Probability & statistics

① Random variables

expectation

$$E(X) = \int_{-\infty}^{\infty} x \cdot f(x)\, dx = \sum x \cdot f(x)$$

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx = \sum g(x) \cdot f(x)$$

$$E(aX+b) = a\,E(X) + b$$

$$E(X \pm Y) = E(X) \pm E(Y)$$

variance

$$V(X) = E(X^2) - [E(X)]^2$$

$$= \sum (x_i - \mu_X)^2 f(x)$$

$$= \int_{-\infty}^{\infty} (x_i - \mu_X)^2 f(x)\, dx$$

$$V(aX+b) = a^2 V(X)$$

$$V(aX \pm bY) = a^2 V(X) + b^2 V(Y) + 2ab\, cov(X,Y)$$

② sampling

bias $E(\hat{\theta} - \theta) \Rightarrow$ bias $= 0$ then unbiased

variance $E(\hat{\theta}^2) - E(\hat{\theta})^2 \Rightarrow$ smaller then more efficient

MSE $MSE = bias^2 + variance$

consistency as $n \to \infty$, $\hat{\theta} \to \theta$

$\hat{\theta}$ is consistent iff $\lim_{n\to\infty} P(\hat{\theta} - \theta = 0) = 1$

④ 2D random variables

covariance

$$cov(X,Y) = E(XY) - \mu_X \mu_Y$$

$$cov(aX+b, cY+d) = ac\, cov(X,Y)$$

X, Y independent $\longrightarrow cov(X,Y) = 0$, not converse

correlation

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$$\hat{\rho}_{X,Y} = \frac{s_{xy}}{s_x s_y} = \frac{\frac{1}{n-1}\sum (x - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum (x-\bar{x}) \cdot \frac{1}{n-1}\sum (y_i - \bar{y})}$$

③ distribution of $\bar{X}$

expectation & variance

under SRS, X_i is i.i.d. $\quad \bar{X} = \frac{1}{N}\sum X_i$

$$E(\bar{X}) = \mu_X \qquad V(\bar{X}) = \frac{\sigma^2}{N}$$

$$s_X^2 = \frac{1}{n-1}\sum (X_i - \bar{X})^2$$

$$SE(\bar{X}) = \frac{s_X}{\sqrt{n}}$$

central limit theorem

As $n \to \infty$, $\bar{X} \to N(\mu_X, \frac{\sigma^2}{n})$

$n \geq 100$, $\bar{X} \sim N(\mu_X, \frac{\sigma^2}{n})$ approximately

law of large numbers

As $n \to \infty$, $P(\bar{X} - \mu_X = 0) \to 1$

$\hookrightarrow \bar{x} \to \mu_X$

hypothesis testing

1. H_0, H_1
2. $\bar{X} \sim N(\mu_{X,H_0}, \frac{\sigma^2}{n})$ approximately,

   $\bar{X} \sim T_{n-1}(\mu_{X,H_0}, \frac{s_X^2}{n})$
3. test statistic $= \dfrac{\bar{x} - \mu_{X,H_0}}{\frac{\sigma}{\sqrt{n}}}$ (N) or $\dfrac{\bar{x} - \mu_{X,H_0}}{\frac{s_x}{\sqrt{n}}}$ (t_{n-1}); $SE(\bar{X})$
4. calculate p-value

confidence interval

90% 95% 99%

1.64, 1.96, 2.58

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{X} \pm t_{n-1,\alpha/2}\, SE(\bar{X})$$

$\Rightarrow$ in the long run, $1-\alpha$ % of all samples will give an interval that contains the true μ

Linear regression

① OLS estimation of bivariate linear regression

linear model

$$E(Y) = E(E(Y \mid X = x))$$

$$E(Y_i \mid X_i) = \beta_0 + \beta_1 X_i$$

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

OLS estimation

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad \hat{u}_i = Y_i - \hat{Y}_i$$

$$\min \sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2$$

$$\frac{\partial Q}{\partial \hat{\beta}_0} = -2 \sum Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = 0$$

$$\frac{\partial Q}{\partial \hat{\beta}_1} = -2 \sum X_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} = \frac{\frac{1}{n-1} \sum (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum (X_i - \bar{X})^2}$$

$$\frac{1}{n} \sum \hat{u}_i = 0$$

$$\frac{1}{n} \sum \hat{Y}_i = \bar{Y}, \ \bar{Y} \text{ lies on line}$$

$$\sum u_i X_i = 0$$

$$\text{cov}(\hat{u}, X) = 0$$

② sampling distributions

$$E(\hat{\beta}_1) = \beta_1$$

larger $V(X)$, smaller $V(\hat{\beta}_1)$

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

Assumptions

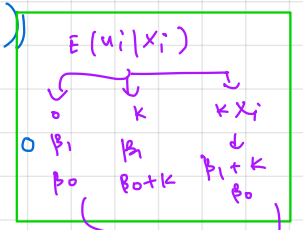
1. $E(u_i \mid X = x) = 0$
2. (X_i, Y_i) are independtly & identically distributed
3. large outliers are rare

bias

$$E(\hat{\beta}_1 - \beta_1 \mid X_1, X_2 \dots X_n) = E\left(\frac{\sum (X_i - \bar{X}) u_i}{\sum (X_i - \bar{X})^2} \mid X_1, X_2 \dots\right)$$

$$= \frac{\sum (X_i - \bar{X}) E(u_i \mid X_1, X_2 \dots)}{\sum (X_i - \bar{X})^2}$$ (i.i.d.)

$$= \frac{\sum (X_i - \bar{X}) E(u_i \mid X_i)}{\sum (X_i - \bar{X})^2} = 0$$



$\rho_{Xu} = 0$ $\quad \rho_{Xu} \neq 0$

if constant, can take out of the sum

consistency

$$E(\hat{\beta}_1) = \beta_1 \ \& \ V(\hat{\beta}_1) \propto \frac{1}{n} \Rightarrow \lim_{n \to \infty} \hat{\beta}_1 \to \beta_1$$

③ homo & heteroskedasticity

$V(u_i \mid X_i = x) = \sigma_u^2$ constant

$V(u_i \mid X_i = x) \neq k$

iff homoskedastic:

$$V(\hat{\beta}_1) = \frac{1}{n} \frac{\sigma_u^2}{\sigma_X^2} = \frac{1}{n} \cdot \frac{\frac{1}{n-2} \sum \hat{u}_i^2}{\frac{1}{n} \sum (X_i - \bar{X})^2}$$

heteroskedastic robust and homoskedastic estimators of $V(\hat{\beta}_1)$ converge as $n \to \infty$

④ studying r/s: measures of fit

R^2 measures fraction of sample variance explained by model. unit-less & directionless.

$$Y_i = \hat{Y}_i + \hat{u}_i$$

predicted variance

$$R^2 = \frac{\text{explained sum of squares}}{\text{total sum of squares}} = \frac{\frac{1}{n-1} \sum (\hat{Y}_i - \bar{\hat{Y}})^2}{\frac{1}{n-1} \sum (Y_i - \bar{Y})^2}$$

$$= 1 - \frac{\text{sum of squared residuals}}{\text{total sum of squares}} = \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}$$

actual variance

fraction of variance not explained by model

$$= \frac{\text{cov}(X, Y)}{\sqrt{V(X)}\sqrt{V(Y)}} = \frac{S_{xy}}{s_x s_y}$$

$ESS + SSR = TSS$

Standard error of regression measure of spread of observations about regression line wrt to Y

ie. sample st. dev of OLS residuals

$$SER = \sqrt{\frac{1}{n-2} \sum (\hat{u}_i - \bar{\hat{u}}_i)^2} = \sqrt{\frac{1}{n-2} \sum \hat{u}_i^2}$$

$= 0$ due to OLS

Root mean squared error

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

⑤ studying r/s: hypothesis testing

1. $H_0: \beta_1 = k$, $H_1: \beta_1 \neq k$
2. compute $SE(\hat{\beta}_1)$

$$se(\hat{\beta}_1) = \sqrt{\frac{var(x_i - \bar{x})(u_i)}{var(x_i)^2}}$$

scaling affects here ⇒ take out, k^2 on var, k on $se(\hat{\beta}_1)$

$$= \sqrt{\frac{\frac{1}{n-2}\sum (x_i - \bar{x})^2 \hat{u}_i^2}{[\frac{1}{n}\sum (x_i - \bar{x})]^2}}$$

