Econometrics 1. Introduction & Review of Probability and Statistics

Types of Data Sets (P5-6)

- Cross-Sectional Data
  - Data collected at a particular point in time across different economic units.
  - The ordering of the data doesn't matter in these samples and contains no information
- Time Series Data
  - Data collected across time for the same economic unit.
  - eg. { macroeconomic data: inflation, unemployment, GDP
    financial data: interest rates, exchange rates, stock market indices
- Panel Data
  - Observations of a collection of units over several time periods.
  - eg. { Household expenditure and income surveys over several years
    Market shares of firms in different years
    Trade deficits of countries over several years

Review of Probability (Chapter 2)

$$\text{Skewness} = \frac{E[(Y-\mu_Y)^3]}{\sigma_Y^3}$$ how much a distribution deviates from symmetry

- $=0$: symmetric
- <0 : left tail is longer
- >0 : right tail is longer

$$\text{Kurtosis} = \frac{E[(Y-\mu_Y)^4]}{\sigma_Y^4}$$ how much mass is in its tails / how much of the variance arises

- $=3$ normal distribution
  Mesokurtic
- >3 : heavy tails ("leptokurtotic")
  more values in the dist. tails and more values close to the mean (ie. sharply peaked with heavy tails)
- <3 : ("platykurtic") fewer values in the tails & fewer values close to the mean (i.e. flat peak and more dispersed scores with lighter tails)

$X \perp\!\!\!\perp Z \Rightarrow Cov(X,Z) = 0.$ $\quad corr(X,Z) = \frac{cov(X,Z)}{\sqrt{Var(X)\,var(Z)}} = \frac{\sigma_{XZ}}{\sigma_X\sigma_Z} = \gamma_{XZ}$

$W = X_1^2 + \cdots + X_n^2 \sim \chi_n^2$, $EW = n$, $Var\,W = 2n$.

$T = \frac{Z}{\sqrt{W/n}} \sim t_n$. T_∞ is the same as the standard normal dist.

$F = \frac{W_1/n}{W_2/m} \sim F_{n,m}$ $F_{n,\infty}$ is the same as χ_n^2/n dist.

Review of Statistics (Chapter 3)

1. Estimation (unbiasedness, consistency, efficiency)

consistency: $\lim_n P(|\hat{\mu} - \mu| < \varepsilon) = 1 : \hat{\mu} \xrightarrow{P} \mu$.

Chebyshev: $P(|X - EX| \geq a) \leq \frac{Var(X)}{a^2}$.

efficiency: $Var(\hat{\mu})$ is smaller than the variance of all other estimators

$E(\bar{Y}) = \mu_Y$, $Var(\bar{Y}) = \frac{\sigma_Y^2}{n}$. inversely proportional to n.

- the spread of the sampling distribution is proportional to $1/\sqrt{n}$
- the sampling uncertainty associated with $\bar{Y}$ is proportional to $1/\sqrt{n}$.

LLN: $Pr(|\bar{Y} - \mu_Y| < \varepsilon) \to 1$ as $n \to \infty$ $\quad (Var(\bar{Y}) = \frac{\sigma_Y^2}{n} \to 0)$

$\bar{Y}$ is BLUE (proof: SW. Ch. 17)

2. Hypothesis Testing

$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \xrightarrow{P} \sigma_Y^2$

$$p\text{-value} = Pr_{H_0}\left[|\bar{Y} - \mu_{Y,0}| > |\bar{Y}^{act} - \mu_{Y,0}|\right] \cong Pr_{H_0}\left[\left|\frac{\bar{Y} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right| > \left|\frac{\bar{Y}^{act} - \mu_{Y,0}}{s_Y/\sqrt{n}}\right|\right]$$

$$= Pr_{H_0}[|t| > |t^{act}|] \approx 2\Phi(-|t^{act}|)$$

$\begin{cases} p\text{-value} < 5\% \\ |t^{act}| > 1.96. \end{cases}$

right tail: $p\text{-value} = 1 - \Phi(t^{act}) \Leftrightarrow t^{act} > 1.645$ (5%)

left tail: $p\text{-value} = \Phi(t^{act}) \Leftrightarrow t^{act} < -1.645$ (5%)

3. Confidence Intervals.

① (5%): $\bar{Y} \pm 1.96 \cdot \frac{s_Y}{\sqrt{n}}$.

