ECONOMETRIC BASIC

ECON 425 LECTURE NOTES

SHIHAO (OWEN) TONG[*]

2021 Winter

University of British Columbia

Lecture given by prof Kevin Song .Template from here. No guarantee for accuracy. For personal use only.

CONTENTS

[*]tongshihaoowen@outlook.com

1 PROBABILITY & STATISTICS INFERENCE

1. Probability Space

1.1 DEFINITION. (Sigma algebra) Given $\mathcal{S}$ the sample space, a collection of subset of $\mathcal{S}$, denoted by $\mathcal{F}$ is called a field if

- $\mathcal{S} \in \mathcal{F}$
- If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$
- if $A, B \in F$, then $A \cup B \in \mathcal{F}$

which means the set $\mathcal{F}$ is closed under operation of union and complement with $\mathcal{S}$ as a subset.

1.2 DEFINITION. (Probability) It is a map defined on the σ-field which is

$$\mathcal{P} : \mathcal{F} \longmapsto [0, 1] \tag{1}$$

1.3 DEFINITION. (Probability space) The probabilty space is a tuple $(\mathcal{S}, \mathcal{F}, \mathcal{P})$ where $\mathcal{P}$ is the probability measure

Then the discrete and continuous random variables are discussed. The way so specify whether the r.v is discrete or not is to see whether the r.v is in a finite set (i.e countably finite). The thing interest here is the representation of a probability. Random variables are still defined as usual. Let $B \subset \mathbb{R}$. Then

$$\mathcal{P}_X(B) = \mathcal{P}\{s \in \mathcal{S} : X(s) \in B\} = \mathcal{P}\{X \in B\} \tag{2}$$

in a discrete case. Then the random vector is

$$\mathbf{X} : \mathcal{S} \longmapsto \mathbb{R}^d \tag{3}$$

i.e $\mathbf{X}(s^*) = (1.7,2,3.4) \in \mathbb{R}^3$. Then this is understood to be

$$\begin{pmatrix} X_1(s^*) \\ X_2(s^*) \\ X_3(s^*) \end{pmatrix}$$

which means the n dimensional vector consists of n random variables. One realization on n random variables.

Then talked about *independence* implies 0 covariance between two random variables while inverse is not true. Another interest thing is the way that estimator is defined which is

1.4 DEFINITION. An estimator of a target parameter θ is a known function

of observed random variables. It is also better to understand to be a map

$$\widehat{\theta} : \mathcal{S} \longmapsto \mathbb{R} \tag{4}$$

Then some criteria or evaluating an estimator are discussed.

1.5 DEFINITION. (Mean square error)

$$\text{MSE} = \mathbb{E}[|\widehat{\theta} - \theta|^2] = Var(\widehat{\theta}) + Bias^2(\widehat{\theta}) \tag{5}$$

Proof of the second equality is simple. Bias is defined as

$$Bias(\widehat{\theta}) = \mathbb{E}(\widehat{\theta} - \theta) \tag{6}$$

Then converges in probability, convergence in distribution are discussed.

1.6 DEFINITION. (Consistency of an estimator) An estimator is said to be consistent if

$$\widehat{\theta} \longrightarrow_p \theta \tag{7}$$

Comments: $\lim_{n\to\infty} \text{MSE}(\hat{\theta}) = 0$ is stronger than consistency which is

$$\lim_{n\to\infty} \text{MSE}(\hat{\theta}) = 0 \longrightarrow \text{consistency}$$

however inverse is not necessarily true

$$\text{consistency} \nrightarrow \lim_{n\to\infty} \text{MSE}(\hat{\theta})$$

sometimes variance goes to 0 while bias does not, which implies the estimator is converging to the wrong theta. Then law of large number (WLLN) and Central limit theorem are given.

1.7 PROPOSITION. *The sample mean* $\overline{X}$ *is an consistent estimator of* μ.

Then we use the mean and variance of the estimator to construct confidence interval. Comments on confidence interval: The C.I is constructed using information from estimator which is indeed a random variable. The way we say the "frequency" 95% will fall in this range is that we repeat to take samples and compute the estimator $\widehat{\theta}$ as an estimate if θ and 95% fall in the range. that is 95% of the realization of $\widehat{\theta}$ fall in the range. Not quite accurate

Another new thing is that for the part of CLT, the numerator, can also be expressed as

$$\sqrt{n}(\widehat{\theta} - \theta_0) \approx \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \tag{8}$$

with w_i i.i.d and $\mathbb{E}(x_i) = 0$which is called the **asymptotic linear representation**. more to add

2. **Hypothesis Testing**

   As usual, review the 2 types of error

   - **Type one error** Reject H_0 as it is true
   - **Type two error** Fail to reject H_0 as it is flase

   Notice the action "reject" which means you take the other option.

   No new things for Hypothesis testing. Notice the essential element mentioned in notes of Stat 305.

3. **Conditional**

   Useful properties for conditional expectation:

   - $\mathbb{E}(a_1 Y_1 + a_2 Y_2|X) = a_1\mathbb{E}(Y_2) + a_2\mathbb{E}(Y_2)$
   - If X and Y are independent then $\mathbb{E}(Y \mid X) = \mathbb{E}(Y)$
   - $\mathbb{E}(Yg(X) \mid X) = g(X)\mathbb{E}(Y \mid X)$
   - $\mathbb{E}_X[\mathbb{E}_Y(Y \mid X)] = \mathbb{E}(Y)$

   The last one is the iterated expectation. Most important one. It is related an interpreted with "information". When there are more than one r.v in the condition, only the one with less information survive for example

   $$\mathbb{E}[\mathbb{E}(Y \mid \cos(x)) \mid X] = \mathbb{E}(Y \mid \cos(X))$$

   1.8 THEOREM.

   $$\mathbb{E}[(Y - g(X))^2] \geq \mathbb{E}[(Y - E[Y \mid X])^2] \tag{9}$$

   proof on class note 09-27.

2 REGRESSION

Mostly talk about the linear regression model. The general form

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{10}$$

New thing is the terminology exogeneous.

2.1 DEFINITION. We say X_i is an exogeneous variable if

$$Cov(u_i, X_i) = 0 \tag{11}$$

2.2 DEFINITION. (Identify) Given the joint distribution of random variables leading to a unique determination of the value of parameter, then we say the parameters are identified.Notice the identification can not include any r.v that are not observable. For example if the condition $c\mathbb{E}(u_i \mid X_i)$ does not hold, then the identification process will be

$$Cov(Y_i, X_i) = \beta_1 Var(X_i) + Cov(u_i, X_i)$$

then β_a is not identifiable.

Assumptions for the model (10) in order to have identified (β_0, β_1):

- $\mathbb{E}(u_i \mid x_i) = 0$ this leads to $\mathbb{E}(u_i) = 0$ and $Cov(u_1, x_i) = 0$ The $\mathbb{E}(u_i) = 0$ is simply from iteration expectation. The second one is from

  $$Cov(u_i, x_i) = \mathbb{E}(x_i u_i) - \mathbb{E}(x_i)\mathbb{E}(u_i) = \mathbb{E}(x_i u_i)$$

  then appeal to iterated expectation

  $$\mathbb{E}(x_i u_i) = \mathbb{E}(\mathbb{E}(x_i u_i \mid x_i)) = \mathbb{E}(x_i \mathbb{E}(u_i \mid x_i)) = 0$$

- $Var(x_i) > 0$. No multicolinerality. (i.e if $Var(x) = 0$ then it becomes a constant regressor then it can not be specified between β_0).

Then start to identify those β parameter. This is a way to identify the parameter.

- β_1: Try the covariacne between Y_i and X_i

  $$\text{Cov}(Y_i, X_i) = \text{Cov}(\beta_0 + \beta_1 X_i + u_i, X_i)$$

  $$= \text{Cov}(\beta_0, X_i) + \text{Cov}(\beta_1 X_i, X_i) + \text{Cov}(u_i, X_i)$$

  $$= \beta_1 Var(X_i)$$

  so that is

  $$\beta_1 = \frac{Cov(X_1, Y_i)}{Var(X_i)}$$

  this is the result that β_1 is **Idnetified**. Recall that the OLS method result in the estimator of β_1 to be $\widehat{Cov}(X_1, Y_i)/\widehat{Var}(X_i)$ which now make sence. This is a fabulous example of being identifiable, which is, given model (the joint distribution) and assumptions, β is identified.

- β_0: Instead of using covriance, simply take expectation on both sides
$$\beta_0 = \mathbb{E}(Y_i) - \beta_1 \mathbb{E}(X_i)$$
so β_0 is identified as well.

