# 1. Two variable linear regression analysis

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = \frac{\sum(x_i-\bar{x})y_i}{n-1}$$

average cross product of deviations of x, around its mean, with y, around its mean

$$corr(x,y) = \rho(x,y) = \frac{cov(x,y)}{\sqrt{V(x)\cdot V(y)}}$$

scale free measure

linear association

- $-1 \le \rho(x,y) \le 1$
- $\rho(x,y) = -1 \Rightarrow$ perfect negative association
  $\rho(x,y) = 1 \Rightarrow$ perfect positive association
  $\rho(x,y) = 0 \Rightarrow$ no linear association

$$V(x) = \frac{\sum(x_i-\bar{x})^2}{n-1}$$

<u>Regression</u> $\Rightarrow$ linear causal association

$y_i = \alpha + \beta x_i + \varepsilon_i$

$y_i$ ↓ dependent; $x_i$ ↓ independent/explanatory

CLRM assumptions

$E(\varepsilon_i|x_i) = E(\varepsilon_i) = 0$

$V(\varepsilon_i|x_i) = \sigma^2$ (homoscedastic)

$cov(\varepsilon_i,\varepsilon_j|x_i) = 0$

$\varepsilon_i|x_i \sim N(0,\sigma^2)$

<u>Estimation</u>

$y_i = \hat{y}_i + e_i = a + bx_i + e_i$

OLS = minimising RSS ( $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$ ) to obtain the parameter estimates

$$b = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$$ (sample slope coefficient)

$a = \bar{y} + b\bar{x}$ (intercept)

$e_i = y_i - (a+bx_i)$ (residuals)

$$s^2 = \frac{\sum_{i=1}^{n} e_i^2}{DoF} = \frac{RSS}{DoF}$$ (estimated variance)

for residuals $(n-2)$

Restrictions: $\sum_{i=1}^{n} e_i = 0$, $\sum_{i=1}^{n} x_i e_i = 0$

OLS properties

Best (minimum variance)

Linear (linear function of the error term)

Unbiased ( $E(a|x) = \alpha$, $E(b|x) = \beta$ )

Estimators

$$V(b|x) = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}$$

$$V(a|x) = \sigma^2 \cdot \frac{\sum_{i=1}^{n} x_i^2}{n\cdot\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

<u>Hypothesis testing</u>

1. $H_0: \beta = \beta_0$
2. $H_1: \beta \neq \beta_0$
3. significance level, $c \Rightarrow \pm t^{c/2}_{DoF}$, $DoF = n - p$

$p$: no of estimated parameters OR no of restrictions on the residuals

$$t = \frac{b-\beta_0}{s_b} \sim t_{DoF},$$

$$s_b = \sqrt{\frac{s^2}{\sum(x_i-\bar{x})^2}}$$

$t < -t^{c/2}_{DoF}$ or $t > t^{c/2}_{DoF} \Rightarrow$ reject $H_0$

$$\boxed{\ln(1+g) \approx g \text{ if } -0.1 < g < 0.1}$$

# 2 Multiple variable regression model

$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + \varepsilon_i$ , $i = 1, 2, \dots, n$ (1)

$\rightarrow DoF = n - \underbrace{(k+1)}_{\text{no of restrictions/ on residuals/ no of para. estimated}}$

CLRM assumptions
1. $E(\varepsilon_i | x_{1i}, \dots, x_{ki}) = 0$
2. $Var(\varepsilon_i | x_{1i}, \dots, x_{ki}) = \sigma^2$
3. $cov(\varepsilon_i, \varepsilon_j | X) = 0$
4. $\varepsilon_i | X \sim N(0, \sigma^2)$

$RSS = \sum_{i=1}^{n} (y_i - a - b_1 x_{1i} - b_2 x_{2i} - \dots - b_k x_{ki})^2$

$\frac{\partial RSS}{\partial a} = 0 \Rightarrow \boxed{\sum_{i=1}^{n} e_i = 0}$ → residuals always sum up to 0, providing there is a residual in the model

$\frac{\partial RSS}{\partial b_k} = 0 \Rightarrow \boxed{\sum_{i=1}^{n} x_{ki} e_i = 0}$ → cov between the residuals and the explanatory variable is 0 (corr = 0)

OLS estimators → partition regression (to estimate the estimators instead of using OLS)

For $x_1$: ① run the regression $y_i = \delta_0 + \delta_2 x_{2i} + \delta_3 x_{3i} + \dots + \delta_k x_{ki} + \varepsilon_{yi}$ by OLS & save the OLS residuals → $\tilde{y}_i = y_i - (d_0 + d_2 x_{2i} + d_3 x_{3i} + \dots + d_k x_{ki})$