② Compare Means: $H_0: \mu_m - \mu_w = d_0$. $\quad s_m^2 = \frac{1}{n_m - 1}\sum_{i=1}^{n_m}(Y_i - \bar{Y}_m)^2$

$$\begin{cases} \bar{Y}_m - \bar{Y}_w \sim N\left(\mu_m - \mu_w, \frac{\sigma_m^2}{n_m} + \frac{\sigma_w^2}{n_w}\right) \\ t = \frac{\bar{Y}_m - \bar{Y}_w - d_0}{SE(\bar{Y}_m - \bar{Y}_w)}, \text{ where } SE(\bar{Y}_m - \bar{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}} \\ |t| > 1.96 \text{ reject (5\%)} \end{cases}$$

Econometrics 2.1 Introduction to Linear Regression _ OLS Estimation.

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

- Y_i: dependent var. — not itself economically meaningful
- β_0: intercept
- β_1: slope
- X_i: independent var.
- u_i: error (omitted factors or measurement error)

$Y_i = \hat{Y}_i + \hat{u}_i$ = OLS prediction + OLS residual

$\Rightarrow Var(Y) = Var(\hat{Y}_i) + Var(\hat{u}_i)$ (explained SS, residual SS)

SER: standard error of the regression. $SER = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(\hat{u}_i - \bar{\hat{u}})^2} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}$

$(\bar{\hat{u}} = \frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0)$

$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$

$R^2 = 0$: $ESS = 0$

$R^2 = 1$: $ESS = TSS$

$0 \le R^2 \le 1$.

$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i^2}$

Hint: $n-2$ is "degrees of freedom"

The Least Squares Assumptions (Section 4.4)

① $E(u_i | X_i = x) = 0 \Rightarrow \hat{\beta}_1$ is unbiased

"other factors" contained in u_i should be uncorrelated with X_i.

② (X_i, Y_i) are i.i.d. $\Rightarrow$ delivers the sampling distribution of $\hat{\beta}_0$ and $\hat{\beta}_1$.

③ Large outliers in X and/or Y are rare — OLS sensitive to outlier

Outliers can result in meaningless values of $\hat{\beta}_1$.

Technically $E(X^4) < \infty$, $E(Y^4) < \infty$, X, Y are bounded.

The Sampling Distribution of the OLS Estimator (Section 4.5).

$Y_i = \beta_0 + \beta_1 X_i + u_i \Rightarrow Y_i - \bar{Y} = \beta_1 (X_i - \bar{X}) + (u_i - \bar{u})$

$\bar{Y} = \beta_0 + \beta_1 \bar{X} + \bar{u}$

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} = \frac{\sum(X_i - \bar{X})[\beta_1(X_i - \bar{X}) + (u_i - \bar{u})]}{\sum(X_i - \bar{X})^2} = \beta_1 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2} = \beta_1 + \frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2}$$

$$E(\hat{\beta}_1) = \beta_1 + E\left[E\left[\frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2}\right] \middle| X_1, \cdots, X_n\right] = \beta_1. \quad (LSA\ \#1)$$

$$\hat{\beta}_1 - \beta_1 = \frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum v_i}{\frac{1}{n}\sum(X_i - \bar{X})^2}, \text{ where } v_i = (X_i - \bar{X})u_i.$$

If n is large, $\hat{\beta}_1 - \beta_1 \approx \frac{\frac{1}{n}\sum v_i}{\sigma_X^2} \Rightarrow Var(\hat{\beta}_1 - \beta_1) = Var(\hat{\beta}_1) = \frac{Var(v)/n}{(\sigma_X^2)^2} = \frac{1}{n}\cdot\frac{Var[(X_i - \mu_X)u_i]}{\sigma_X^4}$

$Var(\hat{\beta}_1) \propto \frac{1}{n}$, $E(\hat{\beta}_1) = \beta_1 \Rightarrow \hat{\beta}_1 \xrightarrow{p} \beta_1$.