Then construct estimator for those parametor.

- **Sample Analogue Estimator** It should be using sample mean to replace the true expectation (see notes). It can be applied to wide range of model not only regression model where OLS only works. So the estimator is
$$\widehat{\beta}_1 = \frac{\widehat{Cov}(X_1, Y_i)}{\widehat{Var}(X_i)}$$
where
$$\widehat{Cov}(X_i, Y_i) = \frac{1}{n}\sum_{j=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})$$
$$\widehat{Var}(X_i) = \frac{1}{n}\sum_{j=1}^{n}(X_i - \bar{X})^2$$
The error of $\widehat{\beta}_1$ involved is
$$\widehat{\beta}_1 = \beta_1 + \underbrace{\frac{\widehat{Cov}(u_i, X_i)}{\widehat{Var}(X_i)}}_{Error}$$

Proof. Consider $\widehat{Cov}(Y_i, X_i)$. We have
$$\begin{aligned}\widehat{Cov}(Y_i, X_i) &= \widehat{Cov}(\beta_0 + \beta_1 X_i + \mu_i, X_i) \\ &= 0 + \beta_1 \widehat{Var}(X_i) + \widehat{Cov}(\mu_i, X_i)\end{aligned}$$
which implies
$$\widehat{\beta_1} = \frac{\widehat{Cov}(u_i, X_i)}{\widehat{Var}(X_i)} + \beta_1$$

□

2.3 FACT. $\widehat{\beta}_1$ is unbiased estimator for β_1.

Proof.
$$\mathbb{E}\,(\widehat{\beta}_1 - \beta_1) = \mathbb{E}\Bigg(\underbrace{\frac{\widehat{Cov}(u_i, X_i)}{\widehat{Var}(X_i)}}_{\text{call it A}}\Bigg)$$

let $\mathbf{X} = (X_1, X_2, ..., X_n)'$, then

$$= \mathbb{E}_{\mathbf{X}}\Big(\mathbb{E}_{A}\big(A \,|\, \mathbf{X}\big)\Big)$$

then by the fact that $\mathbb{E}(Yg(X) \,|\, X) = g(X)\mathbb{E}(Y \,|\, X)$ we have

$$= \mathbb{E}_{\mathbf{X}}\left(\frac{1}{\widehat{Var}(x_i)}\mathbb{E}\big(\widehat{Cov}(u_1, x_i) \,|\, \mathbf{X}\big)\right)$$

then consider the $\widehat{Cov}(u_1, x_i)$ we have

$$\widehat{Cov}(u_1, x_i) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{\mu})$$

$$= \frac{1}{n}\sum_{i=1}^{n} u_i(x_i - \bar{x}) - \underbrace{\frac{1}{n}\sum_{i=1}^{n}\bar{\mu}(x_i - \bar{x})}_{=0}$$

$$= \frac{1}{n}\sum_{i=1}^{n} u_i(x_i - \bar{x})$$

then we have

$$\mathbb{E}(\widehat{Cov}(u_i, x_i) \,|\, \mathbf{X}) = \frac{1}{n}\mathbb{E}(\sum_{i=1}^{n} u_i(x_i - \bar{x}) \,|\, \mathbf{X})$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(u_i(x_i - \bar{x}) \,|\, \mathbf{X})$$

again apply the property of $\mathbb{E}(Yg(X) \,|\, X) = g(X)\mathbb{E}(Y \,|\, X)$ we have

$$= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\underbrace{\mathbb{E}(u_i \,|\, \mathbf{X})}_{=0} = 0$$

Notice the last under brace we have $\mathbb{E}(u_i \,|\, \mathbf{X})$ to be zero. This can be thought that

$$\mathbb{E}(u_i \,|\, \mathbf{X}) = \mathbb{E}(u_i \,|\, X_i)$$

since for any other X, they do not provide additional information for u_i. □

- **OLS** The result if exactly the same as sample analogue.

Then the generalized slysky lemma is introduced.

2.4 THEOREM. *Let* $X_n \to_d X$ *and* $Y_n \to_p c$. *Then*

$$g(X_n, Y_n) \to_d g(X, c) \tag{12}$$

Notice when apply this theorem, a.s and p are all stronger than d. So when these occurs, slusky lemma also works.

2.5 FACT. (Consistency of $\widehat{\beta_1}$) $\widehat{\beta_1} \longrightarrow_p \beta_1$ which is

$$\lim_{n\to\infty} \mathcal{P}(|\widehat{\beta_1} - \beta| < \epsilon) = 1 \tag{13}$$

Proof. This is same as showing the error term converges in probability to 0. Thus

$$\widehat{\beta_1} - \beta_1 = \frac{\widehat{Cov}(u_i, X_i)}{\widehat{Var}(X_i)} = \frac{1/n \sum (X_i - \overline{x})(u_i - \overline{u})}{1/n \sum (x_i - \overline{x})^2}$$

then we call two facts, which are also easy to check, which are

- **Sample variance converges in probability to population variance**
- **Sample covariance converges in probability to population covariance**

which are all obtained by Slysky's lemma. Then we have

$$\frac{\widehat{Cov}(u_i, X_i)}{\widehat{Var}(X_i)} \to_p \frac{Cov(u_i, Y_i)}{Var(X_i)} = 0$$

by our assumption. So $\widehat{\beta_1} \to_p .\beta_1$. Consistency proved. □

Then given one more assumption, the homostet. $\mathbb{E}(u_i^2 \mid X_i) = \sigma^2$ we are able to show that

$$\sqrt{n}(\widehat{\beta_1} - \beta_1) \to_d N(0, v^2 = \frac{\sigma^2}{Var(X_i)})$$

where we can show $\widehat{v^2} \to_p v^2$ by sample analog estimator. Finally we have

$$\sqrt{n}(\widehat{\beta_1} - \beta_1)\frac{1}{v} \to_d N(0, 1)$$

Proof. Aim to show the convergence of $\sqrt{n}(\widehat{\beta_1} - \beta_1)$. Start with this expression we have

$$\sqrt{n}(\widehat{\beta_1} - \beta_1) = \frac{\sqrt{n}\widehat{Cov}(u_i, X_i)}{\widehat{Var}(X_i)}$$

The denominator is still converges in probability to the population variance. We focus on the numerator.

$$\sqrt{n}\widehat{Cov}(u_i, \mathrm{X}_i) = \frac{1}{\sqrt{n}} \sum u_i(x_i - \overline{x})$$

$$= \frac{1}{\sqrt{n}} \sum u_i(x_i - \mathbb{E}(\mathrm{X}_i)) + \underbrace{\frac{1}{\sqrt{n}} \sum u_i(\mathbb{E}(\mathrm{X}_i) - \overline{x})}_{\text{by Slysky's lemma} \longrightarrow_d 0}$$

the second term can be thought as

$$\sum u_i(\mathbb{E}(\mathrm{X}_i) - \overline{x}) = \underbrace{\left(\frac{1}{\sqrt{n}} \sum u_i\right)}_{\mathrm{CLT} \longrightarrow_d \mathrm{N}(0, var(u_i))} \underbrace{\left(\mathbb{E}(\mathrm{X}_i) - \overline{x}\right)}_{\mathrm{WLLN} \longrightarrow_p 0} \longrightarrow_{\textbf{d or p}} 0$$

where the first part is base on out assumption that the $\mathbb{E}(u_i) = 0$ and the converges in p or d is because converges in d to a constant implies converges in p to the constant. The for the first part we observe that

$$\mathbb{E}\Big(u_i(x_i - \mathbb{E}(\mathrm{X}_i)\Big) = \mathbb{E}\Big(u_i x_i\Big) - \mathbb{E}(u_i)\mathbb{E}(x_i) = Cov(u_i, x_i) = 0$$

again by our assumption. So by CLT, we have

$$\frac{1}{\sqrt{n}} \sum u_i(x_i - \mathbb{E}(\mathrm{X}_i)) \longrightarrow_d \mathrm{N}(0, Var(u_i(x_i - \mathbb{E}(\mathrm{X}_i))))$$

Mon comment: Obtaining an asymptotic normal by CLT from a series is important especially when other asymptotic behaviour can not be obtained. So together we have

$$\sqrt{n}\widehat{Cov}(u_i, \mathrm{X}_i) \longrightarrow_d \mathrm{N}\Big(0, Var\Big(u_i(x_i - \mathbb{E}(\mathrm{X}_i))\Big)\Big)$$

