Revision Notes by Sally Yang

PRINCIPLES OF ECONOMETRICS

LENT TERM

DR MARCIA SCHAFGANS
LONDON SCHOOL OF ECONOMICS 2020/21

LINEAR ALGEBRA

$(A^{-1})' = (A')^{-1}$

$(AB)' = B'A'$

$(AB)^{-1} = B^{-1}A^{-1}$ (if A and B are both invertible)

$|A| = |A'|$

$|AA^{-1}| = |A||A^{-1}| = 1$ (if A is invertible)

$\Rightarrow |A^{-1}| = \frac{1}{|A|}$

$\Rightarrow |A| \neq 0$

Let X be $n \times k$ matrix. $X = \begin{pmatrix} x_{11} \cdots x_{1k} \\ \vdots \\ x_{n1} \cdots x_{nk} \end{pmatrix} = \begin{pmatrix} \longmapsto x_1' \longmapsto \\ \longmapsto x_j' \longmapsto \\ \longmapsto x_n' \longmapsto \end{pmatrix}$, $x_j' = (x_{j1} \cdots x_{jk})$

Then $X'X = \sum_{j=1}^{n} x_j x_j' = \sum_{j=1}^{n} \begin{pmatrix} x_{j1} \\ \vdots \\ x_{jk} \end{pmatrix} (x_{j1} \cdots x_{jk})$ ☞ $X'X$ is symmetric and positive semidefinite. PSI

Positive definite if columns of X lin. indep.

☆ $Z'X = \sum_{j=1}^{n} z_j x_j'$. Letter in same order but transpose second

Usefulness: $\frac{X'X}{n} = \frac{\sum_{j=1}^{n} x_j x_j'}{n}$ is sample average

☆ $(X'X)^{-1}X'y \neq (X^{-1})(X')^{-1}X'y = X^{-1}y$ because X may not be invertible

☆ $X'y = \sum_{j=1}^{n} x_j y_j'$ ← Scalar

$k \times n$ $n \times 1$ ↑ vector

LINEARLY INDEPENDENT $x_1, \ldots, x_k$ are linearly independent if

The linear combination $\alpha_1 x_1 + \ldots + \alpha_k x_k = 0 \Longrightarrow \alpha_1 = \ldots = \alpha_k = 0$

ORTHOGONAL

$x \perp y$ if $\underline{x} \cdot \underline{y} = \underline{x}'\underline{y} = 0$ (for vector)

Matric X is orthogonal if $X' = X^{-1}$. So $XX' = XX^{-1} = I \Longrightarrow |X|^2 = 1 \Rightarrow |X| = \pm 1$

RANK $\rho(A)$ = Number of linearly independent rows

EIGENVALUES $\underline{x}$ (non-zero) is an eigenvector of matrix A if $A\underline{x} = \lambda\underline{x}$ (Eigenvalue $\lambda \in \mathbb{R}$)

Find eigenvalues: $|A - \lambda I| = 0$

DIAGONALISING After finding eigenvectors, eigenvalues, we can write $A = PDP^{-1}$ hence $P^{-1}AP = D$

$D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$, $P = \underset{\text{eigenvectors}}{\begin{pmatrix} \top & & \top \\ \underline{x}_1 & \cdots & \underline{x}_n \\ \bot & & \bot \end{pmatrix}}$

$|A| = |D| = \prod_{i=1}^{n} \lambda_i$

ORTHOGONAL DIAGONALISING $P = P^{-1}$, so $A = PDP^{-1} = PDP'$ or $D = P'AP$

Works for any symmetric matrix A. If A is also positive definite then $\exists Q$ s.t. $A = QQ'$

DEFINITENESS A symmetric matrix A is positive definite if $\underset{n \times 1}{\underline{x}} A \underline{x}' > 0 \ \forall \underline{x} \neq 0$

	Eigenvalues	$\forall \underline{x}$, $\underline{x}A\underline{x}'$
+ve def	>0	>0
+ve semidef	$\geqslant 0$	$\geqslant 0$
−ve def	<0	<0
−ve semidef	$\leqslant 0$	$\leqslant 0$
indef	$\exists >0$, $\exists <0$	$\exists >0$, $\exists <0$

TRACE $\text{tr}(A)$ is sum of diagonal elements $\sum_{i=1}^{n} a_{ii}$

- $\text{tr}(A) = \text{tr}(A')$
- $\text{tr}(A+B) = \text{tr}(A) + \text{tr}(B)$
- $\text{tr}(AB) = \text{tr}(BA)$

  so $\text{tr}(ABCD) = \text{tr}(DABC) = \text{tr}(CDAB) = \text{tr}(BCDA)$

IDEMPOTENT A is idempotent if $AA = A \Rightarrow A^k = AAA\cdots A = A$

- A is always diagonalisable and its eigenvalues are either 0 or 1
- If A is idempotent, $\text{tr}(A) = \rho(A)$ ——

Invertible	Square, $\lvert A\rvert \neq 0$, full row and column rank (no rows/columns of 0s) linearly indep. row/columns, A' invertible. $\lambda_i \neq 0 \; \forall i$
Diagonal	Square, Off-diagonal 0s
Symmetric	Square, $A = A'$
Orthogonal	Square, $AA' = A'A = I$. $A = A^{-1}$
Idempotent	Square, $AA = A$

I satisfies all

VECTOR DIFFERENTIATION

$f: \mathbb{R}^n \to \mathbb{R}$ (input: $n\times 1$ vector; output is scalar)

$$\underset{n\times 1}{\frac{\partial f}{\partial \underline{x}}} = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}_{n\times 1}$$

PROPERTIES

- If $\underline{a}$ and $\underline{x}$ are $n\times 1$ vectors, $\frac{\partial}{\partial \underline{x}} \underline{a}'\underline{x} = \underline{a}$ ($\underline{a}'\underline{x}$: linear form (also $\underline{a}\cdot\underline{x}$))
- $\frac{\partial}{\partial \underline{x}} \underset{n\times 1}{\underline{x}'} \overset{n\times n}{A} \underline{x} = A\underline{x} + A'\underline{x}$
  $= 2A\underline{x}$ when A is symmetric
- $\frac{\partial}{\partial A} x'Ax = xx'$
- $\frac{\partial}{\partial x} x'Ay = Ay$

SQUARE VECTORS

NOT the inner product $\underline{m}'\underline{m}$ (gives scalar) but the outer product $\underline{m}\underline{m}'$ ($n\times n$ matrix!)

MULTIVARIATE STATISTICS

A random vector/matrix $\underset{\sim}{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$ / $X = \begin{pmatrix} x_{11} & \cdots & x_{nk} \end{pmatrix}$ is a vector/matrix where elements are random variables

EXPECTATIONS

$E(\underset{\sim}{x}) = \begin{pmatrix} E(x_1) \\ \vdots \\ E(x_k) \end{pmatrix}$ so $E(\underset{\sim}{\varepsilon}) = 0$ implies $\begin{pmatrix} E(\varepsilon_1) \\ \vdots \\ E(\varepsilon_n) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}$

Rules of expectation work normally: $E(\underset{\text{constants}}{\underset{\sim}{a}} + \overset{\text{random}}{B\underset{\sim}{x}}) = \underset{\sim}{a} + BE(\underset{\sim}{x})$

VARIANCE

$Var(\underset{n\times 1}{\underset{\sim}{x}}) = E\left[(\underset{\sim}{x} - E\underset{\sim}{x})(\underset{\sim}{x} - E\underset{\sim}{x})'\right] = \begin{pmatrix} Var(x_1) & & \\ Cov(x_2, x_1) & \ddots & \\ \vdots & & \\ Cov(x_m, x_1) & \cdots & Var(x_n) \end{pmatrix}$ always positive semidefinite!

Rules of variance: $Var(a + Bx) = Var(Bx) = Bvar(x)B' = B^2 Var(\underset{\sim}{x})$ if B is a scalar

CONDITIONAL EXPECTATIONS

For every x, we have this object $y|X=x = Z$ which is a random variable

$P(Z \le w) = F_Z(w) = P(y|X=x \le w) = P(y \le w | X=x) = \underset{y|X=x}{F}(w|X=x)$ are all equivalent

Density of $f_{Y|X=x}$ is the density of Y conditional on $X=x$

$E(y|X=x) = \int y f(y|X=x)dy$ is a function of x. Same for $Var(Y|X=x)$

PROPERTIES

$E(f(X) + Y|X) = E(f(X)|X) + E(Y|X)$
$= f(X) + E(Y|X)$ since conditioning on X allows us to think of X as non-random anymore.

