

Notes on Econometrics

Victor Li

Autumn Sememster, 2023

Contents

1	preparation	3
2	linear regression	3
2.1	regression	3
2.2	GM theorem and BLUE	4
2.3	R-squared	4
3	hypothesis and test	6
3.1	t-test	6
3.2	confidence interval test	6
4	multi-variate linear regression	6
5	other regression stuff	8
6	Panel data	9
7	Causal effect	10
7.1	IV	10
7.2	DID	11
7.3	RDD	11
8	Time series data	11

1 preparation

one should always be able to differ

- causal relationship
- correlative relationship

data types

- cross-sectional
- time series
- panel

skewness

skewness (standardized 3rd central moment)

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

If $\begin{cases} \approx 0 & \iff \text{almost like normal distribution} \\ > 0 & \iff \text{right side lower, meaning more outliers ar right side} \\ < 0 & \iff \text{left side lower, meaning more outliers ar left side} \end{cases}$

kurtois

kurtosis or peakness (standardized 4rd central moment of samples)

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

If $\begin{cases} \approx 0 & \iff \text{almost like normal distribution} \\ > 0 & \iff \text{higher than} \\ < 0 & \iff \text{lower than} \end{cases}$

2 linear regression

2.1 regression

sample

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X \text{ (regression equation)} \quad (3)$$

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \underbrace{e}_{\text{error}} = \hat{Y} + e \text{ (regression model)} \quad (4)$$

population

$$E(Y|X) = \beta_0 + \beta_1 X = E(Y|X) \text{ (regression equation)} \quad (5)$$

$$Y = \beta_0 + \beta_1 X + \underbrace{\mu}_{\text{disturbance}} = E(Y|X) + \mu \text{ (regression model)} \quad (6)$$

OLS

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

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

$$= \min \sum_{i=1}^n e_i^2 \quad (8)$$

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

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

$$\begin{cases} \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = 1, 2 \dots n \\ \hat{e}_i = Y_i - \hat{Y}_i, i = 1, 2 \dots n \end{cases} \quad (10)$$

moments of estimated coefficients

$$(11)$$

2.2 GM theorem and BLUE

GM theorem:

Under CLRM conditions, OLS method gives estimated coefficients satisfying BLUE.

BLUE:

$$\begin{cases} \text{unbiasedness: } E(\hat{\beta}) = \beta \\ \text{efficiency: } \min \text{var}(\hat{\beta}) \\ \text{(consistency in large samples): } \hat{\beta} \xrightarrow{P} \beta \end{cases}$$

2.3 R-squared

R^2 measures goodness of fit, which is how good the estimated coefficient fit the real world data. It is considered an indicator to judge a model.

Note 1 (degree of freedom).

degree of freedom is actually $\begin{cases} TSS : n - 1 \\ ESS : k \\ SSR : n - k - 1 \end{cases}$ n- Denominators below are actually degree of freedom. Only in large samples approximated by n .

For Y ,

$$\frac{TSS}{n} = var(Y) \quad (12)$$

$$TSS = n \cdot var(Y) = n \cdot \frac{\sum_i^n (Y_i - \bar{Y})^2}{n} = \sum_i^n (Y_i - \bar{Y})^2 \quad (13)$$

$$SE(Y) = \sqrt{var(Y)} = \sqrt{\frac{TSS}{n}} = \sqrt{\frac{\sum_i^n (Y_i - \bar{Y})^2}{n}} \quad (14)$$

For \hat{Y} ,

$$\frac{ESS}{n} = var(\hat{Y}) \quad (15)$$

$$ESS = \sum_i^n (\hat{Y}_i - \bar{Y})^2 \quad (16)$$

$$SE(\hat{Y}) = \sqrt{var(\hat{Y})} = \sqrt{\frac{ESS}{n}} = \sqrt{\frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{n}} \quad (17)$$

For e ,

$$\frac{SSR}{n} = var(e) \quad (18)$$

$$SSR = \sum_i^n (e_i - \bar{e})^2 \quad (19)$$

$$SER = SE(e) = \sqrt{var(e)} = \sqrt{\frac{SSR}{n}} = \sqrt{\frac{\sum_i^n (e_i - \bar{e})^2}{n}} \quad (\text{also the SE of the regression}) \quad (20)$$

Now the R-squared

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{\sum_i^n (Y_i - \bar{Y})^2} \quad (21)$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = \frac{\text{explained by the estimated model}}{\text{total sample data}} \quad (22)$$

Note 2 (SER, standard error of regression). *an indicator measuring the deviation of error term, not the deviation of whole model.*

3 hypothesis and test

t-value

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\text{estimation} - \text{hypothesis}}{\text{standard error of estimation}} \quad (23)$$

p-value

$$p = 2\Phi(-|t|) \quad (24)$$

3.1 t-test

step 1: give a hypothesis

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

step 2: calculate t-value based on hypothesis

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)}$$

step 3: calculate p-value

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

or another step 3: judge by experience

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

3.2 confidence interval test

step 1: give a hypothesis

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

step 2: calculate t-value

step 3: choose significance level

small sample: based on significance level α , degree of freedom $df = n - 2$, two-tale or one-tale $\Rightarrow t_{\frac{\alpha}{2}} =$

large sample: based on significance level α , degree of freedom $df = n$, two-tale or one-tale $\Rightarrow t_{\frac{\alpha}{2}} =$

step 4: calculate CI

$$\hat{\beta} - t_{\frac{\alpha}{2}} \cdot se(\hat{\beta}) \leq \beta \leq +t_{\frac{\alpha}{2}} \cdot se(\hat{\beta})$$

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 (kurtosis)

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

$$\hat{\beta} \xrightarrow{P} \beta + \underbrace{\frac{\sigma_u}{\sigma_X} \rho_{uX}}_{\text{effect of OVB}} \quad (25)$$

meaning OVB causes estimator to be biased and inconsistent

How to overcome OVB?