Since the variance involves u_i, we apply our new assumption here

$$Var\Big(u_i(x_i - \mathbb{E}(\mathrm{X}_i)\Big) = \mathbb{E}\Big(u_i^2(\mathrm{X}_i - \mathbb{E}(\mathrm{X}_i))^2\Big) - 0$$

$$= \mathbb{E}\Big\{\mathbb{E}\Big(u_i^2(\mathrm{X}_i - \mathbb{E}(\mathrm{X}_i))^2 \,\Big|\, \mathrm{X}_i\Big)\Big\}$$

$$= \mathbb{E}\Big\{(\mathrm{X}_i - \mathbb{E}(\mathrm{X}_i))^2 \underbrace{\mathbb{E}\Big(u_i^2 \,\Big|\, \mathrm{X}_i\Big)}_{=\sigma^2}\Big\} = \sigma^2\mathbb{E}\Big((\mathrm{X}_i - \mathbb{E}(\mathrm{X}_i))^2\Big) = \sigma^2 Var(\mathrm{X}_i)$$

Iterated expectation is extremely useful when given assumption is in

conditional form. Finally integrate with the denominator

$$\sqrt{n}(\widehat{\beta}_1 - \beta_1) = \frac{\sqrt{n}\widehat{Cov}(u_i, X_i)}{\widehat{Var(X_i)}} \longrightarrow_d \; N\Big(0, \underbrace{\frac{\sigma^2}{Var(X_i)}}_{\text{Call it } v^2}\Big)$$

Then Consider the sample analog estimator of the variance which is

$$\widehat{v^2} = \frac{\widehat{\sigma^2}}{\widehat{Var}(X_1)} \longrightarrow_d \frac{\sigma^2}{Var(X_i)}$$

by WLLN and slutsky's lemma. Then again by slutsky's lemma

$$\frac{\sqrt{n}(\widehat{\beta}_1 - \beta_1)}{\widehat{v}} \longrightarrow_d N(0,1)$$

where $v^2 = \sigma^2/Var(X_i)$. Notice the last step, the $N(0, v^2)$ is an **asymptotic result**, so we are not able to simply do algebra to obtain the standard normal distribution, we have to appeal to LLN and slutsky's lemma. □

So this implies our confidence interval is exactly a asymptotic result.

2.1 *Bivariate Linear Regression*

Linear model with two variables. Form

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 Z_i + \mu_i \tag{14}$$

Assumption, analogy from univariate model

- $\mathbb{E}(u_i \mid X_i, Z_i) = 0$

  This condition implies that

  $$\mathbb{E}(u_i) = 0 \text{(by iterated expectation)}$$

  $$Cov(u_i, X_i) = Cov(u_i, Z_i) = 0$$

  the second equality is by

  $$\mathbb{E}(u_i X_i) = \mathbb{E}(\underbrace{\mathbb{E}(u_i \mid X_i, Z_i)}_{=0} X_i)$$

  same for Z_i.

- $Var(X_i) > 0, Var(Z_i) > 0$
- X_i and Z_i are not perfectly correlated.

Then we identify the parameter in the same way as we do in univariate regression which is

$$Cov(Y_i, X_i) = \beta_1 Var(X_i) + \beta_2 Cov(Z_i, X_i) \tag{15}$$

$$Cov(Y_i, Z_i) = \beta_1 Cov(X_i, Z_i) + \beta_2 Var(Z_i) \tag{16}$$

Then two equations, two target parameters, we eventually find

$$\beta_1 = \frac{\widetilde{\beta_1} - \rho_{Z_i,X_i} \frac{Cov(Y_i,Z_i)}{\sqrt{Var(Z_i)Var(X_i)}}}{1 - \rho^2_{Z_i,X_i}} \tag{17}$$

where

$$\widetilde{\beta_1} = \frac{Cov(Y_i, X_i)}{Var(X)}, \quad \rho_{Z_i,X_i} = \frac{Cov(Z_i, X_i)}{\sqrt{Var(Z_i)Var(X_i)}}$$

so the above reaffirm that **if X_i and Z_i are perfectly linearly correlated, the the model (parameters) are not able to be identified. This is equivalently shown as**

- **if $\rho^2_{Z,X} = 1$, then β_1 is not identified, which is the case of perfectly linearly correlated.**
- **if $\rho_{Z,X} = 0$, then $\beta = \widetilde{\beta_1}$ which is the case of perfectly non-linearly correlated. Where $\widetilde{\beta_1}$ is regressing Y_i on X_i, with Z_i omitted. β_1 is regressing on both X_i and Z_i. That is, if X and Z are perfectly non-linearly correlated, then the result of its corresponding coefficient is the same as regressing on them respectively in univariate linear model.**

Also by symmetry, we can figure our β_2 in the same manner.

After identification, look for estimator. The general sample analogue estimator is still by replacing whatever we can by its sample counterpart.

Estimator

Assume the true model is the bi-variate model. We still using the sample analogy estimator of the univariate model. The case is, true model is

$$Y_i = \beta_0 + \beta_1 X_i + \beta_0 Z_i + u_i$$

while we mis-specified the model to be

$$Y_I = \gamma_0 + \gamma_1 X_i + v_i$$

(i.e we actually regressing on X_i solely). Then we try the sample analogy

estimator of γ_1 which give

$$\widehat{\gamma_1} = \frac{\widehat{Cov}(Y_i, X_i)}{\widehat{Var}(X_i)} = \frac{\widehat{Cov}(\overbrace{\beta_0 + \beta_1 X_i + \beta_0 Z_i + u_i}^{\text{the true model}}, X_i)}{\widehat{Var}(X_i)}$$

$$= \beta_1 + \beta_2 \frac{\widehat{Cov}(Z_i, X_i)}{\widehat{Var}(X_i)}$$

we observe by WLLN and Slusky's lemma

$$\widehat{\gamma_1} = \beta_1 + \beta_2 \frac{\widehat{Cov}(Z_i, X_i)}{\widehat{Var}(X_i)} \longrightarrow_p \beta_1 + \underbrace{\beta_2 \frac{Cov(Z_i, X_i)}{Var(X_i)}}_{\text{Asymptotic Bias}} \tag{18}$$

More specifically we call this the **Omitted variable bias** since this kind of bias is caused by omitting one variable. The above observation indicates, and reaffirms, that as long as the Z_i and X_i are not perfectly linearly correlated, the sample analogy estimator in uni-variate model will converges to some point away from true parameter in bivariate model. If asymptotic bias is not zero, then the estimator cannot be consistent, which is

- Asymtotic bias = 0 $\not\Longleftrightarrow$ Consistency
- Consistency $\Longrightarrow$ Asymptotic bias = 0

2.6 DEFINITION. The asymptotic bias is defined as

$$\operatorname*{plim}_{n\to\infty} \widehat{\theta} - \theta_0 \tag{19}$$

where $\operatorname{plim}_{n\to\infty} \widehat{\theta} = \gamma$ if $\widehat{\theta} \to_p \gamma$. The θ_0 is target parameter.

2.7 DEFINITION. (partial effect) Let $m(x, z) = \mathbb{E}(X_i = x \mid Z_i = z) = \beta_0 + \beta_1 x + \beta_2 xz$. Then the partial effect is defined as

$$\frac{\partial m}{\partial x}(x, z)$$

and the average partial effect is

$$\mathbb{E}\left(\frac{\partial m}{\partial x}(x, z)\right)$$

which is the average derivative of x.

2.2 *Multivariate Regression*

Assume we have more than 2 random variable. The assumption is still the same as the uni and bivariate model which are

- $\mathbb{E}(u_i \mid \mathrm{X}_i, \mathrm{Z}_i, ...) = 0$
- $\mathbb{E}(u_i)$ which leads to $Cov(u_i, \mathrm{X}_i) = Cov(u_i, \mathrm{Z}_i) = ... = 0$

Then comes to estimation. OLS is applied. While the interest thing here is the way that OLS is motivated, that is directly from the identification results. So the results is the same as minimizing

$$\min_{b_0, b_1, b_2} \mathbb{E}\Big((\mathrm{Y}_i - b_0 - b_1 \mathrm{X}_i - b_2 \mathrm{Z}_i)^2\Big) \tag{20}$$

then the sample counterpart becomes

$$\min_{b_0, b_1, b_2} \frac{1}{n} \sum_{i=1}^{n} \Big((\mathrm{Y}_i - b_0 - b_1 \mathrm{X}_i - b_2 \mathrm{Z}_i)^2\Big)$$

guaranteed by the WLLN and Slusky's lemma. Go along the same way as before, we need second moment assumption while in this case we assume

$$\mathbb{E}\Big(u_i^2 \mid \mathrm{X}_i, \mathrm{Z}_i\Big)$$

which is the conditional hetero assumption. Notice this is stronger than $\mathbb{E}(u_i^2 \mid \mathrm{X}_i) = \sigma^2$. Then in the bivariate model (14), we have

$$\sqrt{n}(\widehat{\beta_1} - \beta_1) \longrightarrow_d \mathrm{N}(0, v^2), \quad v^2 = \frac{1}{1 - \rho_{\mathrm{Z,X}}^2} * \frac{\sigma^2}{\mathrm{V}ar(\mathrm{X}_i)} \tag{21}$$

Then consider the robustness and efficiency.