② run the regression $x_{1i} = \gamma_0 + \gamma_2 x_{2i} + \gamma_3 x_{3i} + \dots + \gamma_k x_{ki} + \varepsilon_{x_{1i}}$ by OLS & save the OLS residuals → $\tilde{x}_{1i} = x_{1i} - (g_0 + g_2 x_{2i} + \dots + g_k x_{ki})$

③ run the regression $\tilde{y}_i = \alpha + \beta_1 \tilde{x}_{1i} + \varepsilon_i$

$\Rightarrow b_1 = \frac{\sum \tilde{y}_i \tilde{x}_{1i}}{\sum \tilde{x}_{1i}^2}$ , $V(b_1) = \frac{\sigma^2}{\sum_{i=1}^{n} \tilde{x}_{1i}^2}$ , $\sum \tilde{x}_{1i}^2$ = variance in $x_1$ not accounted for by the other variables $x_2$ through $x_k$

For $\sigma^2$: $s^2 = \frac{\sum e_i^2}{DoF} = \frac{RSS}{DoF}$

Properties of OLS estimators
- Best: minimum variance
- Linear: linear function of the error term
- Unbiased: $E(a|X) = \alpha$, $E(b_j|X) = \beta_j$
- Estimators

$b_j | X \sim N(\beta_j, V(b_j))$ and $\frac{RSS^2}{\sigma^2} \sim \chi^2_{n-k}$

## Interpreting coefficients

1. $y_i = \alpha + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$

$\frac{\partial y_i}{\partial x_{1i}} = \beta_1 = \frac{\text{change in } y_i}{\text{unit} \uparrow \text{in } x_1} \Big|_{\text{cet. par.}}$

2. $y_i = \alpha + \beta_1 \ln(x_i) + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$

$\frac{\partial y_i}{\partial(\ln x_{1i}/100)} = \frac{\beta_1}{100} = \frac{\text{change in } y}{1\% \uparrow \text{in } x_1} \Big|_{\text{cet. par.}}$

3. $\ln(y_i) = \alpha + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i$

$100\beta_1 = \frac{\%\text{ change in } y_i}{1 \uparrow \text{in } x_1}$ or $100(e^{\beta_1} - 1) = \frac{\%\text{ change in } y_i}{1 \uparrow \text{in } x_1} \Big|_{\text{cet. par.}}$

4. $\ln(y_i) = \alpha + \beta_1 \ln(x_{1i}) + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i$

$\beta_1 = \frac{\%\text{ change in } y_i}{1\% \uparrow \text{in } x_1} \Big|_{\text{cet. par.}}$

5. $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + \varepsilon_i$ (quadratic rel. in $x_{1i}$)

$\frac{\partial y_i}{\partial x_{1i}} = \beta_1 + 2\beta_2 x_{1i} = \frac{\text{change in } y_i}{1 \uparrow \text{in } x_1} \Big|_{\text{cet. par.}}$ (response varies linearly in $x_{1i}$)

6. $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{1i} z_i + \beta_3 x_{3i} + \dots + \beta_k x_{ki} + \varepsilon_i$

$\frac{\partial y_i}{\partial x_{1i}} = \beta_1 + \beta_2 z_i = \frac{\text{change in } y_i}{1 \uparrow \text{in } x_1} \Big|_{\text{cet. par.}}$ (response varies linearly in $z_i$)

# 3. Dummy Variables

## Additive dummy variables

$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad i = 1, 2, \dots, n$

$\ln(w_i) = \alpha + \beta_1 \, school + \beta_2 \, \underline{female} + \varepsilon_i$

additive dummy → vertically shifts the regression line

$\frac{\partial \ln(w_i)}{\partial \, female} = \beta_2 =$ approx change in the expected wages for a female compared to a male for a given value of school

## Multiplicative dummy variables

$\ln(w_i) = \alpha + \beta_1 \, school + \beta_2 \, female + \beta_3 \, \underline{female \times school} + \varepsilon_i$

multiplicative dummy → rotating the regression line

Male

$E(\ln w) = \alpha + \beta_1 school \quad (1)$

$\beta_1$ = proportionate ↑ in wages for a 1 ↑ in school given you are male

Female

$E(\ln w) = \alpha + \beta_2 + (\beta_1 + \beta_3) school \quad (2)$

$(2) - (1) = \beta_2 + \beta_3 \underbrace{school}_{\text{set school} = 0}$

$\beta_2$ = proportionate change in expected wages for a female compared to a male, providing school = 0