$Var(f(X) + Y|X) = Var(Y|X)$
$Var(f(X)Y|X) = f(X)E(Y|X)f(X)'$

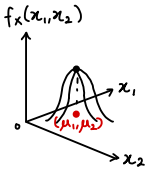
Difference between $E(y|X=x)$ and $E(y|X)$

$E(y|X=x)$ is a function of x, and is not random.
If we let $E(y|X=x) = h(x)$, then $E(y|X) = h(X)$.
So $E(y|X)$ is a random variable that takes the values $E(y|X=x)$ with density function $f_X(x)$
Same for $Var(Y|X=x) = g(x) \Rightarrow Var(Y|X) = g(X)$

MULTIVARIATE NORMAL DISTRIBUTION

$\underset{n\times 1}{x} \sim N(\underset{n\times 1}{\mu}, \underset{n\times n}{Var(x)})$

pdf: $f_x(x) = \frac{1}{\sqrt{\det(2\pi Var(x))}} e^{-\frac{1}{2}(x-\mu)' Var(x)' (x-\mu)}$



If $\underset{m\times 1}{a}$ and $\underset{m\times n}{B}$ are fixed.

$a + Bx \sim N(a + B\mu, BVar(x)B')$

just let $\underset{1\times n}{B} = (0 \cdots 0 \overset{j^{th}\ entry}{1} 0 \cdots 0)$

$\underset{\sim}{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \sim N(\mu, Var(\underset{\sim}{Z}))$ (multivariate normal) implies its elements are univariate normal $Z_j \sim N_1(\mu_j, (Var(Z))_{jj})$

Converse true only if $Z_1, \ldots, Z_n$ are independent

UNCORRELATED ⇒ INDEPENDENCE

If $\underset{\sim}{Z} = \begin{pmatrix} Z_1 \\ \vdots \\ Z_n \end{pmatrix} \sim N(\mu, Var(\underset{\sim}{Z}))$, and $Z_1, \ldots, Z_n$ are identically distributed and all uncorrelated ($cov = 0$)

then it implies independence

REVIEW

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

ASSUMPTION

- Mean independence $E(\varepsilon|X) = 0$

$\Rightarrow E(Y|X) = E(\alpha + \beta X + \varepsilon|X) = \alpha + \beta X + E(\varepsilon|X) = \alpha + \beta X$

$\Rightarrow \beta = \frac{\partial E(Y|X)}{\partial X}$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

"partial effect": effect of X_1 on Y holding all else (X_2) constant

Adding more regressors $\uparrow R^2$

ASSUMPTION

- Mean independence $E(\varepsilon|X_1, X_2) \equiv E(\varepsilon|X) = 0$

$\Rightarrow E(Y|X) = E(\alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon|X) = \alpha + \beta_1 X_1 + \beta_2 X_2 + E(\varepsilon|X) = \alpha + \beta_1 X_1 + \beta_2 X_2$

$\Rightarrow \beta = \frac{\partial E(Y|X)}{\partial X_1}$

SAMPLING Obtain a sample $\{(Y_i, X_i)\}_{i=1}^n$ from population

- Effectively drawing from (X, ε) joint dist.
- In cross-section, maybe can assume (ε_i, X_i) are iid drawings from pop^n
- In time-series, (ε_i, X_i) and (ε_j, X_j) usually related
- X_i, ε_i commonly correlated (OVB)

OLS

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{b_0, b_1}{\arg\min} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

$\hat{\beta}_0, \hat{\beta}_1$: estimators

(realisation of them for particular samples are estimates)

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad \hat{\beta}_1 = \frac{Cov(Y_i, X_i)}{Var(X_i)}$$ ← sample cov / ← sample var

RESIDUALS $\hat{\varepsilon}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$ ☆ NOT same as error ε_i

FITTED VALUES $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

PROPERTIES OF SAMPLING DIST OF $\hat{\beta}_1$

- Unbiasedness $E(\hat{\beta}_1) = \beta_1$ ("correct on average")
- Standard error $SE(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)}$

↑ estimated var of $\hat{\beta}_1$

TERMS

Sample $\{(x_i, y_i)\}_{i=1}^{n}$ $\quad x_i = \begin{pmatrix} x_{1i} \\ \vdots \\ x_{ki} \end{pmatrix}$ Individual characteristics of person i

Joint density of n observations of the same variable, $f(x_1, \ldots, x_n) \equiv f(\underset{\text{vector}}{x})$

Marginal density $f(x_1) = \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)\, dx_2 \ldots dx_n$

Conditional density $f(\varepsilon|x) = \frac{f(\varepsilon, x)}{f(x)}$ ← joint / ← marginal

Under independence,

$f(\varepsilon, x) = f(\varepsilon|x) f(x) = f(\varepsilon) f(x)$ product of marginals

$E(\varepsilon|x) = \int_{-\infty}^{\infty} \varepsilon f(\varepsilon|x)\, d\varepsilon = \int_{-\infty}^{\infty} \varepsilon f(\varepsilon)\, d\varepsilon = E(\varepsilon)$

Covariance matrix of $\varepsilon = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ is $Var(\varepsilon) = E\underset{n\times 1}{\left[(\varepsilon - E(\varepsilon))\underset{1\times n}{(\varepsilon - E(\varepsilon))'}\right]}$ (matrix with Var on diagonal and cov on off-diagonal), $\varepsilon - E(\varepsilon) = \begin{pmatrix} \varepsilon_1 - E(\varepsilon_1) \\ \vdots \\ \varepsilon_n - E(\varepsilon_n) \end{pmatrix}$

$$= \begin{pmatrix} E\left((\varepsilon_1 - E(\varepsilon_1))^2\right) & E\left((\varepsilon_2 - E(\varepsilon_2))(\varepsilon_1 - E(\varepsilon_1))\right) & \vdots \\ \vdots & E\left((\varepsilon_2 - E(\varepsilon_2))^2\right) & \\ E\left((\varepsilon_1 - E(\varepsilon_1))(\varepsilon_n - E(\varepsilon_n))\right) & \cdots & E\left((\varepsilon_n - E(\varepsilon_n))^2\right) \end{pmatrix}$$

vector of constants, vector of variables

$E(c + a'\underset{\text{vector of variables}}{x}) = c + a' E x \quad \left(\sum_{i=1}^{n} a_i x_i = a'x\right)$

$E(c + Ax) = c + AE(x)$

vector of constants, matrix of constants

$$Var\left(c + \sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i^2 Var(x_i) + \sum_{i \neq j}^{n} \sum_{i=1}^{n} a_i a_j Cov(x_i, x_j)$$

$Var(c + a'x) = a' Var(x) a$

$Var(c + Ax) = A\, Var(x) A'$

↑ Covariance matrix!

CLASSICAL (GAUSS-MARKOV) LINEAR REGRESSION

$y = X\beta + \varepsilon$, $y = \underset{n\times 1}{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}$, $X = \underset{n\times k}{\begin{pmatrix} x_{11} & \cdots & x_{1k} \\ \vdots & & \vdots \\ x_{n1} & \cdots & x_{nk} \end{pmatrix}} = (\underline{x}_1, \ldots, \underline{x}_k)$, $\beta = \underset{k\times 1}{\begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}}$, $\varepsilon = \underset{n\times 1}{\begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}}$

n observations
k variables
k+1 parameters (+constant)

- for each observation i, $y_i = x_i'\beta + \varepsilon_i$, $\beta = \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_k \end{pmatrix}$ $x_i = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{ik} \end{pmatrix}$

ASSUMPTIONS

① True model is linear in parameters. $y = X\beta + \varepsilon$, $E(\varepsilon) = 0$
- If $E(\varepsilon) \neq 0$, build it into systemic part of eqn, adding new vars, make $=0$

② No perfect multicollinearity (e.g. dummy variable trap).
- X has full column rank $k \leq n$, lin. indep. rows; $X'X$ invertible in deriving OLS estimator

③ $E(\varepsilon|X) = 0$. Elab below

④ Homoscedasticity, no auto-correlation $\Rightarrow Var(\varepsilon|X) = E(\varepsilon\varepsilon'|X) = \sigma^2 I$

$Var(\varepsilon_i|X) = \sigma^2 \ \forall i$ $Cov(\varepsilon_i, \varepsilon_j|X) = 0 \ \forall i \neq j \Rightarrow Cov(y_i, y_j|X) = 0$

- Reduces to $Var(\varepsilon) = \sigma^2 I$ if X is fixed (non-stochastic)
- Typically works for cross-section but not time series
- Heteroscedastic if $Var(\varepsilon_i|X)$ is some function of X

⑤ Normality $\varepsilon|X \sim N(0, \sigma^2 I)$ A1-4 is Gauss Markov. A1-5 (GM + Normality) is classical lin reg model
- A5 implies A3 and A4; also implies $\varepsilon \sim N(0, \sigma^2 I)$ since $E(\varepsilon|X)$, $Var(\varepsilon|X)$ don't depend on X
  - $\hat{\beta}|X \sim N_k(\beta, \sigma^2(X'X)^{-1}) \Rightarrow \hat{\beta}_j|X \sim N_1(\beta_j, \sigma^2(X'X)^{-1}_{jj}) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2(X'X)^{-1}_{jj}}}|X \sim N(0,1) \Rightarrow \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2(X'X)^{-1}_{jj}}} \sim N(0,1)$
  (marginals normally distributed too; t-testing!)
- Even without A5, if n is large, we can use CLT to say A5 holds "approximately"