2.8 EXAMPLE. Consider the uni and bivariate model which are

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{X}_i + u_i$$

$$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{X}_i + \beta_2 \mathrm{Z}_i + u_i$$

where the univariate one is the true model. Using the bivariate model is still valid. This is the robustness. The second model is a larger model than the first one, so the univariate model can be considered as a special case of the second model. Then if looking at the asymptotic behaviour we will find that the second one has a larger asymptotic variance.

3 CAUSALITY, ENDOGENITY

If we only care about prediction then causality is not a problem since we don't care.

3.1 DEFINITION. (Best linear predictor) The best linear predictor of regression coefficient is the ones that minimize the MSE

$$\min_{\alpha,\beta} \mathrm{E}(\mathrm{Y} - \alpha - \beta \mathrm{X})^2$$

where the model is $\mathrm{Y} = \alpha + \beta \mathrm{X} + u$

3.2 DEFINITION. (Endogenity) This is suggested from $Cov(u_i, \mathrm{X}_i) \neq 0$

The problem caused by Endogenity is that consider uni-variate model then we get

$$\widetilde{\beta} \longrightarrow_p \frac{\mathrm{Cov}(\mathrm{Y_i}, \mathrm{X_i})}{\mathrm{Var}(\mathrm{X_i})} = \beta_1 + \frac{\mathrm{Cov}(\mathrm{u_i}, \mathrm{X_i})}{\mathrm{Var}(\mathrm{X_i})}$$

When the regression is ran and significance is shown. However you are not able to separate which part contributes to it. In extreme case, say $\beta = 0$ however test still pass.
Now we consider the case that $Cov(u_i, \mathrm{X}_i) \neq 0$ which is **endogeneity**. Consider univariate model with non-zero covariance between u_i and X_i (also clearly the $\mathbb{E}(u_i \mid \mathrm{X}_i) \neq 0$) which is

$$\mathrm{Y}_i = \widetilde{\beta}_0 + \widetilde{\beta}_1 \mathrm{X}_i + \widetilde{u}_i$$

. We still identify the parameter in the same was as before. However now the identified parameter becomes

$$\widetilde{\beta}_1 = \frac{\mathrm{Cov}(\mathrm{Y_i}, \mathrm{X_i})}{\mathrm{Var}(\mathrm{X_i})} + \frac{\mathrm{Cov}(\mathrm{u_i}, \mathrm{X_i})}{\mathrm{Var}(\mathrm{X_i})} = \beta_1 + \frac{\mathrm{Cov}(\mathrm{u_i}, \mathrm{X_i})}{\mathrm{Var}(\mathrm{X_i})}$$

and the sample analogy estimator becomes

$$\widehat{\widetilde{\beta}}_1 = \frac{\widehat{\mathrm{Cov}(\mathrm{Y_i}, \mathrm{X_i})}}{\widehat{\mathrm{Var}(\mathrm{X_i})}} + \frac{\widehat{\mathrm{Cov}(\mathrm{u_i}, \mathrm{X_i})}}{\widehat{\mathrm{Var}(\mathrm{X_i})}} \longrightarrow_p \beta_1 + \underbrace{\frac{\mathrm{Cov}(\mathrm{u_i}, \mathrm{X_i})}{\mathrm{Var}(\mathrm{X_i})}}_{\textbf{Endogenity Biase}}$$

easy to see the convergence. Notice the first part converges to β_1. This β_1 is the one under the case that $\mathbb{E}(u_i \mid \mathrm{X}_i) = 0$ which does not exists. So this means if you still use the original estimator which is $\frac{\widehat{\mathrm{Cov}(\mathrm{Y_i}, \mathrm{X_i})}}{\widehat{\mathrm{Var}(\mathrm{X_i})}}$ then as you increases n, the sample size, the estimator converges to β_1 but not $\widetilde{\beta}_1$ and the remaining part is bias.

1. **Source of Endogeneity**

   - **Omitted Variables** Bias cause by laking of variable. Consider true model to be

     $$\mathrm{Y}_i = \beta_0 + \beta_1 \mathrm{X}_i + \beta_2 \mathrm{Z}_i + u_i$$

     where $\mathrm{Cov}(\mathrm{X_i}, \mathrm{u_i}) = \mathrm{Cov}(\mathrm{Z_i}, \mathrm{u_i}) = 0$ while the model we approaches

in practice is

$$Y_i = \beta_0 + \beta_1 X_i + \underbrace{v_i}_{=\beta_2 Z_i + u_i}$$

By doing the same procedures as before we can obtain the bias by

$$\widehat{\beta_1} = \frac{\widehat{Cov(Y_i, X_i)}}{\widehat{Var(X_i)}} \longrightarrow_p \beta_1 + \beta_2 \frac{Cov(X_i, Z_i)}{Var(X_i)}$$

- **Measurable Error Bias**

  The situation is assume the true model is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i^* + u_i \\ &= \beta_0 + \beta_1 X_i - \beta_1 \epsilon_i + u_i \end{aligned}$$

  while the observable, practical model is

$$Y_i = \beta_0 + \beta_1 \underbrace{X_i}_{X_i = X_i^* + \epsilon_i} + v_i$$

  Then the measurable error bias becomes

$$\text{Bias}(\widehat{\beta_1}) = \frac{Cov(v_i, X_i)}{Var(X_i)}$$

  Notice comparing the true and practical model, v_i actually equals $-\beta_1 \epsilon_i + u_i$. So

$$\begin{aligned} Cov(X_i, v_i) &= Cov(X_i^* + \epsilon_i, -\beta_1 \epsilon_i + u_i) \\ &= -\beta_1 \, Cov(X_i^*, \epsilon_i) - \beta_1 \, Var(\epsilon_i) + Cov(\epsilon_i, u_i) \end{aligned} \tag{22}$$

  Then we assume **pure measurement error** which are

  - $Cov(\epsilon_i, X_i^*) = 0$
  - $Cov(\epsilon_i, u_i) = 0$

  So finally $Cov(X_i, v_i) = -\beta_1 \, Var(\epsilon_i)$ and by substituting $Var(X_i) = Var(X_i^*) + Var(\epsilon_i)$

$$\widehat{\beta_1} \longrightarrow_p \beta_1 - \underbrace{\beta_1 \frac{Var(\epsilon_i)}{Var(X_i)}}_{\textbf{Bias}} = \beta_1 \underbrace{\frac{Var(X_i^*)}{Var(X_i^*) + Var(\epsilon_i)}}_{\in (0,1)}$$

  So, conclusion, the measurement error only scale down the target parameter. If the variance of ϵ is so big, then it pulls down the effect to almost zero. So the significance test may not past due the the measurement error. It reduce the power of the test. Not to much of

this we can do with this.

A question raised from lecture. Can we check $\widehat{\text{Cov}(\hat{u}_i, X_i)}$ to check endoginity? No. Since $\widehat{u_i} = Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_i$ and thus by design $\widehat{\text{Cov}(\hat{u}_i, X_i)} = 0$ always. There is no way we can check endoginity. We may need to search for new data sources.

- **Simultaneous Causality Bias** Make up

3.1 *Method to deal with endoginity*

This course mainly talk about instrument variable (IV), a 'tool variable'. Consider univariate model again with endoginity which is

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad \text{Cov}(u_i, X_i) \neq 0$$

The IV need to satisfy two conditions:

- **Validity, IV exogeneity condition** $\text{Cov}(u_i, Z_i) = 0$
- **Relevance, IV relevancy condition** $\text{Cov}(X_i, Z_i) \neq 0$

It is easy to construct such an IV. This two condition leads to identification of β_1

$$\text{Cov}(Z_i, Y_i) = \beta_1 \text{Cov}(Z_i, X_i) \implies \beta_1 = \frac{\text{Cov}(Z_i, Y_i)}{\text{Cov}(Z_i, X_i)} \tag{23}$$

Includes Z_i in the model does not help identification and then cause a new parameter to identify. The procedure is as follows.

