} e_i^2}_{\text{SSR}} \tag{12}$$

Degree of Freedom In statistics, degrees of freedom (df) refers to the number of values in a final calculation that are free to vary. A good way to think about it is as the number of independent pieces of information that you can use to estimate a parameter.

Note 2 (Degree of freedom)**.**

For the decomposition, degree of freedom is actually $\begin{cases} TSS : n-1 \\ ESS : k \\ SSR : n-k-1 \end{cases}$.

Denominators below are actually degree of freedom. Only in large samples approximated by n.

For Y,

$$\frac{TSS}{n} = var(Y) \tag{13}$$

$$TSS = n \cdot var(Y) = n \cdot \frac{\sum_i^n (Y_i - \bar{Y})^2}{n} = \sum_i^n (Y_i - \bar{Y})^2 \tag{14}$$

$$SE(Y) = \sqrt{var(Y)} = \sqrt{\frac{TSS}{n}} = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n}} \tag{15}$$

For $\hat{Y}$,

$$\frac{ESS}{n} = var(\hat{Y}) \tag{16}$$

$$ESS = \sum_i^n (\hat{Y}_i - \bar{Y})^2 \tag{17}$$

$$SE(\hat{Y}) = \sqrt{var(\hat{Y})} = \sqrt{\frac{ESS}{n}} = \sqrt{\frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{n}} \tag{18}$$

For e,

$$\frac{SSR}{n} = var(e) \tag{19}$$

$$SSR = \sum_i^n (e_i - \bar{e})^2 \tag{20}$$

$$SER = SE(e) = \sqrt{var(e)} = \sqrt{\frac{SSR}{n}} = \sqrt{\frac{\sum_i^n (e_i - \bar{e})^2}{n}} \text{ (also the SE of the regression)} \tag{21}$$

R-squared A common and easy way to measure goodness of fit is by using R^2. It is considered an indicator to judge a model.

The R^2, or the coefficient of determination, formalizes this decomposition into a single, intuitive metric. It measures the fraction of the total variance in Y that is explained by the explanatory variable(s) in the model.

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{\sum_i^n (Y_i - \bar{Y})^2} \tag{22}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = \frac{\text{explained by the estimated model}}{\text{total sample data}} \tag{23}$$

Standard Error of the Regression (SER) While R^2 is a relative measure of fit, the SER is an absolute measure. It estimates the standard deviation of the regression disturbance μ. In practical terms, it tells us the typical size of the regression error, or how far our predictions typically are from the actual outcomes.

Definition: Standard Error of the Regression (SER)

$$SER = s_e = \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{\sum_i^n e_i^2}{n-2}} \tag{24}$$

The SER is measured in the same units as the dependent variable, Y.

A lower SER implies a more accurate model in terms of prediction. In our retail example, if sales (Y) are measured in dollars, an SER of \$500 means our model's predictions of daily sales are typically off by about \$500.

3 Hypothesis and Test

Remember the metrics before are used to test how good a model is. When a number is calculated, is it 100 percent convincing? No, they are not. Because they are calculated under assumptions and simplization of reality.

Would a different sample produce a different estimate? The fundamental question of statistical inference is: how confident are we that our estimated relationship is real and not just a fluke of our particular sample? For instance, is the true effect of advertising on sales, β_1, actually zero?

We care about significance in statistics because it provides a way to quantify the likelihood that an observed result in a study is not due to random chance.

Hypothesis testing provides a formal framework to answer this.

3.1 The t-test

The most common method for testing a hypothesis about a single regression coefficient is the t-test.[2] It follows a structured process to determine whether to accept or reject a claim about the true population parameter.

Step 1: State the Hypotheses

We begin by stating a **null hypothesis** (H_0), which represents the "status quo" or a benchmark of no effect, and an **alternative hypothesis** (H_1), which is what we are trying to establish. The most common test is for statistical significance:

$$\begin{cases} H_0 : \beta_1 = 45812 \\ H_1 : \beta_1 \neq 45812 \end{cases}$$

This is a two-sided test, as we are interested in deviations from zero in either direction (positive or negative).

Step 2: Calculate the t-statistic

The t-statistic (or t-value) measures how many standard errors our estimated coefficient, $\hat{\beta}_1$, is away from the value hypothesized under the null. A larger t-statistic implies that our estimate is less likely to have occurred by random chance if the null hypothesis were true.