ZERO CONDITIONAL MEAN $E(\varepsilon|X) = 0$

- Regressors strictly exogenous.
- Ensures X and ε uncorrelated, $Cov(x_{is}, \varepsilon_j) = 0$
- A1+A3 $\Rightarrow E(y|X) = X\beta$ and $\frac{\partial E(y|X)}{\partial x_j} = \beta_j$.
- Need for assumption arises from regressors being stochastic (random). So the x should be conditioned on.
  - If X is fixed (non-stochastic), then it just becomes $E(\varepsilon) = 0$ and no need for A3
- If samples are drawn independently (ε_i not correlated to $x_j \ \forall i \neq j$), A3 reduces to $E(\varepsilon_i|x_i) = 0$
- Independence $\overset{f(y|X=x)=f(y)}{\Rightarrow}$ Mean independence $\overset{(Cov(X,Y))}{\Rightarrow}$ Uncorrelatedness 2020 IRDAP 3
  $E(\xi x) = E(\xi)E(x) = 0$ (Typically too strong for reality) $E(\varepsilon_i|X) = E(\varepsilon_i)$ $= E(YX) - E(Y)E(X)$ $= E(Y|X) = E(XE(Y|X))$ $= E(X)E(Y)$ $Cov(x_{is}, \varepsilon_j) = 0$
- Uncorrelatedness $\Rightarrow$ Mean independence if we have multivariate normal $\underset{\sim}{z}$ with uncorrelated elements (see help lecture)
  - Multivariate normal $\varepsilon|X$ (A5) + uncorrelated elements $Cov(\varepsilon_i, \varepsilon_j) = 0 \ \forall i \neq j$ (A4)
- $Y = X\beta + \varepsilon$, $E(\varepsilon) = 0$ (A1) and $E(\varepsilon|X) = 0$ (A3) $\Leftrightarrow E(Y|X) = X\beta$

LAW OF ITERATED EXPECTATIONS $E(h(x,y)) = E(E(h(x,y)|x))$

assuming these expectations exist
x, y must both be random

OLS ESTIMATOR $\hat{\beta} = \begin{pmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{pmatrix} = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - b_1 x_{i1} - b_2 x_{i2} - \ldots - b_k x_{ik})^2 = \underset{b}{\operatorname{argmin}} (y - Xb)(y - Xb)'$

FOCs: $\begin{cases} \sum_{i=1}^{n} x_{i1}\hat{\varepsilon}_i = 0 \\ \vdots \\ \sum_{i=1}^{n} x_{ik}\hat{\varepsilon}_i = 0 \end{cases}$, $\hat{\varepsilon}_i = y_i - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \ldots - \hat{\beta}_k x_{ik}$

$\Leftrightarrow \underset{k\times n}{X'} \underset{n\times 1}{\hat{\varepsilon}} = 0 \Leftrightarrow X'(y - X\hat{\beta}) = 0 \Leftrightarrow \hat{\beta} = (X'X)^{-1}X'y$
(exists by A2 (otherwise β not identified))

PROPERTIES OF $\hat{\beta}$

MINIMISES RESIDUAL SUM OF SQUARES

RSS is $S(b) = \sum_{i=1}^{n}(y_i - b_1x_{i1} - \dots - b_kx_{ik})^2 = (y-Xb)'(y-Xb) = y'y - \underbrace{2b'X'y}_{\text{scalar}} + b'X'Xb$

When $S(b)$ is at minimum,

Proof

$$\frac{\partial S(b)}{\partial b} = \begin{pmatrix} \frac{\partial S(b)}{\partial b_1} \\ \vdots \\ \frac{\partial S(b)}{\partial b_k} \end{pmatrix} \stackrel{\text{set}}{=} -2X'y + 2X'Xb = 0 \Rightarrow \hat{\beta} = (X'X)^{-1}X'y$$

METHOD OF MOMENTS ESTIMATOR

By A1,3, $E(x_{is}\varepsilon_i) = 0$ and $E(x_{is}\varepsilon_i) = E(x_{is}(y_i - x_i'\beta))$, $s = 1, \dots, n$

In MME, you replace the theoretical expected value with sample average $\frac{1}{n}\sum_{i=1}^{n}x_{is}\hat{\varepsilon}_i = 0$ with $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}_{MME}$

but this coincides with OLS requirements so $\hat{\beta}_{MME} = \hat{\beta}_{OLS}$

MME:
Step 1: replace E with $\frac{1}{n}\sum_{i=1}^{n}$
Step 2: put hats on all parameters, replace ε_i with $\hat{\varepsilon}_i$

- ✯ $E(x_{is}\varepsilon_i) = 0$
- ✯ "Existence of moments" If n^{th} moment exists, $1 - n-1^{th}$ moment exists

UNBIASED $E(\hat{\beta}) = \beta$

finite sample property
i.e. true for any n

Proof

$$E(\hat{\beta}) = E[(X'X)^{-1}X'y] = E[(X'X)^{-1}X'(X\beta+\varepsilon)] = E(\beta) + E[(X'X)^{-1}X'\varepsilon] = \beta + E[(X'X)^{-1}X'\varepsilon] = \beta!$$

($y = X\beta + \varepsilon$ by A.1 (True model))

If X is fixed, $E[(X'X)^{-1}X'\varepsilon] = (X'X)^{-1}X'E(\varepsilon) = 0$

If X is stochastic,

under full (too strong) independence, $E[(X'X)^{-1}X'\varepsilon] = E[(X'X)^{-1}X']E(\varepsilon) = 0$

under mean independence, $E[(X'X)^{-1}X'\varepsilon] \underset{\text{LIE}}{=} E(E[(X'X)^{-1}X'\varepsilon|X]) = E((X'X)^{-1}X'E(\varepsilon|X)) = 0$

$E(\varepsilon_i|X) = E(\varepsilon_i|x_i)$ if the sample $\{(y_i, x_i) : i = 1, \dots, n\}$ is random

VARIABILITY (THEOREM) $Var(\hat{\beta}|X) = \sigma^2(X'X)^{-1}$

finite sample property

Proof

$$Var(\hat{\beta}|X) = Var(\beta + (X'X)^{-1}X'\varepsilon|X) = (X'X)^{-1}X'\,Var(\varepsilon|X)\,[(X'X)^{-1}X']' = (X'X)^{-1}X'\sigma^2 I X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

(by homoscedasticity, no autocorrelation)

$E[Var(\hat{\beta}|X)] = Var(\hat{\beta})$ PS2

✯ Implies that in bivariate regression, $Var(\hat{\beta}_2|X) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \equiv \frac{\sigma^2}{ns_x^2}$ ← sampling variability; ↑ Sample size (should be n-1 for unbiasedness but wtv)

✯ In multivariate regression, $Var(\hat{\beta}_j|X) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2(1 - r^2_{x_j, x_{-j}})} \equiv \frac{\sigma^2}{ns^2_{x_j}(1 - r_j^2)}$

r_j^2: sample correlation (deg of collinearity of regressors) obtained from regressing x_j on all other regressors

$s^2_{x_j} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2$

regressors cannot be highly collinear

We can ↑ precision by

- ↑ sample size,
- ↑ variability of regressors,
- ↓ variance of errors (σ^2),
- ↓ sample correlation r

ESTIMATOR FOR VAR($\hat{\beta}$)

The unbiased estimator for σ^2, $S^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$ (or n-k+1 (+ intercept)), $E(S^2) = \sigma^2$

proof: $\hat{\varepsilon} = M\varepsilon$

$$\widehat{Var}(\hat{\beta}) = S^2(X'X)^{-1}. \quad SE(\hat{\beta}_j) = \sqrt{\widehat{Var}(\hat{\beta}_j)}$$

← j^{th} diagonal element of $\widehat{Var}(\hat{\beta})$

$\hat{\beta}$ IS BLUE

finite sample property

BEST LINEAR UNBIASED ESTIMATOR (BLUE)

i.e. $\hat{\beta}$ has lower variance than any other estimator $\tilde{\beta}$ that is unbiased ($E(\tilde{\beta})=\beta$)

Proof

Let an unbiased estimator be $\tilde{\beta}=Cy=CX\beta+C\varepsilon$.

Then $E(\tilde{\beta})=\beta \Leftrightarrow E(CX\beta+C\varepsilon)=\beta \Leftrightarrow CX\beta+CE(\varepsilon)=\beta \Leftrightarrow CX\beta=\beta \Leftrightarrow CX=I$

($E(\varepsilon)=0$; assume X is fixed)

☆ C^{-1} and X^{-1} may not exist !!!