$\beta_3$ = additional proportionate increase in wages for an extra year of schooling (females compared to males)

## Interactive dummy variables

$\ln(w_i) = \alpha + \beta_1 \, female + \beta_2 \, NM + \beta_3 \, female \times NM + \varepsilon_i$

MALE, MAN

$E(\ln w) = \alpha \quad (1)$

FEM, MAN

$E(\ln w) = \alpha + \beta_1 \quad (2)$

$\beta_1$ = expected proportionate ↑ in wages for female compared to male in MAN

MALE, NM

$E(\ln w) = \alpha + \beta_2 \quad (3)$

$\beta_2$ = expected prop ↑ in wages working in NM comp to MAN given you are a male

Fem, NM

$E(\ln w) = \alpha + \beta_1 + \beta_2 + \beta_3 \quad (4)$

$(4) - (3) \rightarrow \beta_1 + \beta_3 \quad (5)$

$[(4)-(3)] - [(2)-(1)] = \beta_3 =$ Additional proportionate change in expected wages for females compared to males who are in NM compared to the same difference for MAN

# 5. Structural change: Chow Tests 1 & 2

$y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad (1) \quad i = 1, \dots, n$

5 CLRM assumptions hold
OLS estimators are BLUE

If the parameters/coefficients are not constant over the entire sample, we have:

→ efficient, possibly biased

$y_i = \beta_0^1 + \beta_1^1 x_{1i} + \beta_2^1 x_{2i} + \dots + \beta_k^1 x_{ki} + \varepsilon_i^1, \quad i = 1, \dots, n_1 \quad (2a)$ ~~[illegible]~~

$y_i = \beta_0^2 + \beta_1^2 x_{1i} + \beta_2^2 x_{2i} + \dots + \beta_k^2 x_{ki} + \varepsilon_i^2, \quad i = n_1+1, \dots, n \quad (2b)$ → potentially unbiased, might omit irrelevant parameters

<u>Structural change (Chow 1 Test)</u> The coefficients in eq (1) may not be constant for the sample of obs over which we wish to estimate the model $i = 1, \dots, n$

$H_0: \beta_0^1 = \beta_0^2, \dots, \beta_k^1 = \beta_k^2$ ⟶ Include the things that one of the models allows to shift and the other one doesn't

$$F = \frac{(RSS^R - RSS^U)/d}{RSS^U / DoF}$$

$RSS^R = RSS^{(1)}$
$RSS^U = RSS^{2a} + RSS^{2b}$
$d = k+1$
$DoF = n - 2(k+1)$

Alternative test
use of dummies

$y_i = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \dots + \delta_k x_{ki} + \gamma_0 D_i + \gamma_1 x_{1i} D_i + \dots + \gamma_k x_{ki} D_i + \varepsilon_i \quad (5)$

$$D_i = \begin{cases} 1 & i = n_1+1, \dots, n \\ 0 & \text{otherwise} \end{cases}$$

$\boxed{D_i = 0}$

$y_i = \delta_0 + \delta_1 x_{1i} + \delta_2 x_{2i} + \dots + \delta_k x_{ki} + \varepsilon_i \quad (2a)$

$\Rightarrow \delta_0 = \beta_0^1, \dots, \delta_k = \beta_k^1$

$\boxed{D_i = 1}$

$y_i = (\delta_0 + \gamma_0) + (\delta_1 + \gamma_1) x_{1i} + \dots + (\delta_k + \gamma_k) x_{ki} + \varepsilon_i \quad (2b)$

$\delta_0 + \gamma_0 = \beta_0^2, \dots, \delta_k + \gamma_k = \beta_k^2$

$H_0: \gamma_0 = \dots = \gamma_k = 0$

$\Rightarrow y_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad i = 1, \dots, n \Rightarrow$ eq. (1)

$$F = \frac{(RSS^R - RSS^U)/d}{RSS^U / DoF}$$

$RSS^R = RSS^{(1)}$
$RSSU = RSS^{(5)} = RSS^{2a} + RSS^{2b}$
$d = k+1$
$DoF = n - 2(k+1)$.