$\{v_i$ i.i.d., $E(v) = 0$, $Var(v) = \sigma^2 \Rightarrow CLT \Rightarrow \frac{1}{n}\sum_{i=1}^{n} v_i \sim \mathcal{N}(0, \sigma_v^2/n)\}$

$v_i = (X_i - \bar{X})u_i$, $E(v_i) = 0$, $Var(v_i) < \infty$

$$\Rightarrow \hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\sigma_v^2}{n\sigma_X^4}\right)$$

Econometrics 2.2 Introduction to Linear Regression – Hypothesis Testing

Test β_1: $t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}$. reject if $|t| > 1.96$ (5%). Requirement: Large-n : $n = 50$

$$Var(\hat{\beta}_1) = \frac{Var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2} = \frac{\sigma_v^2}{n\sigma_X^4} \Rightarrow \sigma_{\hat{\beta}_1}^2 = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2\right]^2} \Rightarrow SE(\hat{\beta}_1) = \sqrt{\sigma_{\hat{\beta}_1}^2}$$

$$\sigma_{\hat{\beta}_1}^2 = \frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum[(X_i - \bar{X})\hat{u}_i]^2}{\left[\frac{1}{n}\sum(X_i - \bar{X})^2\right]^2}$$

Confidence Interval (5%): $\hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1)$

Regression when X is Binary (Section 5.3)

$$Y_i = \beta_0 + \beta_1 X_i + u_i \quad \begin{cases} X_i = 0, & Y_i = \beta_0 + u_i, \quad E(Y_i | X_i = 0) = \beta_0 \\ X_i = 1, & Y_i = \beta_0 + \beta_1 + u_i, \quad E(Y_i | X_i = 1) = \beta_0 + \beta_1 \end{cases}$$

β_1 = population difference in group means

Heteroskedasticity (异方差) & Homoskedasticity (同方差) (Section 5.4)

Var(u|X = x) is constant (does not depend on X)

If Homoskedasticity $\Rightarrow Var(u_i | X_i = x) = \sigma_u^2$

$$\Rightarrow Var(\hat{\beta}_1) = \frac{Var[(X_i - \mu_X)u_i]}{n(\sigma_X^2)^2} = \frac{E[(X_i - \mu_X)^2 u_i^2]}{n(\sigma_X^2)^2} = \frac{\sigma_u^2}{n\sigma_X^2}$$

Advantage: formula is simpler

Worry: too small

$$\Rightarrow SE(\hat{\beta}_1) = \sqrt{\frac{1}{n} \cdot \frac{\frac{1}{n-2}\sum \hat{u}_i^2}{\frac{1}{n}\sum(X_i - \bar{X})^2}}$$

Additional Theoretical Foundations of OLS (Section 5.5)

The extended Least Square Assumptions:

1. $E(u|X = x) = 0$ 2. (X_i, Y_i) i.i.d. 3. Large outliers are rare ($E(Y^4) < \infty$, $E(X^4) < \infty$)

4. u is homoskedastic 5. u is distributed $N(0, \sigma^2)$

① Under Assumption 1-4: $\hat{\beta}_1$ is the BLUE (Gauss–Markov theorem) $\xrightarrow{\text{prove}}$ Appendix 5.2

smallest $Var(\hat{\beta}_1)$. $\hat{\beta}_1 - \beta_1 = \frac{\sum(X_i - \bar{X})u_i}{\sum(X_i - \bar{X})^2} = \frac{1}{n}\sum w_i u_i$, where $w_i = \frac{X_i - \bar{X}}{\frac{1}{n}\sum(X_i - \bar{X})^2}$