Thus, $\tilde{\beta}=CX\beta+C\varepsilon=I\beta+C\varepsilon=\beta+C\varepsilon$

Then $Var(\tilde{\beta})=Var(\beta+C\varepsilon)=C\,Var(\varepsilon)C'=C\sigma^2IC'=\sigma^2CC'$ (A4)

$Var(\tilde{\beta})-Var(\hat{\beta})=\sigma^2(CC'-(X'X)^{-1})=\sigma^2(CC'-I(X'X)^{-1}I')=\sigma^2(CC'-CX(X'X)^{-1}X'C')=\sigma^2C(I-X(X'X)^{-1}X')C'$ (PSI $=M$, symmetric idempotent)

$=\sigma^2CMM'C'=\sigma^2D'D\geq 0$ (positive semidefinite by PSI)

$Var(\tilde{\beta})\geq Var(\hat{\beta}) \Leftrightarrow \hat{\beta}$ is efficient

COROLLARIES

1. Any linear transformation of $\hat{\beta}$ is also the BLUE of that linear transformation of β (BLUE of $c'\beta$ is $c'\hat{\beta}$)
2. Any particular element of $\hat{\beta}$, $\hat{\beta}_j$, estimates β_j at least as efficiently as any other linear unbiased estimator

$\hat{\beta}$ IS MVUE (minimum variance unbiased estimator)

Assume normality ($\varepsilon|X\sim N(0,\sigma^2I)$; if not, CLT) to obtain sampling distribution for $\hat{\beta}$ and enable hypothesis testing

Then by A1-5, $\hat{\beta}|X\sim N(\beta,\sigma^2(X'X)^{-1})$

and $\hat{\beta}_j|X\sim N(\beta_j,\sigma^2c_{jj})$, $c_{jj}=[(X'X)^{-1}]_{jj}$

$\hat{\beta}$ IS MLE Makes OLS best amongst linear and non-linear unbiased estimators ($\Rightarrow$ MVUE)

Given normality assumption A5, $y|X\sim N(X\beta,\sigma^2I)$, then by independence, joint distribution of $y_1,\dots,y_n$ is

$$L(\alpha,\beta,\sigma^2)\equiv f(y_1,\dots,y_n;\beta,\sigma^2)=\prod_{i=1}^{n}\frac{1}{2\pi\sigma^2}e^{-\frac{1}{2\sigma^2}(y_i-x_i\beta)^2}$$

$$\log L(\alpha,\beta,\sigma^2)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-x_i\beta)^2=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}S(\hat{\beta})$$

(maximise $-S(\hat{\beta})\Leftrightarrow$ minimise $S(\hat{\beta})$ (OLS))

To maximise log-likelihood, FOCs:

$$\frac{1}{2\hat{\sigma}^2_{MLE}}\left(\frac{\partial S(\hat{\beta})}{\partial\hat{\beta}}\right)=0 \leftarrow \text{Same as OLS, so } \hat{\beta}=\hat{\beta}_{MLE}$$

$$-\frac{n}{2\hat{\sigma}^2_{MLE}}+\frac{1}{2\hat{\sigma}^4_{MLE}}S(\hat{\beta})=0 \Rightarrow \hat{\sigma}^2_{MLE}=\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2=\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n}=\hat{\sigma}^2_{MME}$$

GEOMETRIC ASPECTS OF OLS

FOC of OLS is $X'\hat{\varepsilon}=0$ (so residuals and regressors are orthogonal)
If our model has an intercept, X will have a column of 1s. Then:

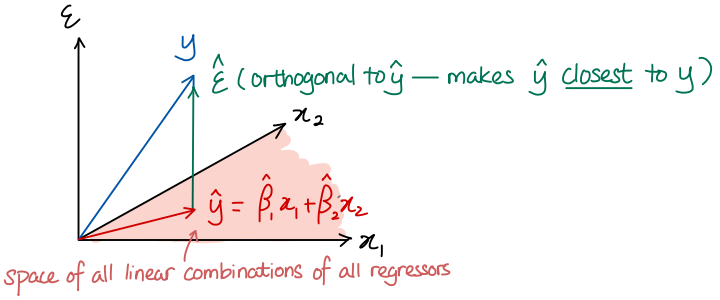
- $\sum_{i=1}^{n}\hat{\varepsilon}_i=0$ (Least squares residuals sum to 0)
- $\bar{\hat{y}}=\bar{y}$
- $\bar{y}=\bar{X}\hat{\beta}$ (Regression "line" passes through the mean)

Residual maker $M = I_n - X(X'X)^{-1}X'$ — $M_X Z$ gives residuals of reg Z X
Projection matrix $P = X(X'X)^{-1}X'$ — Symmetric and idempotent

Estimator: $\hat{\beta} = (X'X)^{-1}X'y$
Fitted values: $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = Py$
Residual: $\hat{\varepsilon} = y-\hat{y} = y-X(X'X)^{-1}X'y = My = M\varepsilon$

- P projects y onto the space of all lin. comb. of regressors
- Orthogonal to the projection space

$PX=X$
$MX=0$ (M orthogonal to X) — regressing X on X gives no residuals
$PM=0$
$\hat{y}'\hat{\varepsilon}=0$ (OLS residuals and fitted values are orthogonal)
$\hat{\varepsilon}'\hat{\varepsilon} = \varepsilon'M\varepsilon$ $\quad E(\varepsilon'M\varepsilon)= E(tr(\varepsilon'M\varepsilon))$



OLS is a problem of finding the point in the space spanned by $x_1, \ldots, x_k$ that is "closest" to y. The solution is $\hat{y}$, which minimises $\|y-\hat{y}\|$

Mathematical Intuition: Projection Theorem. Project y onto subspace of X.

FINITE SAMPLE PROPERTIES OF $s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} = \frac{\varepsilon' M \varepsilon}{n-k}$

ESTIMATOR FOR σ^2 $\quad s^2 = \frac{RSS}{n-k} = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$

$RSS = \hat{\varepsilon}'\hat{\varepsilon} = (My)'(My) = (M(X\beta+\varepsilon))'(M(X\beta+\varepsilon)) = (MX\beta+M\varepsilon)'(MX\beta+M\varepsilon)$
$= (M\varepsilon)'(M\varepsilon) = \varepsilon'M'M\varepsilon = \varepsilon'MM\varepsilon = \varepsilon'M\varepsilon$

UNBIASEDNESS $\quad E(s^2) = \sigma^2$ by A1-4

$E(s^2) = E\left(\frac{\varepsilon'M\varepsilon}{n-k}\right)$

If X is fixed

$E(\varepsilon'M\varepsilon) = E(tr(\varepsilon'M\varepsilon)) = E(tr(M\varepsilon\varepsilon')) = tr(E(M\varepsilon\varepsilon')) = tr(ME(\varepsilon\varepsilon')) \overset{A4}{=} tr(M\sigma^2 I)$ (bc trace is a sum and E is linear operator)
$= \sigma^2 tr(M) = \sigma^2(n-k)$ PS1 Q2

If X is stochastic,

$E(\varepsilon'M\varepsilon) = E(E(tr(\varepsilon'M\varepsilon)|X)) = E(tr(ME(\varepsilon\varepsilon'|X))) \overset{A4}{=} E(tr(M\sigma^2 I)) = \sigma^2 n - k$

SAMPLING DISTRIBUTION $\quad \frac{(n-k)s^2}{\sigma^2} \sim \chi^2_{n-k}$ by A1-5

If X (and thus M) is fixed, $\frac{(n-k)s^2}{\sigma^2} = \frac{\varepsilon'M\varepsilon}{\sigma^2} \underset{PS2}{\sim} \chi^2(n-k)$ (A5: $N(0,\sigma^2 I)$; idempotent; n k λs with = 1 (PS1))

If X (and thus M) is stochastic, $\frac{(n-k)s^2}{\sigma^2} \Big| X \sim \chi^2(n-k)$

ZERO COVARIANCE $\quad \hat{\beta}$ and s^2 are independent

Prove $Cov(\hat{\beta}, \hat{\varepsilon}) = 0 \overset{+A5}{\Longrightarrow} \hat{\beta}$ and $\hat{\varepsilon}$ are independent
$\Rightarrow \hat{\beta}$ and any function of $\hat{\varepsilon}$ (incl. s^2) are independent

$Cov(\hat{\beta}, \hat{\varepsilon}) = E((\hat{\beta} - E(\hat{\beta}))(\hat{\varepsilon} - E(\hat{\varepsilon}))') = E((\hat{\beta}-\beta)(M\varepsilon - E(M\varepsilon))) = E(((X'X)^{-1}X'\varepsilon)\varepsilon'M')$
($\hat{\beta}$ unbiased (A1-3); $= ME(\varepsilon) = 0$ if X is fixed)
if X is fixed $= (X'X)^{-1}X'E(\varepsilon\varepsilon')M' = \sigma^2(X'X)^{-1}X'M' = \sigma^2(X'X)^{-1}(MX)' = 0$ ($E(\varepsilon\varepsilon') = \sigma^2 I$; $MX = 0$)

REGRESSION ANATOMY

PARTITION Let $y = \underset{n\times k}{X}\beta + \varepsilon = (\underset{n\times k_1}{X_1}\ \underset{n\times k_2}{X_2})\begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}^{k_1\times 1}_{k_2\times 1} + \varepsilon$, $k_1+k_2=k$

$= X_1\beta_1 + X_2\beta_2 + \varepsilon$

- Maybe we're only interested in a subset of parameters
- Maybe k_1+k_2 is too large and $X'X$ too large to invert (e.g. fixed effects panel)