# 6. Misspecification

**Ommission of relevant variables** (eg. has excluded some variables that may be important)

$y_i = \beta_0 + \beta_1 x_i + v_i$ , but the true model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i$

OLS $\Rightarrow$ $\boxed{E(b_1) = \beta_1 + \beta_2 \frac{cov(x_i, z_i)}{var(x_i)}}$ $\Rightarrow$ Biased, unless $\beta_2 = 0$ / $cov(x_i, z_i) = 0$

$\Rightarrow$ standard errors, t-ratios are wrong $\Rightarrow$ Hypothesis testing wrong

**Detecting omitted relevant variables**

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \underbrace{u_i}_{}$ , but the true model is $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$

$u_i = \varepsilon_i + \beta_2 z_i$

A test does not exist

**Incorrect functional form** assumes a linear relationship between $y_i$ and $x$, but it should be (e.g. quadratic)

$y_i = f(x_i; \beta) + \varepsilon_i$ (TRUE) Nonlinear)

$y_i = \beta_0 + \beta_1 x_{1i} + u_i$ (estimated incorrectly linear model)

$y_i = \beta_0 + \beta_1 x_i + u_i$ FALSE

$\Rightarrow$ $y_i = \beta_0 + \exp(\beta_1 x) + \varepsilon_i$ TRUE

$z = \beta_1 x \Rightarrow \phi(z) = 1 + z + \frac{z^2}{2} + R_2$

$\exp(\beta_1 x) = 1 + \beta_1 x + \underbrace{\frac{\beta_1^2 x^2}{2}}_{\beta_2 x^2} + R_2$

Taylor Series

$\phi(z) = \left[\frac{\phi(z_0)}{0!} + \frac{\phi'(z_0)}{1!}(z - z_0)^1 + \dots + \frac{\phi^n(z_0)}{n!}(z - z_0)^n\right] + R_n$

Ex: $\phi(z) = \exp(z)$, $z_0 = 0$

$\Rightarrow \phi(z) = \left[\frac{\exp(0)}{0!} + \frac{\exp(0)}{1!}(z - 0) + \dots + \frac{\exp(0)}{4!}(z - 0)^4\right] + R_4$

$\Rightarrow y_i = (\beta_0 + 1) + \beta_1 x_i + \beta_2 x_i^2 + w_i$ , $w_i = \varepsilon_i + R_2$

RESET TEST $\longrightarrow$ save the residuals

$e = y - (\alpha + \beta_1 x_1 + \beta_2 x_2 + \varepsilon_i)$

Run

$e = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \gamma_1 y^2 + \gamma_2 y^3 + \gamma_3 y^4$

Test $\gamma_1 = \gamma_2 = \gamma_3 = 0$

stop at quadratic

**Inclusion of irrelevant variables**

$y_i = \beta_0 + \beta_1 x_i + u_i$ TRUE

$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + v_i$ ESTIMATED

Partitioned regression:

$R^2_{x_1} \to R^2$ from a reg of $x_1$ on $z$

If $R^2_{x_1} = 0$ then $\Sigma(x_{1i} - \bar{x}_1)^2 = \Sigma(\hat{x}_{1i} - \bar{x}_1)^2$ and there is no cost of estimating the Ⓔ model

t-test of the individual coeff.

# 7. Heteroscedasticity

variance of the error term is non-constant over the sample ( $V(\varepsilon_i|X) \neq \sigma^2$ )

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + \varepsilon_i \quad \Big\} \quad V(\varepsilon_i|X) = E(\varepsilon_i^2|X) = \sigma_i^2$$

The other CLRM assumptions hold

non constant

Assuming

$$V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_i^2 = \alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \dots + \alpha_p z_{pi}$$

$$\Rightarrow \varepsilon_i^2 = \alpha_0 + \alpha_1 z_{1i} + \dots + \alpha_p z_{pi} + \underbrace{\varsigma_i}_{\text{well behaved error process}}$$

$H_0$: $\alpha_1 = \dots = \alpha_p = 0 \quad \rightarrow \quad \varepsilon_i^2 = \alpha_0 + \varsigma_i$ (the variance of the error term depends solely on a constant and therefore is not heteroscedasticity)

$H_1$: $\alpha_j \neq 0$.