⇒ under H_0, since $n \geq 100$, $\hat{\beta} \sim N(\beta_1, se(\hat{\beta}_1))$ approximately

3. compute t-statistic

$$t = \frac{\hat{b}_1 - \beta_{1,H_0}}{se(\hat{\beta}_1)} \sim \text{approx normal}$$

$$\hat{y}_i = \hat{b}_1 x_i + \hat{b}_0 \qquad \hat{u}_i = y_i - \hat{y}_i$$

$$\hat{b}_1 = \frac{s_{xy}}{s_x^2} = \frac{\frac{1}{n-1}\sum (x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n-1}\sum (x_i - \bar{x})^2}$$

$$\hat{b}_0 = \bar{y} - \hat{b}_1 \bar{x}$$

4. compute p-value
5. if two tailed, compute confidence interval

$$\hat{b}_1 \pm t_{n-1, \alpha/2}\, se(\hat{\beta}_1) \approx \hat{b}_1 \pm z_{\alpha/2} \cdot se(\hat{\beta}_1)$$

1.64, 1.96, 2.58
90% 95% 99%

⑥ omitted variable bias

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \rho_{xz} \cdot \frac{\sigma_z}{\sigma_x} + E\left(\frac{\sum (x_i - \bar{x}) u_i}{\sum (x_i - \bar{x})^2}\right)$$

omitted absorbed in $\tilde{u}_i$

$$= \beta_1 + E\left(\frac{\sum (x_i - \bar{x})(\beta_2 z_i + u_i)}{\sum (x_i - \bar{x})^2}\right)$$

$$= \beta_1 + \rho_{x\tilde{u}} \cdot \frac{\sigma_{\tilde{u}}}{\sigma_x}$$

when does it occur?

1. the omitted variable is correlated with the included regressor(s)
2. the omitted variable is a determinant of the dependent variable

⑦ unit change

$Y \to aY \Rightarrow \hat{\beta}_1 \to a\hat{\beta}_1$, $\hat{\beta}_0 \to a\hat{\beta}_0$

$X \to bX \Rightarrow \hat{\beta}_1 \to \frac{\hat{\beta}_1}{b}$, $\hat{\beta}_0$ unchanged

⑧ interpretation of regression line

continuous variable

slope: $\frac{\Delta E(Y)}{\Delta X}$

intercept: $E(Y \mid X = 0)$

binary variable

$\hat{\beta}_1$: difference in $E(Y)$ when X is present vs. not. (i.e. diff in avg. Y)

$\hat{\beta}_0$: $E(Y \mid X = 0)$

predictions

x_i should be SRS from same distribution, and within range of data

Multiple regression

① functional form & OLS

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i \dots + \varepsilon_i$$

OLS

$$\hat{Y} = \hat{\beta_0} + \hat{\beta_1} X_i \dots \Rightarrow E(Y \mid X_{1i}, X_{2i} \dots)$$

$$\hat{u_i} = Y_i - \hat{Y_i}$$

$$S = \sum_{i=1}^{n} \hat{u_i}^2 \quad \left\{ \begin{array}{l} \frac{\partial S}{\partial \beta_0} = 0 \\ \frac{\partial S}{\partial \beta_1} = 0 \\ \vdots \end{array} \right\} \Rightarrow \text{solve}$$

② multicolinearity

perfect multicolinearity — one regressor is a perfect linear function of other regressor(s)

eg. $X_1 = kX_2$

imperfect multicolinearity — regressors are said to exhibit imperfect multicolinearity when they are highly correlated

③ Assumptions of multivariable regression

1. $E(u_i \mid X_{1i}, X_{2i}, X_{3i} \dots) = 0$
2. $(X_{1i}, X_{2i} \dots)_j$ — each observation — is IID
3. large outliers are not common
4. **no perfect collinearity** causes divide by 0

④ estimator distribution

↳ same as bivariate case

imperfect multicolinearity

↳ when 2 or more regressors are highly correlated, then during OLS estimation, in keeping all else constant, $V(X_j)$ very small.

$$V(\hat{\beta_j}) \propto \frac{1}{V(X_j)}$$

$\Rightarrow \hat{\beta_j}$ still unbiased & consistent, just larger $SE(\hat{\beta_j})$ since $V(X_j)$ is smaller

⑤ Binary multiple regression

Dummy variable trap — if there are G binary categories and there are G regressors + intercept, fail due to perfect multicolinearity because $\beta_0 \Leftrightarrow \beta_0(1) \Leftrightarrow X_{0,j} = 1$, 1 will always be scalar multiple of all other X_i

$G-1$ regressors + $\hat{\beta_0}$ — exclude $\hat{\beta_0}$

⑥ Hypothesis testing: F test

joint hypothesis testing

1. H_0 : $\underbrace{\beta_j = k_j}_{\text{restriction}}$ ∧ $\beta_m = k_m \dots$ } q restrictions

   H_1 : at least 1 of q restrictions false

2. calculate q t-statistics and sum them to get F test stat

$$F \sim \frac{\chi^2_m}{m} \Big/ \frac{\chi^2_n}{n} + \hat{\beta_j}^2 \sim \chi^2_1 \text{ approximately}$$

t-stat $\left\{ \sum_{j}^{n} \frac{b_{j,0} - \hat{b_j}}{se(\hat{\beta_j})} \sim \frac{\chi^2_q}{q} \right.$

$$f = \frac{\chi^2_q}{q} \Big/ \frac{\chi^2_n}{n}$$ ← sample size

single restriction of multiple $\hat{\beta_i}$ eg. $\beta_1 = \beta_2$

1. reorganise eqn so that $(\beta_1 - \beta_2)$ becomes coefficient
2. t-test it

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

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

$$= \beta_0 + \gamma X_{1i} + \beta_2 V_i, \quad V = X_{1i} + X_{2i}$$

⑦ measures of fit

SER $SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2}$

$$s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum (\hat{u}_i - \cancel{\bar{\hat{u}}_i}^{=0})^2 = \frac{SSR}{n-k-1}$$

$n-k-1$ ← num. of non-β_0 regressors

adjusted $\bar{R}^2$

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}\frac{\sum \hat{u}_i^2}{\sum (y_i-\bar{y})^2}$$

$$= 1 - \frac{n-1}{n-k-1}\frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_y^2}$$

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

---

non-linear regression

① models & effect of $\Delta\{X\}$

$$y_j = f(\{(x_1, x_2 \ldots)\}y_j) + u_j$$

some non-linear functional form

effect of ΔX on Y — take the difference

$$\Delta\hat{y} = f(x_1 + \Delta x_1, x_2 + \Delta x_2 \ldots) - f(x_1, x_2 \ldots)$$

standard error of estimated effect

↳ in the linear case, $\hat{y} = \hat{\beta}_0 + \sum \hat{\beta}_i x_i$

$\Rightarrow \Delta\hat{y} = \hat{\beta}_1(\Delta x_j) \Rightarrow SE(\Delta\hat{y}) = SE(\hat{\beta}_i)\cdot \Delta x_i$

↳ in non-linear case, depends on functional form

eg. $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x^2 \ldots$

$\Rightarrow \Delta\hat{y} = \hat{\beta}_1(\Delta x) + \hat{\beta}_2((x+\Delta x)^2 - x^2)$

$SE(\Delta\hat{y}) = SE(\hat{\beta}_1)\cdot\Delta x + SE(\hat{\beta}_2)((x+\Delta x)^2 - x^2)$

---

② polynomial models

$$y_j = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2 \ldots$$

$\Rightarrow$ we still treat as multiple regression, so multivariate OLS as per usual

$\Rightarrow$ when $\Delta x \Rightarrow$ remember to sum effects of <u>all</u> β_j at that <u>range</u> difference