FOC (as usual): $\begin{pmatrix}X_1'\\ X_2'\end{pmatrix} X'X\hat\beta = X'y \Rightarrow \begin{pmatrix}X_1'X_1 & X_1'X_2\\ X_2'X_1 & X_2'X_2\end{pmatrix}\begin{pmatrix}\hat\beta_1\\ \hat\beta_2\end{pmatrix} = \begin{pmatrix}X_1'y\\ X_2'y\end{pmatrix}$

$\Rightarrow \begin{aligned} X_1'X_1\hat\beta_1 + X_1'X_2\hat\beta_2 &= X_1'y \\ X_2'X_1\hat\beta_1 + X_2'X_2\hat\beta_2 &= X_2'y\end{aligned}$

PS3 $\Rightarrow$ $\begin{cases}\hat\beta_1 = (X_1'M_2X_1)^{-1}X_1'M_2y, & \text{With } M_2 = I - X_2(X_2'X_2)^{-1}X_2'\\ \hat\beta_2 = (X_2'M_1X_2)^{-1}X_2'M_1y, & \text{With } M_1 = I - X_1(X_1'X_1)^{-1}X_1'\end{cases}$ Consistent (2018 3b)

BIVARIATE EXAMPLE $y = \beta_1 + \beta_2 x + \varepsilon$, let $X_1 = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} = \iota_{n\times 1}$, $X_2 = \begin{pmatrix}x_1\\ \vdots\\ x_n\end{pmatrix} = x$

Then $\hat\beta_2 = (x'Mx)^{-1}x'My$ where $M = I_n - \iota(\iota'\iota)^{-1}\iota' = I_n - \frac{1}{n}\iota\iota'$

$x'Mx = (Mx)'(Mx) = \sum_{i=1}^n (x_i-\bar x)^2$ PS1 Q2

$x'My = (Mx)'(My) = \sum_{i=1}^n (x_i-\bar x)(y_i-\bar y)$

$\hat\beta_2 = \frac{\sum_{i=1}^n (x_i-\bar x)(y_i-\bar y)}{\sum_{i=1}^n (x_i-\bar x)^2}$ as expected

ORTHOGONAL REGRESSORS e.g. MECE dummies (female/male, seasons)

$y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$, $X_1'X_2 = 0$ $(X_1 \perp\!\!\!\perp X_2)$ ★ This reg has no constants

$\hat\beta_1 = (X_1'M_2X_1)^{-1}X_1'M_2y = (X_1'X_1)^{-1}X_1'y$

- can just ignore the uninteresting set of dummies, regress y on X_1 only
- But $X_1'X_2 \neq 0$ in general! Ignoring causes OVB

RESIDUAL-BASED MODEL $y = X_1\beta_1 + X_2\beta_2 + \varepsilon$

$\Rightarrow M_2y = M_2X_1\beta_1 + M_2\varepsilon$ $(M_2X_2 = 0)$

$\Rightarrow y^* = X_1^*\beta_1 + M_2\varepsilon$ (error)

$y^* = M_2y$ contains residuals from reg y on X_2

$X_1^* = M_2X_1$ contains residuals from k_1 regressions of X_1 on X_2 ?

Essentially, M_2 finds the transformation to the original regression that removes β_2

If $X_2 = \begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$ (so β_2 is just the constant), $M_2y = y - \bar y$ PS1

FRISCH-WAUGH-LOVELL THEOREM We can get $\hat\beta_1$ by performing OLS on the residual based model

- Furthermore, residuals from the residual based model $M_2\hat\varepsilon = \hat\varepsilon$ (from full regression)! Identical RSS!
- Residual based regression removes the correlation that X_1 and y have with X_2 before estimating β_1
- We can also regress $y = X_1^*\beta_1 + e$ but residuals incorrect for obtaining s^2

SE OF RESIDUAL-BASED MODEL $\mathrm{Var}(\hat\beta_1) = \sigma^2(X_1'M_2X_2)^{-1}$, $\widehat{\mathrm{Var}}(\hat\beta_1) = s^2(X_1'M_2X_2)^{-1}$

$s^2 = \frac{\hat\varepsilon'\hat\varepsilon}{n-k_1-k_2}$, $\hat\varepsilon = y - X_1\hat\beta_1 - X_2\hat\beta_2 = y^* - X_1^*\hat\beta_1$ by FWL

↑ Still controlling for X_2 (implicitly/explicitly)

DESEASONALISATION

time series/panel with seasonal fluctuations

If we have $y = S\gamma + M\delta + \varepsilon$, where S is a complete set of seasonal dummies (mutually orthogonal!)
X are economic variables

By FWL we can estimate δ by $M_S y = M_S X\delta + M_S\varepsilon$
$M_S y$ is residuals from regressing y on seasonal dummies, e.g.

$$y_t = \gamma_1 s_{1t} + \gamma_2 s_{2t} + \gamma_3 s_{3t} + \gamma_4 s_{4t} + \varepsilon_t \quad \text{where } s_{jt} = \begin{cases} 1 & \text{if } t \text{ lies in the } j^{th} \text{ season} \\ 0 & \text{otherwise} \end{cases}$$

Then $\hat{\varepsilon}_t = y_t - (\gamma_1 s_{1t} + \gamma_2 s_{2t} + \gamma_3 s_{3t} + \gamma_4 s_{4t})$

Since $s_{1t}, s_{2t}, s_{3t}, s_{4t}$ are mutually orthogonal, by FWL, we can estimate γ_1 by $y_t = \gamma_1 s_{1t} + v_t$

$$\text{This means that } \hat{\gamma}_1 = (s_1's_1)^{-1}s_1'y \equiv \frac{\sum s_{1t}y_t}{\sum s_{1t}^2} = \frac{\sum s_{1t}y_t}{\sum s_{1t}} \quad (\text{since } s_{1t} = 0 \text{ or } 1)$$
$$= \bar{y}_1, \text{ average of } y \text{ values falling in first season}$$

By similar logic, $\hat{\varepsilon}_t = y_t - \bar{y}_1 s_{1t} - \bar{y}_2 s_{2t} - \bar{y}_3 s_{3t} - \bar{y}_4 s_{4t}$

$$M_S y = \hat{\varepsilon}_t = \begin{cases} y_t - \bar{y}_1 & \text{if } t \text{ is first season} \\ y_t - \bar{y}_2 & \text{if } t \text{ is second season} \\ y_t - \bar{y}_3 & \text{if } t \text{ is third season} \\ y_t - \bar{y}_4 & \text{if } t \text{ is fourth season} \end{cases}$$

This deseasonalises the y variable by subtracting seasonal averages from y_t before further regressing on economic variables ($M_S y = M_S X\delta + M_S\varepsilon$)

- Many time series already deseasonalised before access

PANEL

with unobserved heterogeneity

Individual heterogeneity (fixed effect)

$$y_{it} = \underset{n\times k}{\alpha_i} + \underset{k\times 1}{Z_{it}'}\beta + \varepsilon_{it} \qquad i = 1, \dots, n$$

k+n parameters too many!

$$\underset{1\times nT}{\begin{pmatrix} y_{11} \\ \vdots \\ y_{nT} \end{pmatrix}} = \begin{pmatrix} \underset{\perp}{Z_{it}'} & \underset{\perp}{d_{1,it}} & \underset{\perp}{d_{2,it}} & \cdots & \underset{\perp}{d_{n,it}} \end{pmatrix}\begin{pmatrix} \beta \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} + \varepsilon_{it}, \quad d_{j,it} = \begin{cases} 1 \ \forall t & \text{if } i=j \\ 0 \ \forall t & \text{otherwise} \end{cases}$$

$$= (Z \; D)\begin{pmatrix} \beta \\ \alpha \end{pmatrix} + \varepsilon_{it} = Z\beta + D\alpha + \varepsilon_{it}$$

We want $M_D y$, the residual from reg y on D, to estimate β with $M_D y = M_D Z\beta + M_D\varepsilon_{it}$
Since $d_{j,it}$ are mutually orthogonal, by FWL we can estimate α_j by $y_{it} = d_{j,it}\alpha_j + v_{it}$

$$\text{Then } \alpha_j = (d_{j,it}'d_{j,it})^{-1}d_{j,it}'y = \frac{\sum_t\sum_i d_{j,it}y_{it}}{\sum_t\sum_i d_{j,it}^2} = \frac{\sum_{t=1}^{T}\sum_{i=1}^{n} d_{j,it}y_{it}}{\sum_t\sum_i d_{j,it}} \quad \text{Since } d_{j,it} = 0,1$$

$$= \frac{\sum_{t=1}^{T} y_{jt}}{T} = \bar{y}_j \quad \text{(individual j's average over time)} \quad \text{since } d_{j,it}y_{it} = \begin{cases} y_{it} & \text{if } i=j \\ 0 & \text{otherwise} \end{cases}$$

$$\Rightarrow \hat{v}_{it} = y_{it} - \bar{y}_i$$

$$M_D y = \begin{pmatrix} \hat{v}_{11} \\ \vdots \\ \hat{v}_{1T} \\ \vdots \\ \hat{v}_{nT} \end{pmatrix} = \begin{pmatrix} y_{11} - \bar{y}_1 \\ \vdots \\ y_{1T} - \bar{y}_1 \\ \vdots \\ y_{nT} - \bar{y}_n \end{pmatrix} \qquad \text{Similarly, } M_D Z = \underset{1\times k \text{ vectors}}{\left((Z_{11} - \bar{Z}_1)' \cdots (Z_{1T} - \bar{Z}_1 \cdots Z_{nT} - \bar{Z}_n\right)'}$$

$$y_{it} - \bar{y}_i = (z_{it} - \bar{z}_i)'\beta + M_D\varepsilon_{it}$$

OVB

True (Long): $y = X\beta + Z\delta + \varepsilon$ $\quad E(\hat{\beta}) = \beta \quad Var(\hat{\beta}) = \sigma^2 I$

Misspecified (Short): $y = X\beta + v$

assume X, Z fixed for simplicity

The short $\tilde{\beta} = (X'X)^{-1}X'y$

$= (X'X)^{-1}X'(X\beta + Z\delta + \varepsilon)$

$= \beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'\varepsilon$ is biased!