② Under Assumption 1-5: $\hat{\beta}_1$ has smallest var of all consistent estimator (linear or nonlinear)

③ not-so-good thing about OLS. —
- G-M condition not hold in practice → (heteroskedasticity)
- ② require normal errors.
- 异方差 ⇒ weighted least squares more efficient

Econometrics 3.1 Introduction to Multiple Regression (Chapter 6)

Omitted Variable Bias (Section 6.1)

u arises because of factors influence Y are not included → omitted variables.

The bias in the OLS estimator ⟶ omitted variable bias.

omitted variable "Z" { A determinant of Y (i.e. part of u) 缺一不可
· Correlated with X (i.e. corr(Z, X) ≠ 0).

$\hat{\beta}_1 - \beta_1 = \frac{\sum (X_i - \bar{X}) u_i}{\sum (X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum v_i}{\frac{1}{n}\sum (X_i - \bar{X})^2}$ ($v_i = (X_i - \bar{X}) u_i \approx (X_i - \mu_X) u_i$).

Under Assumption 1, $E[(X_i - \mu_X) u_i] = \text{cov}(X_i, u_i) = 0$.

what if $E[(X_i - \mu_X) u_i] = \text{cov}(X_i, u_i) = \sigma_{Xu} \neq 0$?

In general, $\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum (X_i - \bar{X}) u_i}{\frac{1}{n}\sum (X_i - \bar{X})^2} \xrightarrow{P} \frac{\sigma_{Xu}}{\sigma_X^2} = \left(\frac{\sigma_u}{\sigma_X}\right) \cdot \left(\frac{\sigma_{Xu}}{\sigma_X \sigma_u}\right) = \left(\frac{\sigma_u}{\sigma_X}\right) \rho_{Xu}$ (↓ 0?)

$\hat{\beta}_1 \xrightarrow{P} \beta_1 + \frac{\sigma_u}{\sigma_X} \rho_{Xu}$.

To overcome: ① Run a randomized controlled experiment
② Add omitted variable → multiple regression

The population multiple regression model (Section 6.2)

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$.

$\beta_1 = \frac{\Delta Y}{\Delta X_1}$ holding X_2 const. $\beta_2 = \frac{\Delta Y}{\Delta X_2}$ holding X_1 const.

β_0 = predicted value of Y when $X_1 = X_2 = 0$

OLS. $\hat{\beta} = (X'X)^{-1} X' Y$

Measures of Fit for Multiple Regression (Section 6.4)

Actual = predicted + residual: $Y_i = \hat{Y}_i + \hat{u}_i$.

{ SER = std. deviation of $\hat{u}_i$ (d.f. correction)
RMSE = std of $\hat{u}_i$ (d.f. NOT correction)
R^2. adj-R^2 : adj-$R^2 = \bar{R}^2 < R^2$.

$SER = \sqrt{\frac{1}{n-k-1}\sum \hat{u}_i^2}$ $RMSE = \sqrt{\frac{1}{n}\sum \hat{u}_i^2}$

$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS}$. $ESS = \sum (\hat{Y}_i - \bar{\hat{Y}})^2$, $SSR = \sum \hat{u}_i^2$, $TSS = \sum (Y_i - \bar{Y})^2$.

$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k-1}\right) \frac{SSR}{TSS}$

The least squares Assumption (Section 6.5)

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$, $i = 1, \cdots, n$.

① $E(u | X_1 = x_1, \cdots, X_k = x_k) = 0$ ② $(X_{1i}, \cdots, X_{ki}, Y_i)$ i.i.d. ③ $E(X_{1i}^4) < \infty, \cdots, E(X_{ki}^4) < \infty, E(Y_i^4) < \infty$
④ No Multicollinearity.

