Notes on Econometrics

Victor Li

Spring Sememster, 2024

Contents

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)

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

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

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

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

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

2 linear regression

2.1 regression

sample

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

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

population

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

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

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

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

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

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

so OLS is basically an optimization problem

FOC: $\begin{cases} \frac{\partial \min}{\partial \hat{\beta}_0} = 0 \\ \frac{\partial \min}{\partial \hat{\beta}_1} = 0 \end{cases}$ $\Rightarrow$ 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 $\hat{e}_i$

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

moments of estimated coefficients

(11)

2.2 GM theorem and BLUE

GM theorem:
Under CLRM conditions, OLS method gives estimated coefficients satisfying BLUE.
BLUE:
$\begin{cases} \text{unbiasedness: } E(\hat{\beta}) = \beta \\ \text{efficiency: } \min var(\hat{\beta}) \\ \text{(consistency in large samples): } \hat{\beta} \xrightarrow{P} \beta \end{cases}$

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.

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

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

$$\frac{ESS}{n} = var(\hat{Y}) \tag{14}$$

$$ESS = \sum_i^n (\hat{Y}_i - \bar{Y})^2 \tag{15}$$

$$\frac{SSR}{n} = var(e) \tag{16}$$

$$SSR = \sum_i^n (e_i - \bar{e})^2 \tag{17}$$

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_i^n (\hat{Y}_i - \bar{Y})^2}{\sum_i^n (Y_i - \bar{Y})^2} \tag{18}$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} = \frac{\text{explained by the estimated model}}{\text{total sample data}} \tag{19}$$

indicators concerning small sample features (substituting strange denominator with n when dealing with large sample)

$$s_Y^2 = \frac{1}{n-1} TSS = \frac{\sum_i^n (Y_i - \bar{Y})^2}{n-1} \tag{20}$$

$$SER = \sqrt{\frac{SSR}{n-2}} = \sqrt{\frac{SSR}{n-2}} \tag{21}$$

> **Note 1** (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

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\text{estimation} - \text{hypothesis}}{\text{standard error of estimation}} \tag{22}$$

p-value

$$p = 2\Phi(-|t|) \tag{23}$$

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

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

> **Note 2** (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 $\Rightarrow \begin{cases} E(\mu|X) \neq 0 \text{ endogeneity} \\ R^2 \text{ is lower than it should be} \end{cases}$

$$\hat{\beta} \xrightarrow{P} \beta + \frac{\sigma_u}{\sigma_X} \rho_{uX} \tag{24}$$

meaning OVB causes estimator to be biased and inconsistent

How to overcome OVB?

- More control variables
- IV
- Panel Fixed Effect model

Adjusted R-squared

$$\bar{R}^2 = 1 - \frac{RSS/n-k-1}{TSS/n-1} = 1 - \frac{n-1}{n-k-1}\frac{RSS}{TSS} \tag{25}$$

> **Note 3** (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:

$$\min_{\{\beta_0,\dots,\beta_k\}} \sum_i^n (Y_i - \hat{Y}_i)^2 \tag{26}$$

or in matrix form

$$\min(Y - X\hat{\beta})^2 \tag{27}$$

results

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{28}$$

degree of freedom is actually $n-k-1$

$$SER = \sqrt{\frac{RSS}{n-k-1}} \tag{29}$$

Joint hypothesis test

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

F-test

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

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

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

simplified F statistics when homoskedasticity
$H_0 : \beta_1 = 0, \beta_2 = 0, \dots, \beta_q = 0$ and $H_1 : \dots$
q = number of tested coefficients

unrestricted regression: $Y = Y(X) + Y(H_0)$
restricted regression: $Y = Y(X)$

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

Tests for Single Constraints Involving Multiple Coefficients
change the original $Y = \beta_0 + \beta_1 X + \beta_2 Y + u$
to

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

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

---

non-linear regression
probit model
logit model
extreme model

---

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

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

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

$$y_{it} = \beta_1 X_{it} + \beta_2 D_i + \mu \tag{36}$$

time fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_t + \mu \tag{37}$$

individual and time fixed effect

$$y_{it} = \beta_1 X_{it} + \beta_2 D_i + \beta_3 G_t + \mu \tag{38}$$

Note 4 (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 IV

IV conditions

$$corr(Z, \mu) = 0 \tag{39}$$
$$corr(Z, X) \neq 0 \tag{40}$$

2SLS
For a $Y = \beta_0 + \beta_1 X + \mu$ and $X = \pi_0 + \pi_1 Z + \nu$
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

8 Causal effect

treatment D
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

$$\begin{aligned} \tau_{att} &= E(\delta_i | D_i = 1) \\ &= E(Y_i^1 - Y_i^0 | D_i = 1) \\ &= E(Y_i^1 | D_i = 1) - E(Y_i^0 | D_i = 1) \end{aligned} \tag{41}$$

ATU

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

ATE

$$\begin{aligned}\tau_{ate} &= E(\delta_i) \\ &= E(Y_i^1 - Y_i^0) \\ &= E[E(Y_i^1 - Y_i^0|D_i)] \\ &= E(Y_i^1 - Y_i^0|D_i = 1) \cdot Pr(D_i = 1) + E(Y_i^1 - Y_i^0|D_i = 0) \cdot Pr(D_i = 0) \\ &= \tau_{att} \cdot Pr(D_i = 1) + \tau_{atu} \cdot Pr(D_i = 0)\end{aligned} \tag{43}$$

Matching
RCT
DID
DDD
SCM

9 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
unit root test
used for testing stationary relationship. if unit root is tested true in the time series, the time series is not stationary.
cointegration test
focus on the long-term stationary relationship between multiple non-stationary time series
Granger test
used for causal test in sationary time series