$E(\tilde{\beta}) \neq \beta$

$E(\tilde{\beta}) = E(\beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'\varepsilon)$

$= \beta + (X'X)^{-1}X'Z\delta + (X'X)^{-1}X'E(\varepsilon)$

$= \beta + (X'X)^{-1}X'Z\delta$

bias!! rep indirect effect of X on y through its correlation with Z

"capturing too much"

CONDITIONS $X'Z \neq 0$ (X affects Z)

$\delta \neq 0$ (Z is relevant)

SE IS BIASED $Var(\tilde{\beta}) = \sigma^2(X'X)^{-1}$ which is "correct", but how we estimate σ^2 is wrong!

$\tilde{s}^2 = \frac{\hat{v}'\hat{v}}{n-k_x}$, $\hat{v} = y - X\tilde{\beta}$ is not unbiased! se/t/F-stats based on $\tilde{s}^2$ invalid!

$y = X\beta + v \Rightarrow \hat{v} = M_x y \Rightarrow \hat{v} = M_x(X\beta + Z\delta + \varepsilon) = M_x Z\delta + M_x\varepsilon$

Therefore, $\hat{v}'\hat{v} = (M_x Z\delta + M_x\varepsilon)'(M_x Z\delta + M_x\varepsilon)$

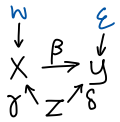
$= \varepsilon' M_x \varepsilon + \delta Z M_x Z\delta + 2\delta' Z' M_x \varepsilon$

$E(\hat{v}'\hat{v}) = E(\varepsilon M_x \varepsilon) + E(\delta Z M_x Z\delta) + 2\delta' Z' M_x E(\varepsilon)$ assuming X, Z fixed

$= \sigma^2(n-k_x) + E(\delta Z M_x Z\delta)$ by A1 (↓ usual process)

So $E(\tilde{s}^2) = E\left(\frac{\hat{v}'\hat{v}}{n-k_x}\right) \neq \sigma^2$

OVB VIOLATES A3



True: $y = X\beta + Z\delta + \varepsilon$ $\quad E(\varepsilon) = 0$

$\Rightarrow v = Z\delta + \varepsilon$

Short: $y = X\beta + v$

Auxiliary: $X = Z\gamma + w$ $\quad E(w) = 0$

Assume $w \perp \varepsilon$, so $w'\varepsilon = 0$

If $E(v|X) = 0$, by LIE, $E(X'v) = E(E(X'v|X)) = E(X'E(v|X)) = 0$ BUT

However, $E(X'v) = E[(Z\gamma + w)'(Z\delta + \varepsilon)]$

$= E(\gamma Z'Z\delta + \gamma' Z\varepsilon + w'Z\delta + w'\varepsilon)$

$= E(\gamma Z'Z\delta) + \gamma' Z\underbrace{E(\varepsilon)}_{=0} + \underbrace{E(w')}_{=0} Z\delta + \underbrace{E(w'\varepsilon)}_{=0}$

$= E(\gamma Z'Z\delta)$

$\neq 0$

Thus, $E(v|X) \neq 0$. A3 is broken!

INCLUDING IRRELEVANT VARIABLES

True : $y = X\beta + \varepsilon$

Misspecified: $y = X\beta + Z\delta + v$

Note: Z is irrelevant to y:

$\tilde{\tilde{\beta}} = (X'M_Z X)^{-1} X'M_Z y$ is still unbiased and $\tilde{\tilde{s}}^2$ is unbiased, so inference is valid

That said, $\tilde{\tilde{\beta}}$ not the most efficient (power)

$\tilde{\tilde{\beta}}$ IS UNBIASED

$$\tilde{\tilde{\beta}} = (X'M_Z X)^{-1} X'M_Z (X\beta + v) = \beta + (X'M_Z X)^{-1} X'M_Z v$$

$$E(\tilde{\tilde{\beta}}) = E\left[\beta + (X'M_Z X)^{-1} X'M_Z v\right] = \beta + (X'M_Z X)^{-1} X'M_Z E(v) = \beta$$

(X, Z fixed)

SE IS UNBIASED

$$\begin{aligned} Var(\tilde{\tilde{\beta}}) &= Var(\beta + (X'M_Z X)^{-1} X'M_Z v) \\ &= (X'M_Z X)^{-1} X'M_Z \, Var(v) \, M_Z' X (X'M_Z X)^{-1} \quad X, Z \text{ fixed} \\ &= \sigma^2 (X'M_Z X)^{-1} \checkmark \end{aligned}$$

$s^2 = \dfrac{\hat{v}'\hat{v}}{N - k_x - k_z}$ is unbiased since

$$\hat{v} = M_{xz} y = M_{xz}(X\beta + \varepsilon) \overset{AI}{=} M_{xz}\varepsilon$$

$$\Rightarrow E(\hat{v}'\hat{v}) = E(\varepsilon M_{xz} \varepsilon) \overset{\text{usual process}}{=} \sigma^2 (N - k_x - k_z) \checkmark$$

EFFICIENCY LOSS

$Var(\tilde{\tilde{\beta}}) \geq Var(\hat{\beta})$ $\quad \sigma^2 (X'M_Z X)^{-1} \geq \sigma^2 (X'X)^{-1}$

- Select relevant variables based on economic arguments, not statistical ones

To prove: $(X'M_Z X)^{-1} - (X'X)^{-1}$ is positive semidefinite

$\Updownarrow$

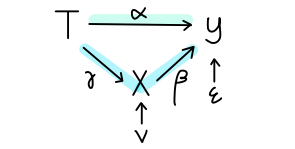
$X'X - X'M_Z X$ is positive semidefinite

$$\begin{aligned} X'X - X'M_Z X &= X'X - X'X + X'Z(Z'Z)^{-1} Z'X \\ &= D'D \text{ with } D = (Z'Z)^{-\frac{1}{2}} Z'X \text{ positive semidefinite} \end{aligned}$$

PSI

MEDIATION

direct effect : α
indirect effect : $\gamma'\beta$



Assume
$\varepsilon \perp T \quad T \perp V$
$V \perp \varepsilon \quad X \perp \varepsilon$

$$\underset{n\times 1}{y} = \delta + T\delta + \underset{k_x\times 1}{X\beta} + \varepsilon$$
$$X = \underset{n\times k_x}{\tau} + \underset{n\times k_T}{T}\underset{k_T\times k_x}{\gamma} + V$$

GM assumptions are satisfied!
e.g. A3: $E(\varepsilon|T,X)=0$

If we exclude X: $y = \delta + T\alpha + (\tau + T\gamma + V)\beta + \varepsilon$
$= \underbrace{\delta + \tau\beta}_{\text{constant}} + T\underbrace{(\alpha+\gamma\beta)}_{\text{coefficient}} + \underbrace{V\beta+\varepsilon}_{\text{error}}$

This means that the coefficient from reg y on T gives the total effect $\alpha+\gamma\beta$
reg y on T, X gives β
reg X on T gives γ

Observing y, T, X allows us to differentiate direct/indirect effect
Observing y, T only gives total effect estimate

This means that if you include "controls" that are actually mediators, the resultant decrease in the coefficient on T just means the direct effect < total effect. Be careful about interpretation!