③ logarithmic models

1) properties of logarithms

$$\ln\left(\frac{a}{b}\right) = \ln a - \ln b$$

$$\ln(ab) = \ln a + \ln b$$

$$\ln(a^b) = b \ln a$$

logarithms & percentages

$$\ln(x+\Delta x) - \ln(x) = \ln \frac{x+\Delta x}{x}$$

$$= \ln\left(1 + \frac{\Delta x}{x}\right)$$

$$\approx \frac{\Delta x}{x} \quad \text{when } \Delta x \text{ is small}$$

2) three log models

Case	Population Regression Model	Interpretation of β_1
I. linear-log	$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i$	A 1% change in X is associated with a change in Y of $0.01\beta_1$
II. log-linear	$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$	A change in X by one unit ($\Delta X = 1$) is associated with a $100\beta_1$% change in Y
III. log-log	$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$	A 1% change in X is associated with a β_1% change in Y

linear-log model

$$Y_j = \beta_0 + \sum \beta_i \ln(X_{ij}) + u_j$$

$$Y_i + \Delta Y_i = \beta_0 + \beta_1 \ln(X_i + \Delta X)$$

$$\Delta Y_i = \beta_1 [\ln(X_i + \Delta X) - \ln(X)] \approx \beta_1 \left(\frac{\Delta X_i}{X_i}\right) \approx \beta (0.01) K$$

⇒ A K% change in X predict an average change in Y by $\beta_1 \times K$%

log-linear model

$$\ln(Y_j) = \beta_0 + \sum \beta_i X_{ij} + u_j$$

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 (X + \Delta X)$$

$$\ln(Y + \Delta Y) - \ln(Y) = \beta_1 \Delta X \approx \frac{\Delta Y}{Y}$$

⇒ A unit change in X is associated w a β_1% change in Y on <u>average</u>

log-log model

$$\ln(Y_j) = \beta_0 + \sum \beta_i \ln(X_{ij}) + u_j$$

$$\ln(Y + \Delta Y) = \beta_0 + \beta_1 \ln(X + \Delta X)$$

$$\ln(Y + \Delta Y) - \ln(Y) = \beta_1 [\ln(X + \Delta X) - \ln(X)]$$

$$\frac{\Delta Y}{Y} \approx \beta_1 \frac{\Delta X}{X}$$

⇒ a K% change in X is associated w an average $\beta_1 \cdot K$% change in Y

predicted values of Y when Y is in log

⇒ taking the exponent is biased

eg. $\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i$

$$\hookrightarrow Y_i = e^{\beta_0 + \beta_1 X_i} \cdot e^{u_i}$$

scaled bias

$$E(Y_i | X_i) = E\left(e^{\beta_0 + \beta_1 X_i} | X_i\right) \cdot E\left(e^{u_i} | X_i\right)$$

⇒ even if $E(u_i | X_i) = 0$, $E(e^{u_i} | X_i) \neq 1$ in general

④ Interaction between variables

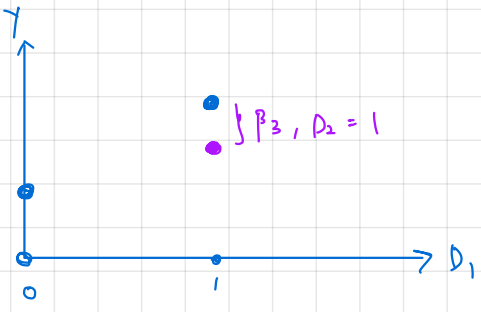
binary-binary

$$Y_i = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_1 D_2 + u_i$$

$$E(Y_i \mid D_1 = 1, D_2 = d_2) = \beta_0 + \beta_1 + \beta_2 d_2 + \beta_3 d_2$$

$$E(Y_i \mid D_1 = 0, D_2 = d_2) = \beta_0 + \beta_2 d_2$$

Expected difference when $D_1 = 1$ vs. $D_1 = 0$ $\Rightarrow$ $\beta_1 + \beta_3 d_2$



$\Rightarrow$ effect on Y of change in D_1 now depends on D_2

$\Rightarrow$ β_3 is the difference in effect on Y of D_1 when $D_2 = 1$ vs. $D_2 = 0$

---

continuous-continuous

↳ effectively finer granularity of binary-binary case

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

(extra effect from changing both X_1 or X_2)

↳ effect of X_1 holding X_2 constant

$$\frac{dy}{dX_1} = \beta_1 + \beta_3 X_2$$

$\Rightarrow$ $\beta_3 \cdot X_2$: the expected difference in effect of a unit change in X_1 given $X_2 = X_2$

---

continuous - binary

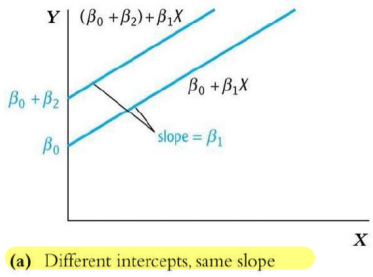
↳ inclusion of binary term allows for different offset

(1. test if they have diff. intercept ie. $\beta_2 = 0$)

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i$$

$D_i = 0$: $y_i = \beta_0 + \beta_1 x_i + u_i$

$D_i = 1$: $y_i = (\beta_0 + \beta_2) + \beta_1 x_i + u_i$



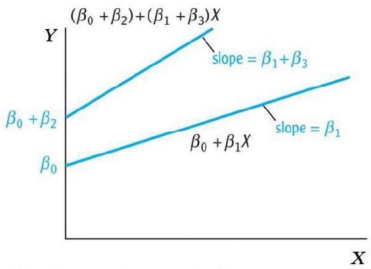
(a) Different intercepts, same slope

↳ inclusion of continuous-binary interaction allows the effect of change in X to be different (ie. different gradient) (2. test if they have diff gradient ie. $\beta_3 = 0$ (t-test))

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 X_i D_i + u_i$$

if $D_i = 1$: $Y_i = \beta_0 + (\beta_1 + \beta_3) X_i + \beta_2$

if $D_i = 0$: $y_i = \beta_0 + \beta_1 X_i$



(b) Different intercepts, different slopes

$\Rightarrow$ β_3 - difference in effect of unit change in X_2 for observations where $D_i = 1$ vs. $D_i = 0$

Assessing regression

① Validity of a study

internal validity — for a study to be internally valid:

1. the estimator of the causal effect of interest should be unbiased and consistent
2. hypothesis tests have the correct significance level; standard errors are correctly estimated (hetero vs. homoskedasticity)

external validity — A study is externally valid if its inferences can be generalised to other populations & settings

---

② Threats to internal validity

omitted variable bias $E(u_i|x_i) \neq 0$

1. the omitted variable is correlated with the included regressor(s)
2. the omitted variable is a determinant of the dependent variable

⇒ solutions: include omitted variables, use panel data, instrumental variables or RCT

---

incorrect functional form

↳ bias can arise if the wrong functional form is used. we will observe trends in the residuals. solution: fit better models

---

measurement error

errors in x_i

↳ in general, $\hat{\beta}_i$ will be biased and inconsistent

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$
$$= \beta_0 + \beta_1 \tilde{X}_i + [\beta_1 (x_i - \tilde{x}_i) + u_i]$$

$\tilde{X}_i$: observed data; v_i: $\beta_1 (x_i - \tilde{x}_i) + u_i$

errors in y_i

↳ in general, no bias to $\hat{\beta}_i$, but $\text{Var}(\hat{\beta}_i)$ will be larger

$$Y_i = \beta_0 + \beta_1 x_i + u_i + w_i$$

⇒ so long as measurement error on Y_i is random, $E(w_i|x_i) = 0$, so $E(v_i|x_i) = E(w_i|x_i) + E(u_i|x_i) = 0$

⇒ check corr. between $\tilde{x}_i$ and v_i $= \dfrac{\text{cov}(\tilde{x}_i, v_i)}{\sigma_{\tilde{x}_i} \sigma_{v_i}}$

1. $\text{cov}(\tilde{x}_i, v_i) = 0$ iff $E(\quad | \quad) = 0$, then $\hat{\beta}_1$ unbiased & consistent

2. $$\text{cov}(\tilde{x}_i, v_i) = \text{cov}(\tilde{x}_i, \beta x_i - \beta \tilde{x}_i + u_i)$$
$$= \text{cov}(\tilde{x}_i, \beta x_i) - \text{cov}(\tilde{x}_i, \beta \tilde{x}_i) + \text{cov}(\tilde{x}_i, u_i)$$
$$= \beta(\text{cov}(\tilde{x}_i, \beta x_i) - \sigma_{\tilde{x}_i}^2)$$

$\text{cov}(\tilde{x}_i, u_i) = 0$, since $E(u_i|x_i) = 0$

$\sigma_{\tilde{x}}^2$ if $\tilde{x}_i = x_i$

in general, $\neq 0$, so $E(v_i|x_i) \neq 0$

sample selection bias — in some cases, SRS is thwarted because the dependent variable is only observed for a restricted sample eg. education on earnings — we only study the sub-group that is employed → we fail to study effect of education earnings had that group worked.