- More control variables
- IV
- Panel Fixed Effect model

Adjusted R-squared

$$\bar{R}^2 = 1 - \frac{RSS/n - k - 1}{TSS/n - 1} = 1 - \frac{n - 1}{n - k - 1} \frac{RSS}{TSS} \quad (26)$$

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:

$$\min_{\{\beta_0, \dots, \beta_k\}} \sum_i^n (Y_i - \hat{Y}_i)^2 \quad (27)$$

or in matrix form

$$\min (Y - X\hat{\beta})^2 \quad (28)$$

results

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (29)$$

Joint hypothesis test

$$\begin{cases} H_0 : \beta_1 = 0 \& \beta_2 = 0 \\ H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or both } \neq 0 \end{cases} \quad (30)$$

F-test

$$F = \frac{1}{2} \left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1, t_2} t_1 t_2}{1 - \hat{\rho}_{t_1, t_2}^2} \right), \text{ where } \hat{\rho}_{t_1, t_2}^2 \text{ is estimated correlative coefficient} \quad (31)$$

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

$$F = \frac{1}{2}(t_1^2 + t_2^2) \quad (32)$$

simplified F statistics when homoskedasticity

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

q = number of constraints

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

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

$$F = \frac{(R_{\text{unrestricted}}^2 - R_{\text{restricted}}^2)/q}{(1 - R_{\text{unrestricted}}^2)/(n - k_{\text{unrestricted}} - 1)} \quad (33)$$

Tests for Single Constraints Involving Multiple Coefficients

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

to

$$\begin{aligned} Y &= \beta_0 + (\beta_1 - \beta_2)X + \beta_2(X + Y) + u \\ &= \beta_0 + \gamma X + \beta_2 W + u \end{aligned} \quad (34)$$

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 multicollinearity

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

non-linear regression

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

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

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

multicollinearity

interaction term

1) two dummies (DID)

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

$$\begin{aligned} \text{effect of } D_2 &= E(Y|D_1, D_2 = 1) - E(Y|D_1, D_2 = 0) \\ &= (\beta_0 + \beta_1 D_1 + \beta_2 + \beta_3 D_1) - (\beta_0 + \beta_1 D_1) \\ &= \beta_2 + \beta_3 D_1 \end{aligned} \quad (35)$$

2) dummy and continuous variable

3) two continuous variables

6 Panel data

fixed effect

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

individual fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_i + \mu_{it} \quad (36)$$

time fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_t + \mu_{it} \quad (37)$$

individual and time fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_i + \beta_3 D_t + \mu_{it} \quad (38)$$

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

random effect

random effect is used when $\text{corr}(X, u) = 0$

Hausman test

used to decide whether to use fixed effect or random 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

$$\begin{aligned} \tau_{att} &= E(\delta_i | D_i = 1) \\ &= E(Y_i^1 - Y_i^0 | D_i = 1) \\ &= E(Y_i^1 | D_i = 1) - E(Y_i^0 | D_i = 1) \end{aligned} \quad (39)$$

ATU

$$\begin{aligned} \tau_{atu} &= E(\delta_i | D_i = 0) \\ &= E(Y_i^1 - Y_i^0 | D_i = 0) \\ &= E(Y_i^1 | D_i = 0) - E(Y_i^0 | D_i = 0) \end{aligned} \quad (40)$$

ATE

$$\begin{aligned} \tau_{ate} &= E(\delta_i) \\ &= E(Y_i^1 - Y_i^0) \\ &= E[E(Y_i^1 - Y_i^0 | D_i)] \\ &= E(Y_i^1 - Y_i^0 | D_i = 1) \cdot Pr(D_i = 1) + E(Y_i^1 - Y_i^0 | D_i = 0) \cdot Pr(D_i = 0) \\ &= \tau_{att} \cdot Pr(D_i = 1) + \tau_{atu} \cdot Pr(D_i = 0) \end{aligned} \quad (41)$$

7.1 IV

IV conditions

$$corr(Z, \mu) = 0 \quad (42)$$

$$corr(Z, X) \neq 0 \quad (43)$$

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

$$y_{it} = \beta x_{it} + \gamma_1 D_i + \gamma_2 D_t + \mu_{it} \quad (44)$$

7.3 RDD

$$Y_i = \alpha + \beta D_i + f(X_i) + \mu_i \quad (45)$$

- 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 $X_i \geq c$

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

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

8 Time series data

auto regression

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

Stationary

for $\{X_1, \dots, X_t, \dots\}$ 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 multiple stationary time series