Definition: t-statistic

$$t = \frac{\hat{\beta}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} = \frac{\text{Estimation} - \text{Hypothesized Value}}{\text{Standard Error of Estimation}} \tag{25}$$

where $se(\hat{\beta}_1)$ is the standard error of our coefficient estimate, a measure of its sampling variability. When testing for significance, β_{1,H_0} is 0.

Step 3: Make a Decision

We have two common, and equivalent, ways to decide whether our t-statistic is "large enough" to reject the null hypothesis.

The p-value Approach: The p-value is the probability of observing a t-statistic as extreme as, or more extreme than, the one we calculated, assuming the null hypothesis is true.

$$p = 2\Phi(-|t|) \tag{26}$$

Here, Φ is the cumulative distribution function of the standard normal distribution (a good approximation for the t-distribution in large samples). We compare the p-value to a pre-determined **significance level** (α), usually 0.05 (5%), 0.01 (1%), or 0.10 (10%).

- If $p < \alpha$, we **reject the null hypothesis**. The result is "statistically significant at the α level." We have strong evidence that β_1 is not zero.
- If $p \geq \alpha$, we **fail to reject the null hypothesis**. The result is "not statistically significant." We do not have sufficient evidence to claim that β_1 is different from zero.

[2] Take a good look at the word "single" and remember what it stands.

$$p = 2\Phi(-|t|) \begin{cases} < \alpha \iff \text{at reject area} \iff \text{reject null hypothesis } H_0 \iff X \text{ is significant} \\ > \alpha \iff \text{at accept area} \iff \text{accept null hypothesis } H_0 \iff X \text{ is not significant} \end{cases}$$

The Critical Value Approach: Alternatively, we can find a critical value, t_c, from a t-distribution table (or software) that corresponds to our chosen significance level α and degrees of freedom ($df = n - 2$).

- If $|t| > t_c$, our calculated statistic falls in the "rejection region." We **reject the null hypothesis**.
- If $|t| \leq t_c$, our statistic falls in the "acceptance region." We **fail to reject the null hypothesis**.

$$|t| \begin{cases} > t_{\frac{\alpha}{2}} \iff \text{at reject area} \iff \text{reject null hypothesis } H_0 \iff X \text{ is significant} \\ < t_{\frac{\alpha}{2}} \iff \text{at accept area} \iff \text{accept null hypothesis } H_0 \iff X \text{ is not significant} \end{cases}$$

3.2 Confidence Interval

While a t-test gives a yes/no answer about a single hypothesized value, a confidence interval provides a more informative range of plausible values for the true population parameter, β_1.

Constructing a Confidence Interval A 95% confidence interval is constructed by taking our point estimate and adding and subtracting a margin of error, which is determined by the critical t-value and the standard error of the estimate.

Formula: (1-α)% Confidence Interval

$$CI = [\hat{\beta}_1 - t_c \cdot se(\hat{\beta}_1), \quad \hat{\beta}_1 + t_c \cdot se(\hat{\beta}_1)] \tag{27}$$

For a 95% confidence interval, $\alpha = 0.05$, and t_c is the critical value leaving $\alpha/2 = 2.5\%$ in each tail of the t-distribution. For large samples, $t_c \approx 1.96$.

A confidence interval can also be used for hypothesis testing. To test the null hypothesis $H_0 : \beta_1 = 0$, we simply check if 0 lies within the interval.

- If the interval **does not** contain 0, we can reject H_0 at the corresponding significance level.
- If the interval **does** contain 0, we fail to reject H_0.

This provides a measure of both statistical significance and the practical range of uncertainty around our estimate. A very wide interval, even if it excludes zero, signals that our estimate is imprecise.

Note 3 (one-tale or two-tale?)**.** *depending on the hypothesis*

4 multi-variate linear regression

New assumption for MLR:

non-zero finite fourth order moment (kurtois)

OVB, Omitted Variable Bias $\Rightarrow \begin{cases} E(\mu|X) \neq 0 \text{ endogeneity} \\ R^2 \text{ is lower than it should be} \end{cases}$

$$\hat{\beta} \xrightarrow{P} \beta + \underbrace{\frac{\sigma_u}{\sigma_X}\rho_{uX}}_{\text{effect of OVB}} \tag{28}$$

meaning OVB causes estimator to be biased and inconsistent

How to overcome OVB?