- **1st stage regression** We first regression X_i on Z_i by simple linear regression

$$X_i = \pi_0 + \pi_1 Z_i + v_i \tag{24}$$

where the basic assumptions for identification (consistency) are still hold. Then we have

$$\widehat{X_i} = \widehat{\pi_0} + \widehat{\pi_1} Z_i \tag{25}$$

with

$$\widehat{\pi_1} = \frac{\widehat{\text{Cov}(X_i, Z_i)}}{\widehat{\text{Var}(Z_i)}}$$

- **2nd stage regression** Then we regress on the purified X which is $\widehat{X_i}$

$$Y_i = \beta_0 + \beta_1 \widehat{X_i} + u_i$$

the sample counterpart is

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1}\widehat{X_i}$$

where

$$\begin{aligned}\widehat{\beta_1} &= \frac{\widehat{Cov(Y_i, X_i)}}{\widehat{Var(X_i)}} \\ &= \frac{\widehat{Cov(Y_i, \widehat{\pi_0} + \widehat{\pi_1} Z_i)}}{\widehat{\pi_1^2 Var(Z_i)}} = \frac{\widehat{Cov(Y_i, Z_i)}}{\widehat{Var(Z_i)}} \frac{\widehat{Var(Z_i)}}{\widehat{Cov(X_i, Z_i)}} \\ &= \frac{\widehat{Cov(Y_i, Z_i)}}{\widehat{Cov(X_i, Z_i)}} \longrightarrow_p \beta_1 \quad \text{(from 23)}\end{aligned}$$

Notice in the 1st stage, we are NOT assuming any causal relation at all. It is pure auxiliary.

Notice that

$$\widehat{\beta_1} = \frac{\widehat{Cov(Y_i, Z_i)}}{\widehat{Cov(X_i, Z_i)}} = \frac{\widehat{Cov(\beta_0 + \beta_1 X_i + u_i, Z_i)}}{\widehat{Cov(X_i, Z_i)}} = \beta_1 + \frac{\widehat{Cov(u_i, Z_i)}}{\widehat{Cov(X_i, Z_i)}}$$

Notice that the true model is still $Y_i = \beta_0 + \beta_1 X_i + u_i$. It is true but with endogeinity. Then

$$\sqrt{n}(\widehat{\beta_1} - \beta_1) = \sqrt{n}\frac{\widehat{Cov(u_i, Z_i)}}{\widehat{Cov(X_i, Z_i)}} \longrightarrow_d N(0, v^2)$$

then

$$\begin{aligned}\sqrt{n}\widehat{Cov(u_i, Z_i)} &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i \left(Z_i - \bar{Z}_i\right) \\ &= \underbrace{\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i \left(Z_i - \mathbb{E}\left(Z_i\right)\right)}_{\text{CLT} \to_d N(0, Var(u_i(Z_i - \mathbb{E}[Z_i])))} + \underbrace{\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} u_i\right)}_{\text{CLT} \to 0} \underbrace{\left(\mathbb{E}\left(Z_i\right) - \bar{Z}\right)}_{\text{LLN} \to 0}\end{aligned}$$

the denominator converges in p to $Cov(X_i, Z_i)$ which eventually generate the limit distribution.

Comparing LS and 2SLS Further more to get the limit distribution more

neat we assume $\mathbb{E}(u_i^2 \mid X_i, Z_i) = \sigma^2$. We have seen two results

$$\sqrt{n}(\widehat{\beta}_{2SLS} - \beta_1) \longrightarrow_d N(0, \frac{\text{Var}(u_i(Z_i - \mathbb{E}[Z_i]))}{\text{Cov}(X_i, Z_i)^2} = \frac{\sigma^2 \text{Var}(Z_i)}{\text{Cov}(X_i, Z_i)^2})$$

$$\sqrt{n}(\widehat{\beta}_{LS} - \beta_1) \longrightarrow_d N(0, \frac{\sigma^2}{\text{Var}(X_i)})$$

where the 2SLS has a larger SE by the following inequality

3.3 THEOREM. *(Cauchy-Schwory Inequality) Given two random variable* X_i *and* Y_i

$$\text{Var}(X)\,\text{Var}(Y) \geq \text{Cov}(X, Y)^2 \tag{26}$$

which suggests that using IV will increase the SE so then lose power.

As we goes to the model with more than one endogenous variables, we need more IV to achieve more equations for identification by using $\text{Cov}(u_i, X_j)$. So we need at least same number of IV as the number of endogenous for identification purpose. Lets set up the model to be

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \sum_{j=k+1}^{r} \beta_j X_{ij} + u_i$$

let the first k to be endogenous and $r - k$ exogenous r.v.

The case can also be **over identified** which means the number of IV is greater than the number of exogenous. When add more IV, there will be more restriction which is the $\text{Cov}(\text{IV}, u_i) = 0$. With more IV which leads to a smaller SE and higher power.

3.2 *Checking the Validity of Instrumental Variable*

We are going to test whether the two assumptions of IV, the relevancy and validity of the IV.

Relevancy

Let assume the whole model to be

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + ... + \beta_{1+r} W_{ri} + u_i \tag{27}$$

where X_i is endogenous and the remaining are exogenous. Then our first stage regression becomes

$$X_i = \pi_0 + \sum_{j=1}^{m} \pi_j Z_{ji} + \sum_{q=1}^{r} \pi_{m+q} W_{qi} + v_i \tag{28}$$

Notice we includes all other exogenous regressors in the first stage regression. Then we will test on the following

$$\begin{aligned} &H_0 : \pi_1 = ... = \pi_m = 0 \\ &H_a : \text{At least one equality above is wrong} \end{aligned} \tag{29}$$

We apply F – *test* here. Notice in practice, the programming is in default testing on

$$H_0 : \pi_1 = ... = \pi_m = \pi_{m+1} = ... = \pi_{m_r} = 0$$

This includes all the coefficient of the exogenous variables. This is WRONG. Logic is this: We are happy to see the null to be rejected since this confirms our guess that at least one of the instrument variables is relevant. However, this can be leaded by the non-zero of those coefficients of exogenous r.v which actually indicates irrelevancy of any of the IV.

Weak instrument This is caused by the weak correlation between the IV and endogenous r.v. Usually this leads to the problem that the asymptotically distribution of $\sqrt{n}(\widehat{\beta} - \beta)$ does not approximate normal distribution. So the asymptotic results is misleading.

Validity

Test $\text{Cov}(u_i, Z_i) = 0$ or not. First idea on doing $\widehat{\text{Cov}(\widehat{u_i}, Z_i)} = 0$ not works. The reason is the same as the discussion on page 15. In short, the parameter is obtained by the same setting so they are numerically identity. This means the following test

$$H_0 : \text{Cov}(Z_i, u_i) = 0$$

$$H_a : \text{Cov}(Z_i, u_i) \neq 0$$

where $Y_i = \beta_0 + \beta_1 X_i + u_i$ and $X_i = \pi_0 + \pi_1 Z_i + v_i$ is **not testable**. So

3.4 FACT. For IV validity test, when our case is **just-identified**, the test is not testable. In order to be testable, we need the case to be **over-identified**.

Let consider the over identified case. Let

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$X_i = \pi_0 + \pi_1 Z_{i1} + \pi_2 Z_{i2} + v_i$$

which test

$$H_0 : \text{Cov}(Z_{i1}, u_i) = \text{Cov}(Z_{i2}, u_i) = 0$$

$$H_a : \text{At least one equality above is not true}$$

Go to notes 5 for the J-test and its construction. make up later

4 PANEL MODEL

Panel data or time panel data, is that data observed in a series of time and cut at a time point so you get a panel data. If the time length is very large so you get a **long panel** which is usually dealt by time series analysis (i.e t is large) and the **short panel** is your sample size n is large. Also we have the so-called **rotational cross section data** which can be thought as a subset of panel data that each column are drawn randomly from a the column of panel data. This means horizontally along the time dimension, the data is not from the same household while the panel data make sure that horizontal dimension are from the same household so calls that **with-in group**, correlated and vertically called **between group** and i.i.d.

We should notice that the usual asymptotic analysis may be not applicable here. Regular method and theorem usually need n get to infinity while we have another time dimension here. So we should be careful about which dimension you are applying the asymototic analysis on. In this course, we deal with panel data.

4.1 *Linear Panel Regression with Fixed Effect*

Assume panel model. Let i indexing the i th sample and t indexing time at t. The model becomes

$$Y_{it} = X_{it}\beta + \alpha_i + u_{it} \tag{30}$$

where:

- β is the causal effect of X_{it} as usual and observable
- Y_{it} also observable
- α_i is the fixed effect, unobservable, time-INvariant heterg. The fixed effect α_i are potentially correlated with X_{it} and this generate endoginity.