#1. $E(u|X_1=x_1, \cdots, X_k=x_k)=0$
If omitted variable (1) belongs in u (2) correlated with X $\Rightarrow$ condition fails

#2. $(X_{1i}, \cdots, X_{ki}, Y_i)$ i.i.d.

#3. large outliers are rare

#4. No perfect multicollinearity.

The sampling Distribution of the OLS Estimator (SW section 6.6).

$$\hat{\beta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta+U) = \beta + (X'X)^{-1}X'u$$

$$\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0, Q^{-1}\Sigma_v Q^{-1})$$

Multicollinearity (Section 6.7)

The Dummy Variable Trap

Solution ① Omit one of the groups ② Omit the intercept

Perfect multicollinearity : "drop" one

Imperfect multicollinearity → highly correlated → req. coefficients imprecisely estimated

$$Var(\hat{\beta}_1) = \frac{\sigma_u^2}{n} \cdot \left[\frac{1}{1-\rho_{X_1,X_2}^2}\right] \cdot \frac{1}{\sigma_{X1}^2}$$ 课本(英) P215

Econometrics 3.2 Multiple Regression — Hypothesis Testing (Chapter 7)

Hypothesis Tests and CI for a single coefficient (Section 7.1)

$$t = \frac{\hat{\beta}_1 - \beta_{1,0}}{SE(\hat{\beta}_1)}, \quad CI: \hat{\beta}_1 \pm 1.96 \cdot SE(\hat{\beta}_1). \quad (5\%)$$

$$\widehat{Var}(\hat{\beta}) = \frac{1}{n} \cdot \left(\frac{1}{n}\sum X_i X_i'\right)^{-1} \left(\frac{1}{n-k-1}\sum X_i X_i' \hat{u}_i^2\right) \left(\frac{1}{n}\sum X_i X_i'\right)^{-1}$$

$$SE(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)}, \quad SE(\hat{\beta}_0) = \sqrt{\widehat{Var}(\hat{\beta}_0)}.$$

Tests of Joint Hypotheses (Section 7.2)

$H_0: \beta_1 = \beta_2 = 0$ vs. H_1: either $\beta_1 \neq 0$ or $\beta_2 \neq 0$.

(b) 錯誤: $P(|t_1| > 1.96 \text{ or } |t_2| > 1.96) = P(|t_1| > 1.96, |t_2| > 1.96) + P(|t_1| > 1.96, |t_2| \leq 1.96) + P(|t_1| \leq 1.96, |t_2| > 1.96)$

$= 0.05 \times 0.05 + 0.05 \times 0.95 + 0.95 \times 0.05 = 0.0975 \rightarrow \neq 0.05$

Solution ① Bonferroni (Appendix 7.1) ② F-statistic.

$$F = \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1 - \hat{\rho}_{t_1,t_2}^2}\right), \quad \hat{\rho}_{t_1,t_2} = \text{correlation between } t_1 \text{ and } t_2$$

若 $\hat{\rho}_{t_1,t_2} \xrightarrow{p} 0$ ($t_1 \perp\!\!\!\perp t_2$) $F \approx \frac{1}{2}(t_1^2 + t_2^2) \rightarrow \frac{\chi_q^2}{q}$ or $F_{q,\infty}$

critical value (PPT 3.2 P14)

If (homoskedasticity-only) F-statistic.

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

(q → number of restrictions under H_0; k → number of regressors.)

$$= \frac{(SSR_{restricted} - SSR_{unrestricted})/q}{SSR_{unrestricted}/(n - k_{unrestricted} - 1)}$$

Testing Single Restrictions on Multiple Coefficients (Section 7.3).