- More control variables
- IV
- Panel Fixed Effect model

Adjusted R-squared

$$\bar{R}^2 = 1 - \frac{RSS/n-k-1}{TSS/n-1} = 1 - \frac{n-1}{n-k-1}\frac{RSS}{TSS} \tag{29}$$

Note 4 (difference between R^2 and $\bar{R}^2$)**.**

$\bar{R}^2 < R^2$

$R^2 \in (0,1)$ whereas $\bar{R}^2$ can be sub zero.

how many variables should i add into the model?

AIC

BIC

OLS in MLR:

$$\min_{\{\beta_0,\ldots,\beta_k\}} \sum_i^n (Y_i - \hat{Y}_i)^2 \tag{30}$$

or in matrix form

$$\min (Y - X\hat{\beta})^2 \tag{31}$$

results

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{32}$$

Joint hypothesis test

$$\begin{cases} H_0 : \beta_1 = 0 \& \beta_2 = 0 \\ H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both} \neq 0 \end{cases} \tag{33}$$

F-test

$$F = \frac{1}{2}(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}), \text{ where } \hat{\rho}_{t_1,t_2}^2 \text{ is estimated correlative coefficient} \tag{34}$$

in large sample $\hat{\rho}_{t_1,t_2}^2 \xrightarrow{P} 0$, therefore

$$F = \frac{1}{2}(t_1^2 + t_2^2) \tag{35}$$

simplified F statistics when homoskedasticity

$H_0 : \beta_1 = 0, \beta_2 = 0, \ldots, \beta_q = 0$ and $H_1 : \ldots$

$q =$ number of constraints

unrestricted regression: $Y = Y(X_1, X_2 \ldots X_n)$

restricted regression: $Y = Y(X_1, X_2 \ldots X_i)$ $s.t.$ $g(X_i, X_{i+1} \ldots X_n) = c$

$$F = \frac{(R^2_{\text{unrestricted}} - R^2_{\text{restricted}})/q}{(1 - R^2_{\text{unrestricted}})/(n - k_{\text{unrestricted}} - 1)} \tag{36}$$

Tests for Single Constraints Involving Multiple Coefficients

change the original $Y = \beta_0 + \beta_1 X + \beta_2 Y + u$

to

$$\begin{aligned} Y =& \beta_0 + (\beta_1 - \beta_2)X + \beta_2(X + Y) + u \\ =& \beta_0 + \gamma X + \beta_2 W + u \end{aligned} \tag{37}$$

now testing $\gamma = 0$ is same as testing $\beta_1 = \beta_2$

5 other regression stuff

dummy variables

$D_i = 0$ or 1

dummy variable trap

For 4 cases, model has 4 cases $\Rightarrow$ perfect multicolineary

To fix it, use $k - 1$ dummies for k cases.

non-linear regression

probit model, $Pr(Y = 1|X_1, X_2 \dots X_n) = \Phi(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)$

logit model, $Pr(Y = 1|X_1, X_2 \dots X_n) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\cdots+\beta_n X_n)}}$

---

heteroskedasticity

$var(u|x)$ is variant to different $x_i, i \in n$

heteroskedasticity causes significance test to be meaningless

how to overcome

- heteroskedasticity-robust standard error regression
- GLS
- clustered heteroskedasticity-robust standard error regression

---

multicolinearity

---

interaction term

1) two dummies (DID)

$Y = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_1 D_2 + u$

$$\begin{aligned} \text{effect of } D_2 =& E(Y|D_1, D_2 = 1) - E(Y|D_1, D_2 = 0) \\ =& (\beta_0 + \beta_1 D_1 + \beta_2 + \beta_3 D_1) - (\beta_0 + \beta_1 D_1) \\ =& \beta_2 + \beta_3 D_1 \end{aligned} \tag{38}$$

2) dummy and continuent variable

3) two continuent variables

6 Panel data

fixed effect

fixed effect is used when $corr(X, u) \neq 0$. we use dummies on individual level to capture fixed effect, eliminating endogeneity.

individual fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_i + \mu_{it} \tag{39}$$

time fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_t + \mu_{it} \tag{40}$$

individual and time fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_i + \beta_3 D_t + \mu_{it} \tag{41}$$

Note 5 (individual fixed effect)**.** *can be used to overcome OVB problem*

random effect

random effect is used when $corr(X, u) = 0$

Hausman test

used to decide whether to use fixed effect or randome effect

The null hypothesis is that there is no difference between random effects and fixed effects. If the null hypothesis is rejected, the fixed effects model is adopted, otherwise the random effects model is adopted.