SELECTING REGRESSORS

GOODNESS OF FIT $R^2 = 1 - \frac{RSS}{TSS}$ ($RSS = \hat{\varepsilon}'\hat{\varepsilon}$, $TSS = \sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2$ also called SST)

★ implies $\hat{\varepsilon}'\hat{\varepsilon} = RSS = TSS(1-R^2)$

- Not suitable in helping to select regressors since additional regressors always $\uparrow R^2$

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

- Allows for a tradeoff between fit and parsimony ($\uparrow k \Rightarrow \downarrow n-k \Rightarrow \uparrow \bar{R}^2$)

SUMMARY

$y = X\beta + Z\delta + \varepsilon$

True state of the world	Include Z	Don't Include Z
$\delta = 0$	Unbiased but higher variance	BLUE
$\delta \neq 0$	BLUE	Biased but lower variance

Unbiasedness is important but so is variance!
Don't just include all potential regressors! Think about the tradeoff

HYPOTHESIS TESTING

under A1-5
assume X fixed for simplicity

SETUP

From A1-5, the exact sampling distribution of

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$
$$\hat{\beta}_i \sim N(\beta_i, \sigma^2 c_{ii}), \quad c_{ii} = [(X'X)^{-1}]_{ii}$$
$$(n-k)\sigma^{-2}s^2 \sim \chi^2_{n-k}$$

$\hat{\beta}$ and s^2 are independent ∎

T-TEST (SINGLE PARAMETER)

$H_0: \beta_i = \mu$ vs $H_1: \beta_i \neq \mu$

If σ^2 is known, we use $z = \frac{\hat{\beta}_i - \mu}{\sigma\sqrt{c_{ii}}} \sim N(0,1)$ under H_0

Reject H_0 (two-sided) if $\left|\frac{\hat{\beta}_i - \mu}{\sigma\sqrt{c_{ii}}}\right| > z_{\frac{\alpha}{2}} \overset{\alpha=5\%}{=} 1.96$

If σ^2 is unknown, we use $t = \frac{\hat{\beta}_i - \mu}{s\sqrt{c_{ii}}} = \frac{\hat{\beta}_i - \mu}{SE(\hat{\beta}_i)} \sim t_{n-k}$ under H_0 (fatter tails (start rejecting later))

Reject H_0 (two-sided) if $\left|\frac{\hat{\beta}_i - \mu}{SE(\hat{\beta}_i)}\right| > t_{n-k,\frac{\alpha}{2}}$

t is derived from $\frac{\hat{\beta}_i - \mu}{s\sqrt{c_{ii}}} = \frac{\left(\frac{\hat{\beta}_i - \mu}{\sigma\sqrt{c_{ii}}}\right) \leftarrow N(0,1)}{\left(\frac{(n-k)\sigma^{-2}s^2}{n-k}\right) \leftarrow \chi^2_{n-k}} \sim t_{n-k}$ under H_0 as $\hat{\beta}$ and s^2 are independent

For small n, critical values are typically larger when σ^2 is unknown
- t-distribution has fatter tails than N
- "Reject the null later"

As $n \to \infty$, $t \to z$ so critical values are identical whether σ^2 is known or not

- If you get a low t with high $se(\hat{\beta})$ but $\hat{\beta}$ far from 0, it's noisy/imprecise data (can't conclude)
- If you get a low t with low $se(\hat{\beta})$ but $\hat{\beta}$ close to 0, you're "not rejecting precisely" (likely no effect)

P-VALUE

$$\Pr\left(|T| \geq \frac{\hat{\beta}_i - \mu}{SE(\hat{\beta}_i)}\right)$$

Lowest level of significance at which you want to reject H_0

CONFIDENCE INTERVAL

$100(1-\alpha)\%$ CI for β_i (σ^2 unknown)

$$\left[\hat{\beta}_i - t_{n-k,\frac{\alpha}{2}} SE(\hat{\beta}_i), \hat{\beta}_i + t_{n-k,\frac{\alpha}{2}} SE(\hat{\beta}_i)\right]$$ derive in PS4 Q3

such that $\Pr(\hat{\beta}_i - t_{n-k,\frac{\alpha}{2}} SE(\hat{\beta}_i) < \beta_i \text{ (true value)} < \hat{\beta}_i + t_{n-k,\frac{\alpha}{2}} SE(\hat{\beta}_i)) = \Pr(-t_{n-k,\frac{\alpha}{2}} < \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)} < t_{n-k,\frac{\alpha}{2}}) = 1-\alpha$

Reject $H_0: \beta_i = \mu$ with sig. level α with t-test $\overset{PS4Q3}{\iff} \mu \notin 100(1-\alpha)\%$ CI

T-TEST (LINEAR RESTRICTION)

$H_0: r_1\beta_1 + r_2\beta_2 + \dots + r_k\beta_k = c$ or equivalently $H_0: r'\beta = c$

$r'\hat{\beta} \sim N(c, \sigma^2 r'(X'X)^{-1}r)$ under H_0

because $\hat{\beta} \sim N(\beta, \text{Var}(\hat{\beta}))$
$\Rightarrow r'\hat{\beta} - c \sim N(r'\beta - c, r'\text{Var}(\hat{\beta})r)$ under H_0

If σ^2 is known, we use $z = \dfrac{r'\hat{\beta}-c}{\sigma\sqrt{r'(X'X)^{-1}r}} = \dfrac{r'\hat{\beta}-c}{SD(r'\hat{\beta}-c)} \sim N(0,1)$ under H_0

If σ^2 is unknown, we use $t = \dfrac{r'\hat{\beta}-c}{s\sqrt{r'(X'X)^{-1}r}} = \dfrac{r'\hat{\beta}-c}{SE(r'\hat{\beta}-c)} \sim t_{n-k}$ under H_0

fatter tails (start rejecting later)

☆ $SE(\hat{\beta}_1+\hat{\beta}_2) \neq SE(\hat{\beta}_1) + SE(\hat{\beta}_2)$

Just reparameterise: rewrite model with $\gamma = \beta_1 + \beta_2 - c$, test $H_0: \gamma = 0$ PS5 Q1b
Allows us to just use the "standard OLS package"

STEPS

1. Define H_0, H_1
2. Calculate test statistic and give its distribution under H_0
3. Give sig. level and critical values
4. State rejection rule and interpret findings
5. Clearly indicate assumptions for test to be valid (A1-5)

IMPORTANT

Single hypothesis tests need to be followed by joint hypothesis test
If near multicollinear, marginal contribution of each regressors may be small (low t) when added last but they may be jointly significant (high F/W)

POWER

Power = $1 - Pr(\text{Type II error})$
= $Pr(H_0 \text{ rejected} \mid H_0 \text{ true})$

e.g. $H_0: \beta = \beta_1$ vs $H_1: \beta \neq \beta_1$

Power of this test when $\beta = \beta_1$ is $Pr\left(\left|\dfrac{\hat{\beta}}{\sigma}\right| \geq z_{\frac{\alpha}{2}} \mid \beta = \beta_1\right)$

$= Pr\left(\dfrac{\hat{\beta}}{\sigma} \leq -z_{\frac{\alpha}{2}} \text{ or } \dfrac{\hat{\beta}}{\sigma} \geq z_{\frac{\alpha}{2}} \mid \beta = \beta_1\right)$

$= Pr\left(\dfrac{\hat{\beta}-\beta_1}{\sigma} \leq -z_{\frac{\alpha}{2}} - \dfrac{\beta_1}{\sigma}\right) + Pr\left(\dfrac{\hat{\beta}-\beta_1}{\sigma} \geq z_{\frac{\alpha}{2}} - \dfrac{\beta_1}{\sigma}\right)$

$\uparrow \beta_1 \Rightarrow \uparrow$ Power $\Rightarrow \downarrow Pr(\text{Type II error})$

$\uparrow \alpha \Rightarrow \uparrow$ Power $\Rightarrow \downarrow Pr(\text{Type II error})$
$= \uparrow Pr(\text{Type I error})$

To ensure test is powerful,

HYPOTHESIS TESTING - MULTIPLE LIN RESTRICTIONS

F-TEST $H_0: \underset{J\times 1}{R\beta} = \underset{J\times 1}{c}$ R is a J×k full rank matrix, $J \leq k < n$

(J = no. of restrictions; no redundant restrictions)

e.g. let $k=4$, $\beta = \begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}$

$R = (0\ 1\ -1\ 0)$ $c = (0)$ then $H_0: \beta_2 = \beta_3$ (single linear restriction)

$R = \begin{pmatrix}0&1&0&0\\0&0&1&0\end{pmatrix}$ $c = \begin{pmatrix}2\\1\end{pmatrix}$ then $H_0: \beta_2 = 2$ and $\beta_3 = 1$

$R = \begin{pmatrix}0&1&0&0\\0&0&1&0\\0&0&0&1\end{pmatrix}$ $c = \begin{pmatrix}0\\0\\0\end{pmatrix}$ then $H_0: \beta_2 = \beta_3 = \beta_4 = 0$

$R = \begin{pmatrix}0&1&-1&0\\0&0&1&-2\end{pmatrix}$ $c = \begin{pmatrix}0\\0\end{pmatrix}$ then $H_0: \beta_2 = \beta_3 = 2\beta_4 = 0$

NOT $\begin{pmatrix}0&1&-1&0\\0&1&0&-2\\0&0&1&-2\end{pmatrix}$ which doesn't have full row rank

WALD TEST

One of two hypothesis testing principles (the other is RRSS/URSS)
Do the estimates come reasonably close to satisfying the restrictions implied by the hypothesis?

$H_0: R\beta = c$ vs $H_1: R\beta \neq c$

We check if discrepancy of $R(\hat{\beta}-c)$ from 0 is statistically significant or just due to sampling error

DISCREPANCY VECTOR $d = R\hat{\beta} - c$.