---

simultaneous causality bias

↳ if $x \to y$ but $y \to x$, then $\hat{\beta}_1$ will be picking up the simultaneous effects
⇒ OLS estimator of $\hat{\beta}_1$ will be biased and inconsistent

Suppose: $y_i = \beta_0 + \beta_1 x_i + u_i$ ⇒ we are fitting this

$x_i = \gamma_0 + \gamma_1 y_i + v_i$

$\gamma_1 > 0$: $u_i \uparrow, y_i \uparrow, x_i \uparrow$; $\rho_{u_i, x_i} > 0$

$\gamma_1 < 0$: $u_i \uparrow, y_i \uparrow, x_i \downarrow$; $\rho_{u_i, x_i} < 0$

$E(u_i | x_i) \neq 0$

↳ solution: RCT

---

③ threats to external validity

↳ whether one can generalize results found in a study to other populations & settings depend on how similar the populations & settings are
eg. results from canada → can apply to SG?

↳ practical significance, beyond statistical significance!

Regression w/ panel data

① fixed effects regression

T=2, Take difference

$$Y_{i1} = \alpha_i + \lambda_1 + \beta_1 X_{i1} + \varepsilon_{i1}$$

$$Y_{i2} = \alpha_i + \lambda_2 + \beta_1 X_{i2} + \varepsilon_{i2}$$

$$Y_{i2} - Y_{i1} = \Delta Y_{it}$$

$$\Delta Y_{it} = \beta_0 + \beta_1 \Delta X_{it} + \Delta \varepsilon_{it}$$

$$(\lambda_2 - \lambda_1) + \beta_1 (X_{i2} - X_{i1}) + (\varepsilon_{i2} - \varepsilon_{i1})$$

using binary variables

↳ model has $T \times n$ different intercepts $(\lambda_t + \alpha_i)$ for each entity-time period pair.

↳ to avoid perfect collinearity, we can model intercepts as $(n-1)$ binary variables, one for each entity and another $(T-1)$ binary variables, one for each period.

$$\beta_0 = (\alpha_1 + \lambda_1)$$

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \ldots + \gamma_2 D_{2i} + \ldots \gamma_n D_{ni} + \delta_2 T_{2t} + \ldots \delta_T T_{Tt}$$

- γ_n: offset for entity n
- D_{ni}: is i the n^{th} entity
- δ_T: offset for time T
- T_{Tt}: is t the T^{th} period

$$Y_{it} = \alpha_i + \lambda_t + \beta_1 X_{it} + \varepsilon_{it}$$

$$\gamma_i = \alpha_i - \alpha_1$$

$$\delta_t = \lambda_t - \lambda_1$$

de-meaning

$$Y_{it} = \alpha_i + \gamma_t + \beta_1 X_{it} + \varepsilon_{it}$$

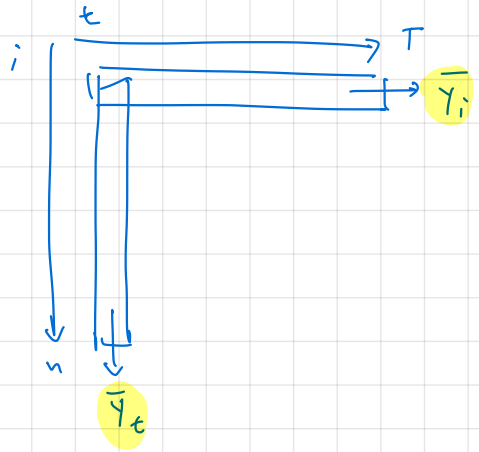
$$\tilde{Y}_{it} = Y_{it} - \bar{Y}_i - \bar{Y}_t$$

$$= \cancel{\alpha_i} + \cancel{\lambda_t} + \beta_1 X_{it} + \varepsilon_{it} - (\cancel{\alpha_i} + \beta_1 \bar{X}_i + \bar{\varepsilon}_i) - (\cancel{\lambda_t} + \beta_1 \bar{X}_t + \bar{\varepsilon}_t)$$

- α_i + $\beta_1 \bar{X}_i$ + $\bar{\varepsilon}_i$: time effect lost, assuming avg. = 0
- λ_t + $\beta_1 \bar{X}_t$ + $\bar{\varepsilon}_t$: entity effect lost, assuming avg. = 0

$$= \beta_1 (X_{it} - \bar{X}_i - \bar{X}_t) + (\varepsilon_{it} - \bar{\varepsilon}_i - \bar{\varepsilon}_t)$$

$$= \beta_1 \tilde{X}_{it} + \tilde{\varepsilon}_{it}$$



$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i = 0 \qquad \frac{1}{T}\sum_{t=1}^{T} \alpha_i = \alpha_i$$

$$\frac{1}{n}\sum_{i=1}^{n} \lambda_t = \lambda_t \qquad \frac{1}{T}\sum_{t=1}^{T} \lambda_t = 0$$

estimated directly by mean if no other regressors

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} = \beta_1 \frac{1}{T}\sum_{t=1}^{T} X_{it} + \frac{1}{T}\sum_{t=1}^{T} u_{it} + \frac{1}{T}\sum_{t=1}^{T} \alpha_i + \frac{1}{T}\sum_{t=1}^{T} \lambda_t$$

(α_i and λ_t terms: 0)

$$\bar{Y}_t = \frac{1}{n}\sum_{i=1}^{n} Y_{it} = \beta_1 \frac{1}{n}\sum_{i=1}^{n} X_{it} + \frac{1}{n}\sum_{i=1}^{n} u_i + \frac{1}{n}\sum_{i=1}^{n} \alpha_i + \frac{1}{n}\sum_{i=1}^{n} \lambda_t$$

(α_i and λ_t terms: 0)

② estimation and estimator distribution

OLS estimation and estimator distributions

$$\hat{\beta}_j^{\text{before attr}} = \sum_{i=1}^{n} \frac{(X_{i2} - X_{i1})(Y_{i2} - Y_{i1})}{\sum_{i=1}^{n} (X_{i2} - X_{i1})^2}$$

$$\hat{\beta}_j^{\text{demean}} = \frac{\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{X}_{ijt}\tilde{Y}_{it}}{\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{X}_{ijt}^2}$$

$$Q_X = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{X}_{it}^2 \qquad \eta_j = \frac{1}{T}\sum_{t=1}^{T} \tilde{X}_{it}\tilde{u}_{it}$$

$$\sqrt{nT}(\hat{\beta}_1 - \beta_1) \sim N\left(0, \frac{\text{Var}(\eta_j)}{Q_X^2}\right) \text{ approximately as } n \to \infty$$

$$E(\hat{\beta}_j) = \beta_1 + \frac{\frac{1}{n}\cdot\frac{1}{T}\cdot\sum_{i=1}^{n}\sum_{t=1}^{T} \tilde{X}_{ijt}\, E(u_{it} \mid X_{ijt})}{\frac{1}{n}\cdot\frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T} X_{ijt}^2}$$

$$\lim_{n\to\infty} \text{Var}(\hat{\beta}_j) = \frac{1}{nT}\frac{\sigma_{u_i}^2}{Q_X^2}$$

$$\text{Var}(\eta_j) = \text{var}\left(\frac{1}{T}\sum_{t=1}^{T} \tilde{X}_{it}\tilde{u}_{it}\right) = \frac{1}{T}\text{var}\left(\tilde{X}_{i1}\tilde{u}_{i1} + \tilde{X}_{i2}\tilde{u}_{i2} \cdots\right)$$

$$= \frac{1}{T}\left(\text{var}(\tilde{X}_{i1}\tilde{u}_{i1}) + \text{Var}(\tilde{X}_{i2}\tilde{u}_{i2}) \cdots + 2\,\text{cov}(\tilde{X}_{i1}\tilde{u}_{i1}, \tilde{X}_{i2}\tilde{u}_{i2}) \cdots\right)$$