As $\varepsilon_i$ is unobserved $\Rightarrow \quad e_i^2 = \alpha_0 + \alpha_1 z_{1i} + \dots + \alpha_p z_{pi} + \varsigma_i$

White's Heteroscedasticity Test (no cross terms)

$z_{1i} = x_{1i}, \dots, z_{ki} = x_{ki}, \; z_{k+1\,i} = x_{1i}^2, \dots, z_{pi} = x_{ki}^2$

ARCH Test

Time series data

$z_{1i} = e_{i-1}^2, \dots, z_{pi} = e_{i-p}^2$

George Căzan

# 8. Errors in variables

measurement errors in one of the variables

$y_i = \alpha + \beta x_i + \varepsilon_i$ $\qquad E(\varepsilon_i | x_i) = 0$

$x_i^* = x_i + u_i$ $\qquad E(u_i | x_i) = 0$

Problem: we don't observe x, we observe $\underset{x \text{ presented to you}}{x^*} = \underbrace{x}_{\text{true } x} + \underbrace{\text{mistakes}}_{\text{assume random}}$

$$y_i = \alpha + \beta(x_i^* - u_i) + \varepsilon_i$$

$$y_i = \alpha + \beta x_i^* + \underbrace{\varepsilon_i - \beta u_i}_{\substack{\text{error term} \\ v_i}}$$

$$E(b_1) = \beta_1 + \frac{\operatorname{cov}(x_i^*, v_i)}{\operatorname{var}(x_i^*)}$$

For OLS to be unbiased and to work, you need

$$\operatorname{cov}(x_i^*, (\varepsilon_i - \beta u_i)) = \operatorname{cov}(x_i^*, \varepsilon_i) - \beta \underbrace{\operatorname{cov}(x_i^*, u_i)}_{>0}$$

$$= -\beta(w) \neq 0$$

$\Rightarrow$ As this is not equal to 0, OLS is biased

UNBIASEDNESS requires that the error term is unrelated to your explanatory variable

---

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

$y_i^* = y_i + u_i \rightarrow$ if the measurement error is on the dependent variable

$$y_i^* - u_i = \alpha + \beta x_i + \varepsilon_i$$

$$y_i^* = \alpha + \beta x_i + (\varepsilon_i + u_i) \rightarrow \text{not a problem}$$

Explanatory variable is not related to the error term

9. RHS Endogenous Variables

whether the RHS variables are taken as given or are endogenously determined / influenced

$Perf = \beta_0 + \beta_1 female + \beta_2 Alex + \beta_3 Att + \varepsilon_{1i}$

$\overset{O}{}$ $\overset{O}{}$ $\overset{U}{}$ $\overset{U}{}$ $\overset{U}{}$

Atted = f( Quality, Time, Interest, Motivation, Ability, Performance, L'SPA)

O = observable
U = unobservable

$Atted = f(Quality, Time, Perf, L'SPA) + \varepsilon_{2i}$

what are the UNOBSERVABLES in $\varepsilon_{1i}$?

$\varepsilon_{1i}$ = Motivation, Interest, Ability, ...

can go either way

(1) $cov(\varepsilon_{1i}, \varepsilon_{2i}) \neq 0$ $(>0)$

$\downarrow$

our guess

$\underset{(1)(2)}{\Rightarrow}$ $cov(Atted, \varepsilon_{1i}) > 0$

(2) $cov(Atted, \varepsilon_{2i}) > 0$

If I increase $\varepsilon_{2i}$, Atted must go up

$\Downarrow$

$E(\varepsilon_{1i} \mid Atted) \neq 0$

If Atted changes, then expected value of $\varepsilon_{1i}$ must change as well

Instrument relevance (the instruments must have a non-zero correlation with the endogenous variables)

$$Atted = \partial_0 + \partial_1 female + \partial_2 A lev + \partial_3 MF + \partial_4 gam + \nu_i$$

$H_0: \partial_3 = \partial_4 = 0$

$\boxed{F > 10}$

You want $\widehat{Atted}$ to explain variation and atted over and above the variables

Instrument exogeneity (the instruments must be unrelated to the error term in the equation of interest)

Take your iv residuals:

$$u^{iv} = \gamma_0 + \gamma_1 female + \gamma_2 A lev + \gamma_3 MF + \gamma_4 game + \eta_i$$

$H_0: \gamma_3 = \gamma_4 = 0$

$$2F \sim \chi^2_{(2-1)}$$

no of instruments

no of instr

no. of RHS problem var

If you reduce the amount of variance in x, you increase the st. error

$$Perf = \gamma_0 + \gamma_1 Alev + \gamma_2 female + \gamma_3 Atted + \gamma_4 e + \eta_i$$

$$e_i = Atted - (\partial_0 + \partial_1 A lev + \partial_2 female + \partial_3 MF + \partial_4 game)$$

$e$ = best guess of $\underbrace{\text{motiv, interest, ability,}}_{\text{UNOBSERVABLES}}$