$H_0: \beta_1 = \beta_2$

Rearrange ("transform") the regression.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i = \beta_0 + (\beta_1 - \beta_2) X_{1i} + \beta_2 (X_{1i} + X_{2i}) + u_i$$

$$= \beta_0 + \underbrace{\gamma_1}_{\text{Test}} X_{1i} + \beta_2 W_i + u_i$$

More in Matrix Setup (Section 18.3)

Joint Hypothesis: $R\beta = r$

heteroskedasticity-robust F: $(R\hat{\beta}-r)'[R\,\widehat{var}(\hat{\beta})R']^{-1}(R\hat{\beta}-r)/q$

(If 4 LS assumptions hold) $\xrightarrow{d}$ $F_{q,\infty}$.

Confidence Sets for Multiple Coefficients (Section 7.4)

Joint confidence set of (β_1, β_2)

Let $F(\beta_{1,0}, \beta_{2,0})$ be the (heteroskedasticity-robust) F-stat.

95% confidence set = $\{\beta_{1,0}, \beta_{2,0} : F(\beta_{1,0}, \beta_{2,0}) < 3.00\}$ ↓ 5% critical value of $F_{2,\infty}$.

an ellipse: $\{\beta_1, \beta_2 : \frac{1}{2}\left(\frac{t_1^2 + t_2^2 - 2\hat{\rho}_{t_1,t_2} t_1 t_2}{1-\hat{\rho}^2_{t_1,t_2}}\right) < 3.00\}$

$$F = \frac{1}{2(1-\hat{\rho}^2_{t_1,t_2})}\left[\left(\frac{\hat{\beta}_2-\beta_{2,0}}{SE(\hat{\beta}_2)}\right)^2 + \left(\frac{\hat{\beta}_1-\beta_{1,0}}{SE(\hat{\beta}_1)}\right)^2 - 2\hat{\rho}_{t_1,t_2}\left(\frac{\hat{\beta}_2-\beta_{2,0}}{SE(\hat{\beta}_2)}\right)\left(\frac{\hat{\beta}_1-\beta_{1,0}}{SE(\hat{\beta}_1)}\right)\right].$$

Digression about measures of fit.

High R^2 ($\bar{R}^2$) { regression explain variation in Y.
- NOT mean eliminated omitted variable bias
- NOT mean have unbiased estimator of causal effect (β_1)
- NOT mean included variables are statistically significant (must use hypotheses tests)

Econometrics 4 Nonlinear Regression

The general nonlinear population regression function

$$Y_i = f_2(X_{1i}, X_{2i}, \cdots, X_{ki}) + u_i, \quad i = 1, \cdots, n.$$

Assumptions.

1. $E(u_i | X_{1i}, \cdots, X_{ki}) = 0$ $\Rightarrow$ f is the conditional expectation of Y given X's.
2. $(X_{1i}, \cdots, X_{ki}, Y_i)$ i.i.d.
3. Big outliers are rare
4. No perfect multicollinearity

① Polynomials ② Logarithmic transformation ("percentages" interpretation)

1. Polynomials $Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \cdots + \beta_r X_i^r + u_i$

Choice of r : ① plot, t-/F-test / Check sensitivity / Judgment ② model selection

2. Logarithmic functions of Y / X.

Reason: $\ln(X+\Delta X) - \ln(X) = \ln\left(1 + \frac{\Delta X}{X}\right) \approx \frac{\Delta X}{X}$

I. $Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$ (linear-log) II. $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$ (log-linear) III. $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$ (log-log)

I. $Y = \beta_0 + \beta_1 \ln(X)$. $\begin{cases} Y + \Delta Y = \beta_0 + \beta_1 \ln(X + \Delta X) \\ Y = \beta_0 + \beta_1 \ln(X) \end{cases} \Rightarrow \Delta Y = \beta_1 \ln\left(1 + \frac{\Delta X}{X}\right) \approx \beta_1 \frac{\Delta X}{X}$ $\left(\beta_1 \approx \frac{\Delta Y}{\Delta X / X}\right)$

$\beta_1 \approx \frac{\Delta Y / Y}{\Delta X}$ (percentage change in X / percentage change in Y)