Endoginity here can be removed without using IV thanks to first, α_i is time-invariant (i.e omitted variable is time invariant) and second, X_{it} is time varying. We assume, for identification

- $\mathbb{E}[u_{it} \mid X_{i1}, X_{i2}, ..., X_{iT}] = 0$ for $\forall i, T$
- $\sum_{t=1}^{T} \mathbb{E}[(X_{it} - \frac{1}{T}\sum_{s=1}^{T} X_{is})^2] > 0$ which means X_{it} should be time-varying.

Notice the term in the second condition is the **within group transformation**

4.1 DEFINITION. (Within group transformation) This is defined for horizontal observation which is for the same household at different time

$$X_{it}^* = X_{it} - \frac{1}{T}\sum_{s=1}^{T} X_{is} = X_{it} - \overline{X}_{iT} \tag{31}$$

actually this transformation center the Then we identify the β by doing within group transformation for Y_{it}

$$\frac{1}{T}\sum_{t=1}^{T} Y_{it} = \left(\frac{1}{T}\sum_{t=1}^{T} X_{it}\right)\beta + \alpha_i + \frac{1}{T}\sum_{t=1}^{T} u_{it} \tag{32}$$

then (30) – (32) we get

$$Y_{it}^* = X_{it}^*\beta + u_{it}^* \tag{33}$$

Notice the within group transformation moves away the fixed effect. Then multiply X_{it}^* on both sides then take expectation

$$\mathbb{E}[Y_{it}^* X_{it}^*] = \beta\,\mathbb{E}[X_{it}^*]^2 + \underbrace{\mathbb{E}[u_{it}^* X_{it}^*]}_{=0}$$

then next step we want to use as much data available as possible so

$$\sum_{t=1}^{T}\mathbb{E}[Y_{it}^* X_{it}^*] = \beta\sum_{t=1}^{T}\mathbb{E}[X_{it}^*]^2$$

so finally

$$\beta = \frac{\sum_{t=1}^{T}\mathbb{E}[Y_{it}^* X_{it}^*]}{\sum_{t=1}^{T}\mathbb{E}[X_{it}^*]^2} \tag{34}$$

The correlation between We are not assuming the distribution of X_{it} are the same across all t.

Then if we try the sample analogue estimator we have

$$\begin{aligned}\widehat{\beta} &= \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} Y_{it}^* X_{it}^*}{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X_{it}^*\right)^2} \\ &= \beta + \frac{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T} u_{it}^* X_{it}^*}{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X_{it}^*\right)^2}\end{aligned} \tag{35}$$

Notice from the first line we can think of $Y^*_{it}X^*_{it}$ as a function or transformation of X^*_{it} so they are still i.i.d and then the estimator is consistent. The asymptotic theory is apply in a sense that time t is fixed and n goes to infinity. Very important. To make the above discussion more clear, lets take the numerator as an example.

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X^*_{it}\right)^2$$

Our objective dataset is a short panel which means the time T is not going to be large. So for each i, household, the gray part is fixed given a T. What goes to infinity for asymptotic analysis is the n, number of household. so

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X^*_{it}\right)^2 \longrightarrow_p \mathbb{E}[\sum_{t=1}^{T}(X^*_{it})^2]$$

Similar for the numerator. Then we can derive the asymptotic distribution

$$\sqrt{n}(\widehat{\beta}-\beta) = \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}u^*_{it}X^*_{it}}{\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{T}\left(X^*_{it}\right)^2}$$

the numerator

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{t=1}^{T}u^*_{it}X^*_{it} \longrightarrow_d N(0, v^2)$$

where

$$v^2 = \frac{\mathrm{Var}(\sum_{t=1}^{T}u^*_{it}X_{it})}{\left(\mathbb{E}\left[\sum_{t=1}^{I}\left(x^*_{it}\right)^2\right]\right)^2} = \frac{\mathbb{E}[\sum_{t=1}^{T}(u^*_{it}X_{it})^2]}{\left(\mathbb{E}\left[\sum_{t=1}^{I}\left(x^*_{it}\right)^2\right]\right)^2}$$

the numerator converges to constant. Then we introduce some more constrain to simplify the expression. Analogue to linear model

- $\mathbb{E}[u^2_{it} \mid X_{i1}, ..., X_{iT}] = \sigma^2$
- $\mathbb{E}[u_{it}u_{is} \mid X_{i1}, ..., X_{iT}] = 0, \ \forall t \neq s$

We can logically knows, even without proof, that the transformed version of the above two assumption also works which is

- $\mathbb{E}[(u^*_{it})^2 \mid X^*_{i1}, ..., X^*_{iT}] = \sigma^2$
- $\mathbb{E}[u^*_{it}u^*_{is} \mid X^*_{i1}, ..., X^*_{iT}] = 0, \ \forall t \neq s$

Then we can simplify the numerator

$$\mathbb{E}\left(\left(\sum_{t=1}^{T} u_{it}^{*} X_{it}^{*}\right)^{2}\right) = \sum_{s=1}^{T}\sum_{t=1}^{T} u_{it}^{*} u_{is}^{*} X_{it}^{*} X_{is}^{*}$$

$$= \sum_{s=1}^{T}\sum_{t=1}^{T} \mathbb{E}\left(\mathbb{E}\left(u_{it}^{*} u_{is}^{*} \mid X_{i1}, X_{i2} \ldots . X_{it}\right) \cdot X_{it}^{*} X_{is}^{*}\right)$$

$$= \sum_{t=1}^{T} \mathbb{E}(\underbrace{\mathbb{E}\left(u_{it}^{+2} \mid X_{i1}, X_{i2} \ldots, X_{it}\right)}_{=\sigma^2}) X_{it}^{2}$$

$$+ \sum_{t\neq s}\sum_{s=1}^{T} \mathbb{E}\underbrace{\left(\mathbb{E}\left(u_{it}^{*} u_{is}^{*} \mid X_{i1}, \cdots, X_{it}\right) X_{it}^{*} X_{is}^{*}\right)}_{=0}$$

$$= \sigma^{2} \sum_{i=1}^{I}\left[\mathbb{E}\left[\left(X_{it}^{*}\right)^{2}\right]\right.$$

so finally

$$v^{2} = \frac{\sigma^{2}}{\sum_{t=1}^{T} \mathbb{E}[(X_{it}^{*})^{2}]}$$

4.2 *Fixed Effect Multiple Regression*

Add one more variable

$$Y_{it} = \beta_1 X_{it} + \beta_2 W_{it} + \alpha_i + u_{it}$$

- $\mathbb{E}[u_{it} \mid X_{i1}, X_{i2}, ..., X_{iT}, W_{i1}, W_{i2}, ..., W_{iT}] = 0$ for $\forall i, T$
- $\sum_{t=1}^{T} \mathbb{E}[(X_{it}^{*})^{2}] > 0$, $\sum_{t=1}^{T} \mathbb{E}[(W_{it}^{*})^{2}] > 0$ there is no multicolinerality

then the way to identify is

$$\sum_{t=1}^{T} \mathbb{E}\left(Y_{it}^{*} X_{it}^{*}\right) = \beta_1 \sum_{t=1}^{T} \mathbb{E}\left[\left(X_{it}^{*}\right)^{2}\right] + \beta_2 \sum_{t=1}^{T} \mathbb{E}\left(X_{it}^{*} W_{it}^{*}\right)$$

$$\sum_{t=1}^{T} \mathbb{E}\left(Y_{it}^{*} W_{it}^{*}\right) = \beta_1 \sum_{t=1}^{T} \mathbb{E}\left[\left(W_{it}^{*} X_{it}^{*}\right)\right] + \beta_2 \sum_{t=1}^{T} \mathbb{E}\left(W_{it}^{*}\right)^{2}$$

So the meaning of the second condition is to

4.3 *First differencing*

This is another way to transform the data. This is differencing between two consecutive time period which is

$$\Delta Y_{it} = \beta \Delta X_{it} + \Delta \alpha_i + \Delta u_{it}$$

where the delta means $\cdot_t - \cdot_{t-1}$ for $t = 1, 2, ..., T$. Then identify β same as before

$$\beta_1 = \frac{\sum_{t=1}^{T} \mathbb{E}[\Delta Y_{it} \Delta X_{it}]}{\sum_{t=1}^{T} \mathbb{E}[\Delta X_{it}^2]}$$

and sample analogue counterpart. The assumptions

- $E(\Delta u_{it} \mid x_{i1}, \ldots, X_{iT}) = 0$
- $\sum_{t=2}^{T} E(\Delta x_{it})^2 > 0$