---

assumptions of fixed effects regression

1. $E(u_{it} \mid X_{i1}, X_{i2} \ldots) = 0$ — for each entity, u_{it} has mean 0 given a specific entity effect and the history of X_j for that state. i.e. no OVB
2. $(X_{i1}, X_{i2} \ldots X_{iT})$ are i.i.d. — entity tuples are independently distributed. i.e. collection of observations for a given entity has no bearing on that of another.
3. (X_{it}, Y_{it}) have finite 4^{th} moments. — no extreme outliers
4. No perfect multicollinearity
5. $\text{corr}(u_{it}, u_{is} \mid X_{ijt}, X_{ijs}, \alpha_i) = 0$ for $t \neq s$ — given some variable X_j, the error terms are uncorrelated over time within a state. i.e. no serial correlation
   - ↳ if holds, can use heteroskedasticity robust SE
   - ↳ also can if just 2 time periods
   - ↳ else use clustered SE

heteroskedasticity robust vs. clustered SE

↳ recall that the heteroskedasticity robust estimator of variance is as:

$$\widehat{SE}(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)} \qquad Var(\hat{\beta}_1) = \frac{1}{n}\sum \frac{Var((X_i - \bar{X})u_i)}{Var(X_i)^2}$$

↳ above, we have derived that in fixed effects regression:

$$\lim_{n\to\infty} Var(\hat{\beta}_j) = \frac{1}{nT}\frac{\sigma_{u_i}^2}{Q_x^2} \qquad Q_x = \frac{1}{nT}\sum_{i=1}^{n}\sum_{t=1}^{T}\tilde{X}_{it}^2$$

$$Var(u_i) = Var\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{X}_{it}\tilde{u}_{it}\right) = \frac{1}{T}Var\left(\tilde{X}_{i1}\tilde{u}_{i1} + \tilde{X}_{i2}\tilde{u}_{i2} \dots\right)$$

$$= \frac{1}{T}\left(Var(\tilde{X}_{i1}\tilde{u}_{i1}) + Var(\tilde{X}_{i2}\tilde{u}_{i2}) \dots + 2\,cov(\tilde{X}_{i1}\tilde{u}_{i1}, \tilde{X}_{i2}\tilde{u}_{i2}) \dots\right)$$

⇒ heteroskedasticity robust estimator does not take into account of correlations in time. So we need one that does.

clustered SE

$$SE_{clustered}(\hat{\beta}) = \sqrt{\frac{1}{nT}\frac{\hat{\sigma}_u^2}{\hat{Q}_x^2}} \qquad \hat{\sigma}_u^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{1}{T}\sum_{t=1}^{T}\tilde{X}_{it}\hat{u}_{it} - \left(\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i\right)\right)$$

($\hat{u}_i$ = $\frac{1}{T}\sum_{t=1}^{T}\tilde{X}_{it}\hat{u}_{it}$; the bracketed term works out to 0 assuming u_i and X_i uncor.)

clustering by state

⇒ i.e. for one i, avg. across all its time, because entity likely dependent across time, but independent across state

Regression w/ dependent variable

① binary regression

↳ interpret the regression as modelling the probability that the dependent variable = 1.

model $f(\{X_j\})$ predicts $E(Y \mid \{X_j\})$

$$E(Y \mid \{X_j\}) = P(Y=1 \mid \{X_j\})(1) + P(Y=0 \mid \{X_j\})(0)$$

---

② linear probability model

↳ use of a linear model to predict probability

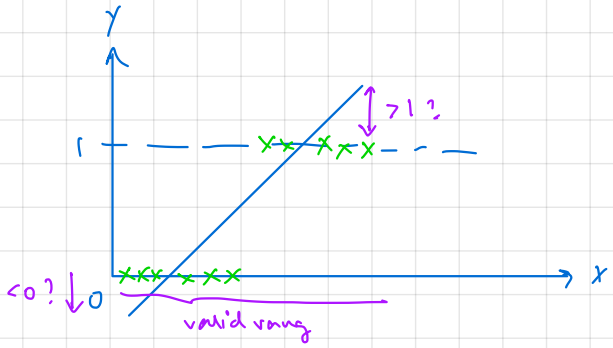
functional form $E(Y \mid \{X_j\}) = P(Y=1 \mid \{X_j\}) = \beta_0 + \sum_j \beta_j X_j$

estimation OLS

estimator distributions same as that of OLS

inference same as that of OLS

shortcomings



1. probabilities cannot exceed 1 or be negative.
2. effects of regressor(s) on probability likely to be non-linear, have saturation, etc.

---

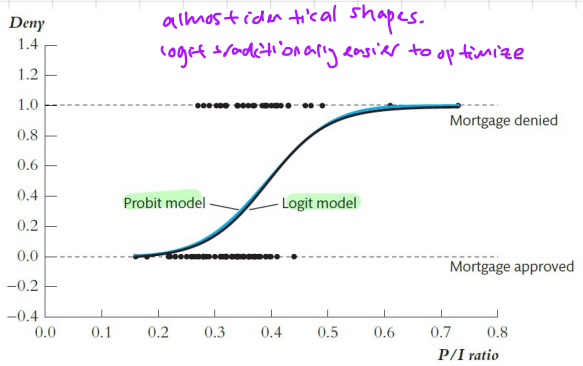
③ probit & logit regression

$z = \beta_0 + \sum_j \beta_j X_j$

$P(Y=1 \mid \{X_j\})$

→ **probit model** $\Phi(z)$, Φ is cumulative standard normal

→ **logit model** $\frac{1}{1+e^{-z}}$, logistic function



④ estimation and inference of probit & logit regression

i) non-linear least squares

↳ in probit & logit regression, the coefficients are non-linear with respect to the prediction, so ordinary least squares no longer works. But the idea of taking derivates to find a closed form solution still applies

$$Q = \sum_{i=1}^{n} \left[y_i - f\left(\hat{\beta}_0 + \sum_j \hat{\beta}_j x_{ij} \right) \right]^2 \quad L^2 \text{ loss}$$

$$\frac{\partial Q}{\partial \hat{\beta}_j} = 0 \implies \text{solve}$$

1. estimators are consistent
2. normally distributed in large samples

} but, MLE estimators have smaller variance and are thus more efficient

ii) maximum likelihood estimation

idea: assuming our data was collected SRS, iid, then can we estimate β_j by finding the coefficients that maximise the probability of observing this dataset?

1. probability of observing $Y_i \mid \{X_{ij}\}$: $p_i = f(\{\beta_j\}, \{X_{ij}\})$ (logit or probit functions)
2. we can model it as $P(Y_i = y_i) = p_i^{y_i}(1-p_i)^{1-y_i}$
3. assuming iid, $P(Y_1 = y_1, Y_2 = y_2 \dots) = \prod_{i=1}^{n} P_i(Y_i = y_i)$
4. convert product to log for convenience

$$P(Y_1 = y_1, Y_2 = y_2 \dots) = \sum_{i=1}^{n} (y_i)\ln(p_i) + (1-y_i)\ln(1-p_i)$$

(can split for convenience of expression)

5. optimise by gradient ascent / descent (if negative log, which STATA does)

iii) MLE estimator distributions

↳ $\hat{\beta}_j^{MLE}$ is unbiased & consistent, and normally distributed in large samples

↳ $\hat{\beta}_j^{MLE}$ achieves minimum asymptotic variance among unbiased estimators, and is thus the most efficient estimator in binary dependent regression

⇒ note that $SE(\hat{\beta}_j^{MLE})$ also needs to be robustly adjusted for heteroskedasticity

⇒ inference is identical to OLS for practical purposes

⑤ measures of fit

R^2 R^2 is a poor measure of fit for the LPM. This also true for the probit and logit models. Ultimately, the data is inherently non-linear and limited; R^2, which measures deviation from a continuous line is thus not a good metric.