II. $\ln(Y) = \beta_0 + \beta_1 X$. $\begin{cases} \ln(Y + \Delta Y) = \beta_0 + \beta_1 (X + \Delta X) \\ \ln Y = \beta_0 + \beta_1 X \end{cases} \Rightarrow \ln\left(1 + \frac{\Delta Y}{Y}\right) \approx \frac{\Delta Y}{Y} \approx \beta_1 \Delta X$ $\left(\beta_1 \approx \frac{\Delta Y / Y}{\Delta X}\right)$

III. $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$. $\beta_1 \approx \frac{\Delta Y / Y}{\Delta X / X}$.

Interactions Between Independent Variables (Section 8.3)

(a) 2 binary. $Y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 (D_{1i} \times D_{2i}) + u_i$

$E(Y_i | D_{1i} = 0, D_{2i} = d_2) = \beta_0 + \beta_2 d_2$ $\quad E(Y_i | D_{1i} = 1, D_{2i} = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2$

$\beta_1 + \beta_3 d_2$ → increment to the effect of D_1 when $D_2 = 1$.

(b) binary & continuous. $Y_i = \beta_0 + \beta_1 D_i + \beta_2 X_i + \beta_3 (D_i \times X_i) + u_i$

$\begin{cases} D = 0: \ Y_i = \beta_0 + \beta_2 X_i + u_i \\ D = 1: \ Y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3) X_i + u_i \end{cases}$

$\begin{cases} Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D X) \\ Y + \Delta Y = \beta_0 + \beta_1 D + \beta_2 (X + \Delta X) + \beta_3 [D (X + \Delta X)] \end{cases} \Rightarrow \Delta Y = \beta_2 \Delta X + \beta_3 D \Delta X \Rightarrow \frac{\Delta Y}{\Delta X} = \beta_2 + \beta_3 D.$

(c) 2 continuous

Econometrics 5 Assessing Studies based on multiple Regression

- ✓ Internal validity : its statistical inferences about causal effects are valid for population being studied
- External validity : its statistical inferences can be generalized to other populations and settings

  - estimator of causal effect : unbiased and consistent
  - hypothesis tests should have the desired significance level

OLS : internal validity { unbiased and consistent; standard errors are computed correctly

Threats:
1. omitted variable bias
2. misspecification of the functional form
3. measurement error
4. sample selection bias
5. simultaneous causality bias

→ $E(u_i | X_{1i}, \dots, X_{ki}) \neq 0$.

Omitted variable bias

(i) a determinant of Y (ii) correlated with at least one included regressor

solution : ① include ② instrumental variable ③ randomized controlled experiment

Misspecification of the functional form

① Continuous : log, interaction, etc. ② discrete : probit/logit

Errors-in-variable bias

$$Y_i = \beta_0 + \beta_1 X_i + u_i = \beta_0 + \beta_1 \tilde{X}_i + \left[\beta_1 (X_i - \tilde{X}_i) + u_i\right] = \beta_0 + \beta_1 \tilde{X}_i + \tilde{u}_i$$

X_i: unmeasured true value of X; $\tilde{X}_i$: imprecisely measured version of X

$$\text{cov}(\tilde{X}_i, \tilde{u}_i) = \text{cov}(\tilde{X}_i, \beta_1 (X_i - \tilde{X}_i + u_i) = \beta_1 \text{cov}(\tilde{X}_i, X_i - \tilde{X}_i) + \text{cov}(\tilde{X}_i, u_i)$$

$$= \beta_1 [\text{cov}(\tilde{X}_i, X_i) - \text{Var}(\tilde{X}_i)] + 0 \neq 0$$

Typically not equal

Sample selection Bias

Simultaneous Causality Bias

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

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

→ large u_i → large Y_i ($\gamma_1 > 0$) ⇒ $\text{corr}(X_i, u_i) \neq 0$.