Notice the first condition allows the $\mathbb{E}(u_{it} \mid X_{i1}, ..., X_{iT}) = v_i$. Then with these assumptions the first differencing is actually equivalent to within group transformation but DIFFERENT variance. Then

$$\sqrt{n}\left(\hat{\beta}_{FD} - \beta\right) = \frac{\frac{1}{\sqrt{n}} \sum_{n=1}^{n} \sum_{t=2}^{T} \Delta u_{it} \Delta X_{it}}{\frac{1}{n} \sum_{i=1}^{n} \sum_{t=2}^{T} (\Delta X_{it})^2}$$

and the denominator by WLLN, the numerator by CLT get

$$v^2 = \frac{\text{Var}\left(\sum_{t=2}^{T} \Delta X_{it} \Delta U_{it}\right)}{\left(\sum_{t=2}^{T} E\left((\Delta X_{it})^2\right)\right)^2}$$

read the note about Endogenous Regressor

4.4 *Fixed Effect with Lagged Dependent Variables (Dynamic Panel Model)*

The model is divided into observable and observable part

$$Y_{it} = X_{it}\beta + \epsilon_{it}$$

where $\epsilon_{it} = \alpha_i + u_{it}$. Assumption

$$\begin{aligned} &\mathbb{E}(u_{it} \mid X_{i1}, ..., X_{iT}, \alpha_i) \\ &\mathbb{E}(u_{it} u_{is} \mid X_{i1}, ..., X_{iT}, \alpha_i) = 0 \quad \forall t \neq s \end{aligned} \tag{36}$$

then these two gives

- $\mathbb{E}(\epsilon_{it}\epsilon_{is} \mid \mathrm{X}_{i1}, ..., \mathrm{X}_{i\mathrm{T}}, \alpha_i) = \alpha_i^2$

The lagged is actually defined to be

$$\mathrm{Y}_{it} = \mathrm{X}_{it}\beta + \epsilon_{it}, \;\; \epsilon_{it} = \beta_2 \mathrm{Y}_{it-1} + \alpha + u_{it}$$

that is we assume the serial correlation comes from two sources, the **state dependence** and **unobserved heterogeneity** ($i.e\alpha_i$). Then our model becomes

$$\mathrm{Y}_{it} = \beta_1 \mathrm{X}_{it} + \beta_2 \mathrm{Y}_{it-1} + \alpha_i + u_{it}$$

$$\mathrm{Y}^*_{it} = \beta_1 \mathrm{X}^*_{it} + \beta_2 \mathrm{Y}^*_{it-1} + \alpha^*_i + u^*_{it}$$

We still can have the assumption of $\mathbb{E}(u_{it} \mid \mathrm{X}_{i0}, ..., \mathrm{X}_{i\mathrm{T}})$. However if we still what to identify β as before we have to have validity of

$$\mathbb{E}(u^*_{it}\mathrm{Y}^*_{it-1}) = 0$$

this is not valid actually since u_{it} involves all those $u_{i1}, .., u_{i\mathrm{T}}$ and Y^*_{it-1} involves all those $u_{i1}, ..., u_{it-1}$. So overlapping makes within group transformation NOT WORKS. By the same idea, first differencing also fails due to overlapping. Thus we need instrument variable here.

The idea is considering the two sources of serial correlation, we want to set up a model to include these two cases at the same time. Still consider FD. This will wipe out the α_i from the first place (i.e FD removes all time invariant variables). So far our equation becomes

$$\Delta \mathrm{Y}_{it} = \beta_1 \Delta \mathrm{X}_{it} + \beta_2 \Delta \mathrm{Y}_{it-1} + \Delta u_{it}$$

so it turns out the lagged term is endogeneous: First we still assume (36) while add the lagged term in to the condition. So we want to check $\mathrm{Cov}(\mathrm{Y}_{\text{it-1}}, \mathrm{u}_{\mathrm{it}})$ is 0 or not. This is

$$\mathbb{E}(\Delta u_{it}, \Delta \mathrm{Y}_{it-1}) = \mathbb{E}\left(u_{it}\mathrm{Y}_{it-1}\right) - \mathbb{E}\left(u_{it}\mathrm{Y}_{it-2}\right) - \mathbb{E}\left(u_{it-1}\mathrm{Y}_{it-1}\right) + \mathbb{E}((u_{it-1}\mathrm{Y}_{it-2})$$

notice all terms are the production of current error term u_{it} and past $\mathrm{Y}_{i\mathrm{T}}$, $\mathrm{T} < t$. We can show this is zero:

$$\begin{aligned}
\mathrm{E}\left(u_{it} \cdot \mathrm{Y}_{it-1}\right) &= \mathrm{E}\left(u_{it}\left(\beta_1 x_{it} + \beta_2 \mathrm{Y}_{it-2} + \alpha_i + u_{it-2}\right)\right) \\
&= \beta_1 \underbrace{\mathrm{E}\left(u_{it}x_{it}\right)}_{=0} + \beta_2 \mathrm{E}(u_{it}\mathrm{Y}_{it-2}) + \underbrace{\mathrm{E}\left(u_{it}u_{it-2}\right)}_{=0} \\
&= \beta_2 \mathrm{E}\left(u_{it}\mathrm{Y}_{it-2}\right) \\
&= \cdots\cdots = \mathrm{E}\left(u_{it}\mathrm{Y}_{i0}\right) = 0
\end{aligned}$$

so we have

$$\mathbb{E}(\Delta u_{it}, \Delta Y_{it-1}) = \mathbb{E}\left(u_{it-1} Y_{it-1}\right)$$

Thus we need instrument variable to remove the exgeneounity. Fortunately, for the dynamic model we use, there exists natural IV so no need to search for additional resources. This can be any lagged Y. Try

$$\text{Cov}(\Delta u_{it}, Y_{it-2})$$

By similar idea above, this is zero and then we can identify the parameter by

$$\text{cov}\left(\Delta Y_{it}, \Delta X_{it}\right) = \beta_1 \text{ var}\left(\Delta X_{it}\right) + \beta_2 \text{ cov}\left(\Delta Y_{it-1}, \Delta X_{it}\right)$$

$$\text{cov}\left(\Delta Y_{it}, Y_{it-2}\right) = \beta_1 \text{ cov}\left(\Delta X_{it}, Y_{it-2}\right) + \beta_2 \text{ cov}(\Delta Y_{it-1}, Y_{i-2})$$

Notice that expectation also works here while covariance also works. Other candidates of IV can be $Y_{i,t-2}, \Delta Y_{i,t-2}, Y_{i,t-3}, \Delta Y_{i,t-3}$... as long as it does not involves the u_{it-1}. Always remember the two condition that IV should satisfy: relevancy and exogenous. When we say good quality IV, we are saying the stronger correlation with regressor. So we can rank all those IV by its covariance with the endogenous regressor. In this case, the one that more close in time has stronger correlation.

5 BINARY CHOICE MODEL

Exactly a classification model. Binary classification. This is

$$Y_i = \mathbf{1}\{\beta_0 + \beta_1 X_i \geq u_i\} = \begin{cases} 1, & \text{if } \beta_0 + \beta_1 X_i \geq u_i \\ 0, & \text{otherwise} \end{cases}$$

This model is motivated by *random utility model* from Mcfadden.

Model Setup

Choice is binary either 1 or 0 (i.e buy or not to buy, yes or no). We assume the utility of chose 1 and 0 respectively as

$$\textbf{Option 1} : U_i(1) = \gamma_{01} + \gamma_{11} X_i + v_{i1}$$

$$\textbf{Option 0} : U_i(0) = \gamma_{00} + \gamma_{10} X_i + v_{i0}$$

Individual will chose option 1 if its utility is greater than option 0. So we take difference of the two utility and get

$$\begin{aligned} Y_i &= \mathbf{1}\{\gamma_{01} - \gamma_{00} + (\gamma_{11} - \gamma_{01}) X_i \geq v_{i1} - v_{i0}\} \\ &= \mathbf{1}\{\beta_0 + \beta_1 X_i \geq u_i\} \end{aligned}$$

In research, we usually start with strong assumptions and then see if we can relax it when doing identification.

Assumptions

- *Conditional Distribution of u_i given* X_i Assume

$$f(u_i \mid X_i) = N(0, 1)$$

this actually implies the independence how? between X_i and u_i. The uncorrelation is not very useful in non-linear model. So we need this Independence. This assumption actually packed up three assumption together, one further by another

  - u_i and X_i are independent
  - $u_i \sim N(\mu, \sigma^2)$
  - $\mu = 0, \sigma = 1$ otherwie not identifiable.