Long panel, $n < t$

Short panel, $n > t$

For short panels, since t is small, it is impossible to explore whether the disturbance term has autocorrelation. For long panels, t is relatively large, so it is necessary to discuss its heteroskedasticity and autocorrelation.

GMM

7 Causal effect

Treatment $D_i = \begin{cases} 1, \text{receiving treatment} \\ 0, \text{not receiving treatment} \end{cases}$

potential untreated outcome $Y_i^0 = Y_i(D = 0)$

potential treated outcome $Y_i^1 = Y_i(D = 1)$

realistic outcome $Y_i = D_i Y_i^1 - (1 - D_i) Y_i^0$

unit treatment effect $\delta_i = Y_i^1 - Y_i^0$

ATT

$$\begin{aligned} \tau_{att} &= E(\delta_i | D_i = 1) \\ &= E(Y_i^1 - Y_i^0 | D_i = 1) \\ &= E(Y_i^1 | D_i = 1) - E(Y_i^0 | D_i = 1) \end{aligned} \tag{42}$$

ATU

$$\begin{aligned}\tau_{atu} &= E(\delta_i|D_i=0)\\ &= E(Y_i^1 - Y_i^0|D_i=0)\\ &= E(Y_i^1|D_i=0) - E(Y_i^0|D_i=0)\end{aligned} \tag{43}$$

ATE

$$\begin{aligned}\tau_{ate} &= E(\delta_i)\\ &= E(Y_i^1 - Y_i^0)\\ &= E[E(Y_i^1 - Y_i^0|D_i)]\\ &= E(Y_i^1 - Y_i^0|D_i=1)\cdot Pr(D_i=1) + E(Y_i^1 - Y_i^0|D_i=0)\cdot Pr(D_i=0)\\ &= \tau_{att}\cdot Pr(D_i=1) + \tau_{atu}\cdot Pr(D_i=0)\end{aligned} \tag{44}$$

7.1 IV

IV conditions

$$corr(Z,\mu) = 0 \tag{45}$$

$$corr(Z,X) \neq 0 \tag{46}$$

2SLS

For a $\begin{cases} Y = \beta_0 + \beta_1 X + \mu \\ X = \pi_0 + \pi_1 Z + \nu \end{cases}$

step 1: regress X on Z, eliminating the part of X related to μ

step 2: regress Y on the estimated $\hat{X}$

step 3: resulting $\hat{\beta} = \frac{s_{YZ}}{s_{XZ}}$

weak IV

First stage least squares has F-value lower than 10. Or first stage regression is not significant.

Identification

n =number of IV and k = number of endogenous variable

an identification problem can be denoted as $\begin{cases} n = k \text{ perfect identification} \\ n > k \text{ over-identification} \\ n < k \text{ unable to identify} \end{cases}$

Sargent test

Hansen J test

C-statistics

7.2 DID

$$y_{it} = \beta x_{it} + \gamma_1 D_i + \gamma_2 D_t + \mu_{it} \tag{47}$$

7.3 RDD

$$Y_i = \alpha + \beta D_i + f(X_i) + \mu_i \tag{48}$$

- D_i denotes if the treatment is received
- X_i contains the treatment variable
- $f(\cdot)$ is to capture the continuity around the cut-off

sharp RD demands $W_i = 1$ if $X_i \geqslant c$

fuzzy RD demands $\lim_{x \to c^+} E(Y|X = x) \neq \lim_{x \to c^-} E(Y|X = x)$

LATE of RDD $\tau = \lim_{x \to c^+} E(Y|X = x) - \lim_{x \to c^-} E(Y|X = x)$

8 Time series data

auto regression

AR(n) $\iff$ $corr(X_t, X_{t-n}) \neq 0$

Stationary

for $\{X_1, \ldots, X_t, \ldots\}$ that any sequence of N period has the same distribution

potential causes for being unstationary:

-

unit root test

study one time series data's stationary relationship. if unit root is tested true in the time series, the time series is not stationary.

common unit root tests are:

- ADF test
- PP test

cointegration test

study multiple non-stationary time series' long-term stationary relationship.

Granger test

used for causal test in mutiple sationary time series

9 Structural Equations