$H_0: \gamma_4 = 0$

ENDOGENEITY $\Rightarrow$ Hausman - Wu Test

## Endogeneity

suspect $\ln P_i$ to be endogenous

$$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln x_i + \varepsilon_i$$

① choose instruments $z_{1i}$, $z_{2i}$ which should be relevant and exogenous

$$\ln P_i = \delta_0 + \delta_1 \ln x_i + \delta_2 z_{1i} + \delta_3 z_{2i} + u_i$$

② Test for relevance of instruments

Test $\delta_2 = \delta_3 = 0$ $\quad F > 10 \Rightarrow$ RELEVANT

- endogeneity (Hausman Wu) Test :

  save $\hat{u}_i$

  $$\ln Q_i = \beta_0 + \beta_1 \ln P_i + \beta_2 \ln x_i + \beta_3 \hat{u}_i + \varepsilon_i$$

  Test $\beta_3 = 0$ (exogeneity)

③ Test for exogeneity of instruments

$$\ln P_i = \delta_0 + \delta_1 \ln x_i + \delta_2 z_{1i} + \delta_3 z_{2i} + v_i$$

$\widehat{\ln P_i}$ by OLS

$$\ln Q_i = \beta_0 + \beta_1 \widehat{\ln P_i} + \varepsilon_i \quad \text{by OLS}$$

$$\hat{\varepsilon}_i = \ln Q_i - (b_0^{iv} + b_1^{iv} \ln P_i \; b_2^{iv} \ln x_i)$$

$$\hat{\varepsilon}_i = \gamma_0 + \gamma_1 \ln x_i + \gamma_2 z_{1i} + \gamma_3 z_{2i} + v_i$$

TEST $\quad H_0: \gamma_2 = \gamma_3 = 0$

$\quad\quad\quad H_1: \gamma_j \neq 0 \quad j = 2, 3$

$$J = 2F \sim \chi^2_{(1)} \rightarrow \text{no of instruments} - \text{no of instrumented variables}$$

# 10. Normality
(non-normality of the error terms)

$y_i = \alpha + \beta x_i + \varepsilon_i$

$b = \beta + \Sigma w_i \varepsilon_i \quad , \quad w_i = \frac{x_i - \bar{x}}{\Sigma (x_i - \bar{x})^2}$

$\downarrow$

OLS estimator

But $\varepsilon_i$ is for example a UNIFORM distribution rather than normal.

If you add up enough distributions together the resulted distribution will always be approx. normal (CLT)

$$\Rightarrow \quad b \overset{a}{\sim} N\left(\beta, \sigma^2 / \Sigma (x_i - \bar{x})^2\right) \text{ if } n > 30$$

Detecting non-normality

Testing the skewness and excess kurtosis of these residuals

$m_j = n^{-1} \sum_{i=1}^{m} e_i^j \quad j = 1,2,3,4$

$\text{skewness} = m_3 / m_2^{3/2}$

$\text{kurtosis} = K = m_4 / m_2^2$

Jarque-Bera test of normality: $\quad JB = \frac{n}{6} \left[ S^2 + (K-3)^2 / 4 \right]$

The most usual cause of non-normality is an outlier in the data set, which shows up as apparent symmetry

# 11. Multicollinearity

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

OLS estimation of above

$$V(b_1) = \sigma^2 / \Sigma(\tilde{x}_{1i} - \bar{\tilde{x}}_1)^2$$

where $\tilde{x}_{1i}$ are the residuals from a regression

Partition regression is: you are interested in the influence of $x_2$

- extract the influence of $x_2$ from the y variable by regressing y on $x_2$ and taking the residuals
- extract the influence of $x_2$ from $x_1$ by regressing $x_1$ on $x_2$ and taking the residuals
- do a regression of the residuals from y on the residuals from $x_1$ and that's OLS

Perfect multicollinearity is when one variable is a perfect function of another variable

↓

should never arise (indicates you included the same variable twice)

imperfect multicollinearity highly correlated variables (close to unity)

↓

can estimate coefficients

Problem

If $x_1$ and $x_2$ are perfectly correlated ⇒ the residuals would be 0 (horizontal line)

If $x_1$ and $x_2$ are imperfectly correlated ⇒ the residuals are tiny

⇓

- the variances are big (big st. errors)
- the t-ratios are small

Detecting multicollinearity
- calc. simple corr. between all expl. variables (concerned if $\rho > 0.85$)

Solutions:

1. Do nothing
2. Drop the variable which is highly correlated
3. Increase sample size
4. (a) Principle component Analysis
   (b) In time series, take the first differential

] Transform the collinear variables by