- $\mathrm{Var}(X_i) > 0$

Identification

We start with the conditional choice probability (CCP).

$$\begin{aligned}
\mathbb{P}(Y_i = 1 \mid X_i = x) &= \mathbb{P}(\beta_0 + \beta_1 X_i \geq u_i \mid X_i = x) \\
&= \mathbb{P}(\beta_0 + \beta_1 x \geq u_i \mid X_i = x) \\
&= \mathbb{P}(u_i \leq t \mid X_i = x) \quad \text{where } t = \beta_0 + \beta_1 x \text{ \& independence} \\
&= \Phi(t) = \Phi(\beta_0 + \beta_1 X_i) \text{ Recall the first assumption}
\end{aligned}$$

Then since CDF for standard normal is strictly increasing, so bijection and so

$$\Phi^{-1}\left\{\mathbb{P}(Y_i = 1 \mid X_i = x)\right\} = \beta_0 + \beta_1 x$$

and so does the random version

$$\Phi^{-1}\left\{\mathbb{P}(Y_i = 1 \mid X_i)\right\} = \beta_0 + \beta_1 X_i$$

then we can identify by

$$\mathrm{Cov}(\Phi^{-1}\left\{\mathbb{P}(Y_i = 1 \mid X_i)\right\}, X_i) = \beta_1 \mathrm{Var}(X_i)$$

and

$$\beta_1 = \frac{\mathrm{Cov}(\phi^{-1}\left\{\mathbb{P}(Y_i = 1 \mid X_i)\right\}, X_i)}{\mathrm{Var}(X_i)}$$

This result is for identification and in this sense is valid. The probability should not be worried in identification since we are assuming, in identification that we know the population Y and X.

In order to interpret the coefficient, we consider the average partial effect (APE). The APE is actually talking about the regressor like how unit of X increment causes the Y to change. The coefficients can be just part of this effect when the model is not linear. In a more general way we consider

$$Y = m(X, U) = \mathbb{E}(Y \mid X, U)$$

as model with no error term which is already the expectation. Then the APE of U at $X = x$ fixed is

$$\mathbb{E}_{U|X}\Big[\frac{\partial m(x, U)}{\partial x}\Big|X = x\Big]$$

So in our case here our m is

$$m(X) = \mathbb{E}(Y_i \mid X_i = x) = \mathbb{P}(Y_i = 1 \mid X_i = x) = p(x)$$

the last equality is due to the Bernoulli distribution of Y_i.Then the APE of X is

$$\text{APE}_X = \mathbb{E}\Big(p'(X)_i\Big) = \mathbb{E}\Big(\phi(\beta_0 + \beta_1 X_i)\beta_1\Big)$$

where we use small phi for density. If we know both beta then the sample analogue estimator works. Or if we have consistent estimator also works which is

$$\widehat{\text{APE}} = \frac{1}{n}\sum_{i=1}^{n} \phi(\widehat{\beta_0} + \widehat{\beta_1} X_i)\widehat{\beta_1}$$

Estimation

We use MLE in the model. Familiar idea just mention the notation. Recall likelihood function is about the parameter so

$$\mathcal{L}(\theta \mid \mathbf{X})$$

is the likelihood. The MLE estimator is

$$\widehat{\theta}_{\text{MLE}} = \text{augmax}_{\theta\in\Theta}\, \mathcal{L}(\theta \mid \mathbf{X})$$

also the equivalent log likelihood

$$\widehat{\theta}_{\text{MLE}} = \text{augmax}_{\theta\in\Theta} \log \mathcal{L}(\theta \mid \mathbf{X})$$

Lecture gives the detailed proof of MLE of the linear regression model which finally shows the equivalence result of LSE and sample analogue. Makeup Notice the assumption of $u_i \sim N(0, \sigma^2)$. MLE here is applied as

$$\mathbb{P}\{Y_i = 1 \mid X_i = x\} = \Phi(\beta_0 + \beta_1 x)$$

$$\mathbb{P}\{Y_i = 0 \mid X_i = x\} = 1 - \Phi(\beta_0 + \beta_1 x)$$

together we have

$$\mathbb{P}\{Y_i = y \mid X_i = x\} = \Phi(\beta_0 + \beta_1 x)^y (1 - \Phi(\beta_0 + \beta_1 x))^{1-y}$$

then our likelihood becomes

$$\begin{aligned}
&\mathcal{L}(b_0, b_1; \mathrm{Y}_1, .., \mathrm{Y}_n \mid \mathrm{X}_i, ..., \mathrm{X}_n) \\
&= \sum_{i=1}^{n} \log p(\mathrm{Y}_i \mid \mathrm{X}_i; b_0, b_1) \\
&= \sum_{i=1}^{n} \Big\{ (\mathrm{Y}_i \log \Phi(b_0 + b_1 \mathrm{X}_i)) + (1 - \mathrm{Y}_i) \log(1 - \Phi(b_0 + b_1 \mathrm{X}_i)) \Big\}
\end{aligned}$$

We can analysis all those properties of MLE estimator while not covered in this course. The resulting function looks like the logistic regression.

Extended to Multiple Variable

The model simply becomes

$$\mathrm{Y}_i = \mathbf{1}\{\beta_0 + \beta_1 \mathrm{X}_i + \beta_2 \mathrm{W}_i \geq u_i\}$$

with assumption $u_i \mid \mathrm{X}_1, \mathrm{W}_i \sim \mathrm{N}(0, 1)$. This also composed three assumption as before. Then the CCP

$$\begin{aligned}
\mathbb{P}\{\mathrm{Y}_i = 1 \mid \mathrm{X}_i = x, \mathrm{W}_i = w\} &= \mathbb{P}\{\beta_0 + \beta_1 \mathrm{X}_i + \beta_2 \mathrm{W}_i \geq u_i \mid \mathrm{X}_i = x, \mathrm{W}_i = w\} \\
&= \mathbb{P}\{\beta_0 + \beta_1 x + \beta_2 w \geq u_i\} \text{See notes by independence} \\
&= \Phi(\beta_0 + \beta_1 \mathrm{X}_i + \beta_2 \mathrm{W}_i)
\end{aligned}$$

The average partial effect. The APE in binary choice model is actually about the objective function which is the probability of the Y to be 1. Analogously the objective function in linear model is the Y itself. Identification is the same. Omitted.

6 CASUAL INFERENCE

We focus on **Potential Outcome Approach**. Motivated by medical study. No need to specify the specific formulae or equation for this. Let consider the experiments. We have two potential outcomes

$$(\mathrm{Y}_{i1}, \mathrm{Y}_{i0})$$

the first state 1 is the potential outcome under treated state and the state 0 is the state under control. So these generate 2 potential outcome. Let $\mathrm{D}_i \in \{0, 1\}$ which indicates whether the sample is treated or under control. Treatment effect is thus

$$\mathrm{Y}_{i1} - \mathrm{Y}_{i0}$$

so we are comparing the same person. The treatment effect is heterogeneous which means the effect is different across every person. The most important

quantity we want to recover from data is the *average treatment effect* which is

$$\text{ATE} = \mathbb{E}[Y_{i1} - Y_{i0}]$$

and average treatment effect under treated

$$\text{ATT} = \mathbb{E}[Y_{i1} - Y_{i0} \mid D_i = 1]$$

like a subgroup of people both under control and treatment. Usually, we can only observe either Y_{i1} or Y_{i0} since usually a person can either be treated or control. So what we observe is

$$Y_i = D_i Y_{i1} + (1 - D_i) Y_{i0}$$

so the table will have missing values. The thing we can observe are Y_i, D_i, X_i where X_i can be demographic information about participants. Let's assume that (Y_i, D_i, X_i) are iid across people however this is still not enough to pin down the ATE. For D_i the decision depends on each person them-self which may depends on different personality. This is more like the bias but not about the iid distribution.

Our real assumptions :

- **Randomized Control Trial**: (Y_{i1}, Y_{i0}, X_i) are independent from D_i

Then we can identify the ATE:

$$\begin{aligned}
\text{ATE} &= E(Y_{i1} - Y_{i0}) \\
&= E(Y_{i1}) - E(Y_{i0}) \\
&= E(Y_{i1} \mid D_i = 1) - E(Y_{i0} \mid D_i = 0) \\
&= E(Y_i \mid D_i = 1) - E(Y_i \mid D_i = 0)
\end{aligned}$$

the equities are all due to independence. Then we can write

$$= \frac{\mathbb{E}(Y_i D_i)}{\mathbb{P}(D_i = 1)} - \frac{\mathbb{E}(Y_i(1 - D_i))}{\mathbb{P}(D_i = 0)}$$

and do sample analogue estimator.