Accuracy Does not give a good indicator of quality of prediction ⟹ whether $p = 51\%$ or 90% is the same.

Pseudo-R^2 compares the likelihood of the estimated model ie. $P(\{b_j\} \mid \{x_{ij}\})$ to the likelihood when none of the X's are included as regressors — that is, probability p_i is fixed

$$\text{pseudo-}R^2 = 1 - \frac{\ln\left(L^{fmax}_{probit}\right)}{\ln\left(L^{fmax}_{bernoulli}\right)}$$

conversion to log likelihood

(analogous to $R^2 = 1 - \frac{ESS}{TSS}$,

$1 - \frac{p'}{\bar{p}}$ → p': explained by model; $\bar{p}$: baseline p of occurance

$$L^{fmax}_{probit} = \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \quad p_i = \sigma\left(\hat{b}_0 + \sum b_j x_{ij}\right)$$

$$L^{fmax}_{bernoulli} = \prod_{i=1}^{n} \hat{p}^{y_i}(1-\hat{p})^{1-y_i}$$

($\hat{p}$: fixed)

⑥ comparing LPM, logit & probit models

↳ all three models are just approximations to the unknown population regression function $E(Y \mid \{X_j\}) = P(Y=1 \mid \{X_j\})$

1. The LPM is easiest to use and interpret, but cannot capture the non-linear nature of the true function

2. The probit and logit models are identical for all intents and purposes, but as with most non-linear models, interpretation is trickier

3. In some datasets, all three models may be very similar

IV regression

① instrumental variables

1. instrument relevance : $\text{corr}(z_i, x_{ji}) \neq 0 \Rightarrow z_i$ predicts x_j
2. instrument exogeneity : $\text{corr}(z_i, u_i) = 0 \Rightarrow z_i$ does not introduce bias

---

② TSLS

↳ idea : to find the unbiased effect of X on Y, we induce unbiased changes in X through an exogeneous variable Z. Z might also predict Y, though. So to estimate $\hat{\beta}_j$, we simply need to account for the effects of Z on Y.

$$x_i = \pi_0 + \pi_1 z_i + v_i \iff z_i = -\frac{\pi_0}{\pi_1} + \frac{1}{\pi_1} x_i + v_i$$

$$y_i = \gamma_0 + \gamma_1 z_i + w_i$$

z_i predicts both x_i and y_i w/o OVB.

$$E(v_i \mid z_i) = 0$$
$$E(w_i \mid z_i) = 0$$

$$y_i = \underbrace{\left(\gamma_0 - \gamma_1 \frac{\pi_0}{\pi_1}\right)}_{\beta_0} + \underbrace{\frac{\gamma_1}{\pi_1}}_{\beta_1} x_i + \underbrace{\left(w_i - \frac{1}{\pi_1} v_i\right)}_{u_i} \Rightarrow E(u_i \mid x_i) = 0 !$$

$$\beta_1 = \frac{\gamma_1}{\pi_1}$$

OLS estimation of $\pi_0, \pi_1, \gamma_1, \gamma_0$.

$$\gamma_1 = \frac{\text{cov}(y_i, z_i)}{\text{var}(z_i)}$$
$$\pi_1 = \frac{\text{cov}(x_i, z_i)}{\text{var}(z_i)}$$

$$\beta_1 = \frac{\gamma_1}{\pi_1} = \frac{\text{cov}(y_i, z_i)}{\text{cov}(x_i, z_i)}$$

$$\hat{\beta}_j^{TSLS} = \frac{s_{yz}}{s_{xz}} = \frac{\frac{1}{n-1}\sum (y_i - \bar{y})(z_i - \bar{z})}{\frac{1}{n-1}\sum (x_i - \bar{x})(z_i - \bar{z})}$$

$$\hat{\beta}_1^{TSLS} \sim N\left(\beta_1, \frac{\text{Var}((z_i - \mu_z) u_i)}{n \cdot \text{cov}(x_i, z_i)^2}\right)$$

assumptions of IV regression

$$y_i = \beta_0 + \underbrace{\sum_{i=1}^{k} \beta_k x_{ki}}_{\text{endogeneous}} + \underbrace{\sum_{j=k+1}^{r} \beta_j w_{ji}}_{\text{exogeneous}} + u_i$$

1. $E(u_i \mid w_{1i} \ldots w_{ji}) = 0$
2. $(y_i, x_i \ldots w_i \ldots)$ is i.i.d.
3. no extreme outliers
4. IV $z_1 \ldots z_k$ are valid

③ Checking instrument relevance

1) weak (irrelevant) instruments

↳ a way to think about instrument relevance w.r.t. $\hat{\beta}_j$ is like sample size — higher corr., larger variation in X explained and triggered

↳ Statistical inference using TSLS relies on $\hat{\beta}_j^{TSLS}$ being approx. normal in large samples. But if instruments are weak, distribution is non-normal, even in large samples, and distributed more like ratio of two normal variables

⇒ not consistent, even if $n \to \infty$

intuition: $\hat{\beta}_j^{TSLS} = \frac{cov(Z,Y)}{cov(Z,X)}$ ← small

very sensitive

2) checking instrument relevance

idea: regress endogenous X on $Z_1 \dots Z_m$, controlled (OVB) with $W_1 \dots W_n$. Then do an F test that coefficients on $Z_1 \dots Z_m$ are zero. If rejection probability is low, instruments are weak

rule of thumb: first stage F test, F stat < 10 ⇒ weak instrument set

↳ bias of TSLS w weak instrument: $E(\hat{\beta}_j^{TSLS}) - \beta_j \approx \frac{E(\hat{\beta}_1^{OLS}) - \beta_j}{E(F) - 1}$

↳ if $E(F) = 10$, then $\frac{E(\hat{\beta}_j^{TSLS}) - \beta_j}{E(\hat{\beta}_1^{OLS}) - \beta_j} \approx \frac{1}{10-1}$ ⇒ that is, bias of TSLS relative to bias of (bad, since simultaneous) OLS is about 10%, which is generally acceptable

⇒ if F > 10, relative bias is < 10%, vice versa

④ checking instrument validity

under, exactly, over identified — no. of IVs less than, equal to, more than endogeneous variables X_j

checking instrument exogeneity: J test of overidentified instruments

↳ idea: exogeneity of IVs mean that they are uncorrelated with u_i. It logically follows that they should be statistically uncorrelated w our estimate $\hat{u}_i$

1. estimate $Y_i = \beta_0 + \beta_1 X_1 \dots + \beta_{k+1} W_{k+1} \dots$ (true X, rather than predicted $\hat{X}$. Take note! Not $\hat{X}$.) $\Rightarrow$ get $\hat{u}_i = Y_i - \hat{Y}_i$
2. regress $\hat{u}_i$ on all supposedly exogeneous variables $\Rightarrow \hat{u}_i = \delta_1 z_1 \dots + \delta_{m+1} W_{m+1} \dots$
3. compute F stat. that $\delta_1 \dots \delta_{m+1}$ all coefficients on IV, are 0.
4. compute J stat., that is $J = mf$ (m: no. of IV)
5. compute p-value. $J \sim \chi^2(m-k \text{ dof})$ (m − k: no. of IV − no. of X endogenous) → reject $\Rightarrow$ not exogeneous

interpreting the J test → what it does say
- hypothesis about instrument exogeneity (all)

what it does not say
- which instruments are the non-exogeneous ones → use theory / intuition to decide removal

effects of OVB w instrument (corr. z_i & w_i)

$$\beta_1 = \frac{\gamma_1}{\pi_1} = \frac{cov(Y_i, z_i)}{cov(X_i, z_i)} = \frac{cov(z_i, \underbrace{\beta_0 + \beta_1 X_i + \beta_2 W_i + u_i}_{\text{"true" functional form}})}{cov(X_i, z_i)}$$

$$= \frac{\beta_1 cov(z_i, X_i)}{cov(X_i, z_i)}$$ ideally since all other $cov(z_i, -) = 0$

$\Rightarrow$ but if z_i corr. to w_i, then

$$\hat{\beta}_1 = \beta_1 \frac{cov(z_i, X_i)}{cov(X_i, z_i)} + \beta_2 \frac{cov(z_i, W_i)}{cov(X_i, z_i)}$$ biased & inconsistent!

Time series

① fundamental concepts

The jth lag Y_{t-j} is the j^{th} lag. We use the lag operator L to denote:

$$LY_t = Y_{t-1} \iff L^j Y_t = Y_{t-j}$$

Differences $\Delta Y_t = Y_t - Y_{t-1}$. $\Delta = (1-L)$

↳ Δ can be iteratively applied. $\Delta^2 Y_t = (1-L)^2 Y_t$

$$= (1 - 2L + L^2) Y_t$$

$$= Y_t - 2Y_{t-1} - Y_{t-2}$$

↳ note that $\Delta^j Y_t$ if $Y_t - Y_j \Rightarrow$ substitute to L, then solve

2) autocovariance and autocorrelation

j^{th} autocovariance the covariance between Y_t and Y_{t-j}.

j^{th} autocorrelation the correlation between Y_t and Y_{t-j}.

$$\rho_j = corr(Y_t, Y_{t-j}) = \frac{cov(Y_t, Y_{t-j})}{\sqrt{Var(Y_t)\, Var(Y_{t-j})}}$$

3) types of autocorrelation

Geometric decay

$\rho(Y_t, Y_{t-k}) \approx c^k$ for some $c < 1$

slow decay power law

$\rho(Y_t, Y_{t-k}) \approx k^d$ for some $d < 0$

---

② stationarity & ergodicity

stationarity a mathematical property that a stochastic process has when all the random variables of that process are identically distributed.

⇒ for any set $X_t \dots X_{t+n}$ in a stochastic process, all $X_t \dots X_{t+n}$ have the same probability distribution

⇒ the joint distribution of $Y_t, Y_{t-1} \cdots Y_{t-K}$ is identical to $Y_{t-h}, Y_{t-h-1} \cdots Y_{t-h-K}$ for any h. All corr. joint distributions are identical & time displacement invariant

> The joint distributions $(Y_{s+1}, Y_{s+2} \cdots Y_{s+T})$ do not depend on s, regardless of T (though may be different for different T)

covariance stationarity a stochastic process exhibits covariance stationarity if its mean, variance and autocovariance do not depend on t and variables Y_t have finite variance.

> $cov(X_t, X_{t+j})$ is independent of t, at most dependent on j
>
> $E(Y_{t+j}) = \mu_j$ is independent of t, at most dependent on j

stationarity up to order m

↳ strict stationarity is hard in practice

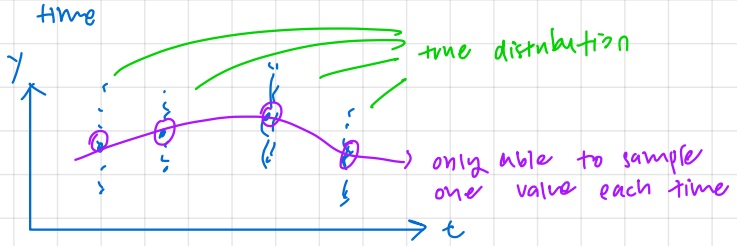
↳ often we simply require that stationarity holds up to a certain point in history

↳ that is, $\forall h \in [0, H]$, $\{Y_{t+i}\}_{i \in [h, T+h]}$, all Y_{t+i} are identically distributed or have the same mean, variance & j^{th} autocovariance in the case of weak (autocovariance) stationarity

ergodicity a stochastic process is said to be ergodic if its statistical properties can be deduced from a single, sufficiently long random sample of the process

why are stationarity & ergodicity important?

↳ with stochastic processes of time, it is by definition that there is no way to collect multiple data points for each variable, since that entails going back in time



↳ stationarity ⇒ Y_t is identically distributed at all t

↳ ergodicity ⇒ w enough time steps, we can estimate the stationary properties!

② Autoregression & stationary processes

(autoregressive distributed lag model)

↳ suppose there is some stable (stationary) relationship between Y_t and its previous p values, and also some other X and their previous k values

↳ that is, the population distribution of Y_t follows

$$Y_t = \beta_0 + \beta_1 Y_{t-1} \dots + \beta_p Y_{t-p} + f_1 X_{t-1} \dots f_K X_{t-K} + u_t$$

(u_t: some error term)

if $E(u_t \mid Y_{t-1}, Y_{t-2} \dots Y_{t-p}, X_{t-1}, \dots X_{t-K}) = 0$, then we can estimate $\hat{Y}_t$ via OLS regression such that $\hat{E}(\hat{Y}_t) = E(Y_t)$.

ergodicity → we set up tuples of $Y_t, Y_{t-1} \dots Y_{t-K}$, and use OLS to estimate $\beta_0, \beta_1 \dots f_1 \dots$

stationarity → no variables dependent on time

least squares assumptions

↳ analogous to those of OLS on cross sectional / panel data

1. $E(u_t \mid Y_{t-1}, Y_{t-2} \dots X_{t-1}, X_{t-2} \dots) = 0$. That is $E(u_t \mid \text{regressors}) = 0$
2. All $Y_t, X_{1t}, X_{2t} \dots$ have a stationary distribution. (No dependence on time needs to be modelled)
3. $(Y_t, X_{1t}, X_{2t} \dots)$ and $(Y_{t-j}, X_{1,t-j}, X_{2,t-j} \dots)$ are independent as j get large. That is, all regressors are ergodic.
4. No perfect multicollinearity or outliers.

③ forecasting errors

forecast error

$$FE = Y_{t+1} - Y_t$$

MSFE

$$MSFE = E\left((Y_{t+1} - \hat{Y}_{t+1})^2\right)$$

RMSFE

$$RMSFE = \sqrt{MSFE}$$

oracle forecast

suppose $\hat{\beta}_0, \hat{\beta}_1 \dots$ are estimated perfectly, with zero variance.

then $\hat{Y}_t = \beta_0 Y_{t+1} + \beta_1 Y_{t-2} \dots + u_t$.

then $E(\hat{Y}_t) = 0$, $Var(\hat{Y}_t) = Var(u_t)$.

⇒ this case is known as the oracle forecast, and this variance and MSFE = $Var(u_t)$ for an oracle model.

1. observe $Var(X) = E(X^2) - E(X)^2$

2. $MSFE = E(Y_T - \hat{Y}_T)^2$

$= E(u_T^2) + E\left(\left((\beta_0 - \hat{\beta}_0) + (\hat{\beta}_1 - \beta_1) Y_{T-1} \dots\right)^2\right)$

$= E(u_T^2)$ for oracle forecast

$= Var(u_t)$ since $E(u_t \mid Y_{T-1} \dots) = 0$

but in the general case, is bias variance trade off.

↑ $MSE = bias^2 + Var + \sigma_u^2$

observe that this is $Var(\hat{\beta}_0 + \dots) = Var(\hat{Y}_T)$

for the same reason that mean = 0, $E(X)^2 = 0$ if unbiased

3) estimation of MSFE

$$MSFE = E(u_T^2) - E\left(\left((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) Y_{T-1} \dots\right)^2\right)$$

$E(u_T^2)$: variance of random error

$E\left(\left((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) Y_{T-1} \dots\right)^2\right)$: accumulated squared error from coefficients (variance of $\hat{Y}$)

recall that ergodicity only maintains for long time horizon! How long? ⇒ could be biased

bias variance tradeoff

$$E\left((y_t + u_t - \hat{y}_t^2)\right) = \left(y_t - E(\hat{y}_t)\right)^2 + E(u_t^2) + E\left((E(\hat{y}) - \hat{y})^2\right)$$

$= bias^2 + noise + variance\ of\ coefficients$

1. approximation by SER — if no. of predictors small compared to sample size, contribution of estimation error will be small

$$\hat{MSFE} \simeq E(u_{t+1}^2) = \sigma_u^2 \approx SER$$

⇒ assumes specific functional form above

2. estimation by final prediction error

↳ assuming errors are homoskedastic

$$\hat{MSFE} = \frac{T+p+1}{T} s_{\hat{u}}^2 = \frac{T+p+1}{T-p-1} \frac{SSR}{T}$$

3. estimate by pseudo OOS

↳ the SER method ignores estimation error
↳ the FPE method requires homoskedasticity
↳ & formulation also require stationarity, since functional form of $\hat{y}$ above based on stationary

↳ pseudo OOS simulates "real time"

Pseudo out-of-sample forecasts are computed using the following steps:

1. Choose a number of observations, P, for which you will generate pseudo out-of-sample forecasts; for example, P might be 10% or 20% of the sample size. Let $s = T - P$.
2. Estimate the forecasting regression using the estimation sample—that is, using observations $t = 1, \ldots, s$.
3. Compute the forecast for the first period beyond this shortened sample, $s + 1$; call this $\tilde{Y}_{s+1|s}$.
4. Compute the forecast error, $\tilde{u}_{s+1} = Y_{s+1} - \tilde{Y}_{s+1|s}$.
5. Repeat steps 2 through 4 for the remaining periods, $s = T - P + 1$ to $T - 1$ (reestimate the regression for each period). The pseudo out-of-sample forecasts are $\tilde{Y}_{s+1|s}, s = T - P, \ldots, T - 1$, and the pseudo out-of-sample forecast errors are $\tilde{u}_{s+1}, s = T - P, \ldots, T - 1$.

$$\widetilde{MSFE}_{POOS} = \frac{1}{P} \sum_{s=T-P+1}^{T} \tilde{u}_{s.}^2$$

P samples for OOS testing, take avg. estimated errors

forecast interval

in practice, it is convenient to assume u_t is normally distributed and heteroskedastic. Under assumptions of stationarity, forecast error is the sum of u_T and estimation error of coefficients.

$$E(Y_+ - \hat{Y}_+) = E(u_T) + E((\beta_0 - \hat{\beta}_0) + (\beta_1 - \hat{\beta}_1) Y_T \cdots)$$

assuming ergodicity, second term is approx. normal by CLT, then if u_t is normal, forecast error is also normal

$$\hat{Y}_T \pm z_\alpha \widehat{RMSE}$$ ← normally distributed

$\widehat{RMSE}$ PPE or $\widehat{RMSE}$ OOS

---

④ choosing order

F-stat similar to multivariate regression

Bayesian information criterion

$$BIC(p) = \ln \frac{SSR(p)}{T} + (p+1) \frac{\ln T}{T}$$ } intuition: SSR ↓, BIC ↓; p ↑, BIC ↑ (penalty)

$$SSR = \sum_{i=1}^{N} \hat{u}_i^2$$ sum of squared residuals

↳ Fit p^{th} model and compute BIC(p). Pick model w lowest BIC.

⇒ $\hat{p}$ which minimizes BIC is a consistent estimator of true lag length, ie. length of lag in underlying population model

⇒ minimising BIC maximises the posterior probability of the model being the true model

Akaike information criterion

↳ but in forecasting, "true model" is irrelevant — all that matters is MSE

$$AIC(p) = \ln \frac{SSR(p)}{T} + (p+1) \frac{2}{T}$$

→ SSR ↓, AIC ↓

→ p ↑, AIC ↑

→ T ↑, AIC ↓

↳ penalty for more p smaller than BIC

⑤ Non-stationarity

deterministic trends — Y_t is a non-random function of time. That is, we can include t as a regressor in the model, and the relationship between Y_t and t is stable

eg. $Y_t = \beta_0 + \alpha_1 t + \alpha_2 t^2 + \beta_1 Y_{t-1} \cdots$

stochastic trends — Y_t is a random function of time. That is, the way Y_t fluctuates is a function of time. i.e. $\text{Var}(Y_t) = f(t)$

⇒ when a trend is present, by definition, $(Y_t, Y_{t-1} \ldots)$ cannot be stationary. That is, the distribution of Y_t must be different at each time step.

random walk model

special case of AR(1) where $\beta_0 = 0$, $\beta_1 = 1$

$Y_t = Y_{t-1} + u_t$, u_t is i.i.d.

1. $E(Y_t) = Y_{t-1} + E(u_t)$, since Y_{t-1} is realised

2. $Y_t = Y_{t-2} + u_{t-1} + u_t \ldots$

$= Y_{t-k} + \sum_{i=1}^{k} u_{t-i} = Y_0 + \sum_{i=1}^{t} u_{t-i}$

$\text{Var}(Y_t) = 0 + t\,\text{Var}(u_t)$ if homoskedastic

3. $\text{cov}(Y_t, Y_{t-k}) = \text{cov}\left(\beta_0 + \sum_{i=1}^{t} u_i, \beta_0 + \sum_{i=1}^{k} u_i\right)$

$= (t-k)\,\text{var}(u_i)$ since $\text{cov}(u_i, u_j) = 0$ for $i \neq j$

random walk with drift

$Y_t = \beta_0 + Y_{t-1} + u_t$, u_t is i.i.d.

1. $E(Y_t) = \beta_0 + Y_{t-1} + E(u_t)$
2. $\text{Var}(Y_t)$ same $= t\,\text{Var}(u_t)$
3. $\text{cov}(Y_t, Y_{t-k}) = (t-k)\,\text{Var}(u_t)$, since $\text{Var}(\beta_0) = 0$

---

⑥ Testing non-stationarity

$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} \ldots + u_t$, u_t is i.i.d.

$= \beta_0 + \beta_1(\beta_0 + \beta_1 Y_{t-2} \ldots) + \beta_2(\beta_0 + \beta_1 Y_{t-3}) \ldots + u_t$

roots are the z that satisfy $1 - \beta_1 z - \beta_2 z^2 \cdots - \beta_p z^p = 0$

p roots. if all $|z| > 1$, stationary.

⟺ sum of coefficients β_1 to $\beta_p = 1$ if not stationary

Dickey Fuller test in the AR(1) model

↳ main idea: not stationary → $\sum_{i=1}^{p} \beta_i = 1$ ⇒ H_0: not stationary, try to reject

$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$

$H_0: \beta_1 - 1 = 0$

$H_1: \beta_1 - 1 < 0$

fit $Y_t - Y_{t-1} = \beta_0 + \beta_1 Y_{t-1} - Y_{t-1} + u_t$

$\Leftrightarrow \Delta Y_t = \beta_0 + (\beta_1 - 1) Y_{t-1} + u_t$

test $\frac{\delta}{SE(\delta)}$, where δ is DF distributed against critical values

Augmented Dickey Fuller test for AR(p)

1. choose p using F test / BIC / AIC
2. ADF test

$\Delta Y_t = \beta_0 + \delta Y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta Y_{t-i}$, $\delta = \sum_{i=1}^{p-1} \beta_i - 1$

$H_0: \delta = \sum_{i=1}^{p} \beta_i - 1 = 0$ → non-stationary

$H_1: \delta < 0$ → not-non-stationary

} test against ADF statistic distribution by $\frac{\delta}{SE(\delta)}$ vs. ADF_α

testing for stochastic & deterministic non-stationarity

↳ we extend the augmented DF test by including function of t

$\Delta Y_t = \beta_0 + (\alpha t)$ + $\delta Y_{t-1} + \sum_{i=1}^{p-1} \gamma_i \Delta Y_{t-i}$, $\delta = \sum_{i=1}^{p-1} \beta_i - 1$

include time term

$H_0: \delta = \sum_{i=1}^{p} \beta_i - 1 = 0$ → non-stationary

$H_1: \delta < 0$ → not-non-stationary

} test against ADF statistic distribution by $\frac{\delta}{SE(\delta)}$ vs. ADF_α