Typically use the quadratic form $d'[Var(d)]^{-1}d$ that is a scalar measure of the size of vector

If σ^2 is known,

Under H_0, $d \sim N(0, Var(d))$, $Var(d) = \sigma^2 R(X'X)^{-1}R'$ (must have full rr for Var to be non-singular)

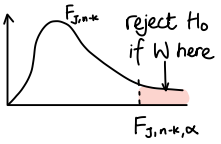
$$\frac{(R\hat{\beta}-c)[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-c)}{\sigma^2} = d'[Var(d)]^{-1}d \sim \chi^2_J$$

(vector of normal vars; non-singular; PS2 Q3)

If σ^2 is unknown,

Under H_0, $$W = \frac{(R\hat{\beta}-c)[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-c)}{S^2 J} = \frac{d'[\widehat{Var}(d)]^{-1}d}{J} \sim F_{J,n-k}$$

Then we reject H_0 if $W > F_{J,n-k,\alpha}$ not $\frac{\alpha}{2}$!! Squaring d makes one-sided tests impossible anyway



WHY KEEP T-TEST? When there is only a single linear restriction such that d is scalar,

Under H_0, $d'[Var(d)]^{-1}d = \left(\frac{d}{SD(d)}\right)^2 \sim \chi^2_1$ if σ^2 is known

$$\frac{d'[\widehat{Var}(d)]^{-1}d}{J} = \left(\frac{d}{SE(d)}\right)^2 \sim t^2_{n-k} = F_{1,n-k}$$

By taking squares, we hide the direction of our violations and makes one-sided tests impossible.

LOSS OF FIT (COMPARE RRSS / URSS)

$$H_0: R\beta = c \quad \text{vs} \quad H_1: R\beta \neq c$$

Does imposing the restrictions in H_0 lead to a significant loss in fit?

① Regress y on X to get the unconstrained least squares estimator $\hat{\beta}$.

② Calculate $URSS = \hat{\varepsilon}'\hat{\varepsilon} = (y - X\hat{\beta})'(y - X\hat{\beta})$

③ Regress y on X subject to the constraint $R\beta = c$ (minimise $(y-X\beta)'(y-X\beta)$ st $R\beta = c$) to get constrained least squares estimator β^*

④ Calculate $RRSS = \varepsilon^{*\prime}\varepsilon^* = (y - X\beta^*)'(y - X\beta^*)$

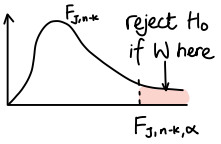
⑤ $F = \dfrac{RRSS - URSS}{URSS}\left(\dfrac{n-k}{J}\right) \sim F_{J,n-k}$ under H_0. (J: no. of restrictions; n−k: dof of unrestricted model)

also written as $\left(\dfrac{R^2_{unconstr} - R^2_{constr}}{1 - R^2_{constr}}\right)\left(\dfrac{n-k}{J}\right) \sim F_{J,n-k}$

- F is a ratio measuring % increase in residual variance (loss in fit) due to imposing H_0.

⑥ Find the critical value K^* st α% of the area under an $F_{J,n-k}$ distribution lies to the right of K^*

⑦ Reject H_0 if $F > K^*$



SIMPLE EXAMPLE - "NOTHING IS HAPPENING HERE" except for the constant

$y = \beta_1 + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i = x_i'\beta + \varepsilon_i$ $\quad H_0: \beta_2 = \beta_3 = \dots = \beta_k = 0$ (k−1 restrictions) vs $H_1: (\beta_2, \dots, \beta_k) \neq (0, \dots, 0)$

$URSS = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 = RSS$

Restricted model is $y_i = \beta_1 + \varepsilon_i \Rightarrow \beta_1^* = \bar{y} \Rightarrow RRSS = \sum_{i=1}^{n}\varepsilon_i^{*2} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = TSS$

$$F = \left(\frac{TSS - RSS}{RSS}\right)\left(\frac{n-k}{k-1}\right) = \frac{R^2}{1-R^2}\left(\frac{n-k}{k-1}\right) \sim F_{k-1,n-k} \text{ under } H_0 \qquad R^2 = 1 - \frac{RSS}{TSS}$$

- More likely to reject H_0 when R^2 is large! Regressors more relevant!

FINDING RRSS

Find β which minimises $(y-X\beta)'(y-X\beta)$ subject to $R\beta = c$

Other than plugging in the restrictions (see above), can also do Lagrangian

$$\mathcal{L}(b, \lambda) = (y - Xb)'(y - Xb) - 2\lambda'(Rb - c)$$

(2 just helps simplify (actual value of λ unimportant))

$$\text{FOCs}: \begin{cases} -2X'y + 2X'X\beta^* - 2R'\lambda^* = 0 \\ -2(R\beta^* - c) = 0 \end{cases}$$

$$\Rightarrow \beta^* = \underbrace{(X'X)^{-1}X'y}_{=\hat{\beta}} + (X'X)^{-1}R'\lambda^*$$

$$R\beta^* = R\hat{\beta} + R(X'X)^{-1}R'\lambda^*$$

$$c = R\hat{\beta} + R(X'X)^{-1}R'\lambda^*$$

$$\lambda^* = (R(X'X)^{-1}R')^{-1}(c - R\hat{\beta}) \quad \leftarrow \text{if } R\hat{\beta} = c,\ \lambda^* = 0$$

$$\beta^* = \hat{\beta} - (X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(c - R\hat{\beta}) \quad \leftarrow \text{if } R\hat{\beta} = c,\ \beta^* = \hat{\beta}$$

λ^* is "shadow prices" of imposing the restrictions — high is bad

LOSS OF FIT = WALD TEST

$$F = W \equiv \frac{d'(\widehat{Var}(d))^{-1}d}{J} \sim F_{J,n-k} \text{ under } H_0$$

$$d = R\hat{\beta} - c$$
$$Var(d) = \sigma^2 R(X'X)^{-1}R'$$
$$\widehat{Var}(d) = s^2 R(X'X)^{-1}R'$$

Allows us to conclude $F = \frac{RRSS-URSS}{URSS}\left(\frac{n-k}{J}\right) \sim F_{J,n-k}$ under H_0

PROOF

$$\frac{RRSS-URSS}{URSS}\left(\frac{n-k}{J}\right) = \frac{RRSS-URSS}{s^2J}$$

$$= \frac{(y-X\beta^*)'(y-X\beta^*) - (y-X\hat{\beta})'(y-X\hat{\beta})}{s^2J}$$

$$= \frac{(y-X\hat{\beta}+D)'(y-X\hat{\beta}+D) - (y-X\hat{\beta})'(y-X\hat{\beta})}{s^2J}, \quad D = X(X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(c-R\hat{\beta})$$

$$= \frac{\cancel{(y-X\hat{\beta})'(y-X\hat{\beta})} + D'D + \cancel{(y-X\hat{\beta})'D} + \cancel{D'(y-X\hat{\beta})} - \cancel{(y-X\hat{\beta})'(y-X\hat{\beta})}}{s^2J}$$

since $(y-X\hat{\beta})'D = \hat{\varepsilon}'X(X'X)^{-1}R'(R(X'X)^{-1}R')^{-1}(c-R\hat{\beta}) = 0$

$$= \frac{D'D}{s^2J}$$

$$= \frac{d'[Var(d)]^{-1}d}{J}$$

$$= W$$

Note: Wald test is preferred to comparing $\frac{RRSS}{URSS}$ (or equivalently $\frac{R_r^2}{R_u^2}$) when there is heteroscedasticity (can use RSE for $\widehat{var}(d)$)

$t^2 = F$

$$F = \frac{RRSS-URSS}{URSS}\left(\frac{n-k}{J}\right) = \frac{d'(\widehat{Var}(d))^{-1}d}{J} = \left(\frac{d}{SE(d)}\right)^2 = t^2$$

(from above; with only 1 restriction ($J=1$))

ASYMPTOTIC THEORY

In the absence of a finite (exact) sampling distribution, we use estimators that are consistent (convergence in probability) and has a limiting distribution (convergence in distribution)
asymptotically unbiased — enables hypothesis testing

"density of estimator converges to a spike at true value"

CONSISTENT

$\hat{\theta}$ is consistent if, as $n \to \infty$, $\operatorname{plim} \hat{\theta} = \theta$ or $\hat{\theta} \xrightarrow{p} \theta$

Formal definition:
Let X_n be a random variable indexed by the size of a sample
X_n converges in probability to X ($X_n \xrightarrow{p} X$) if $\lim_{n\to\infty} P(|X_n - X| > \epsilon) = 0 \quad \forall \epsilon > 0$

PROVING CONSISTENCY

SUFFICIENT CONDITIONS

① $\lim_{n\to\infty} E(\hat{\theta}) = \theta$ (Asymptotic unbiasedness)

② $\lim_{n\to\infty} \operatorname{Var}(\hat{\theta}) = 0$ requires $\operatorname{Var}(\hat{\theta})$ to be finite. Stronger than needed

① + ② ⇒ Chebychev's Inequality ⇒ Consistency
(Chebychev's Inequality: ★ stronger-than-needed concept of convergence)

$$P(|\hat{\theta} - \theta| > \epsilon) \leq \frac{E[(\hat{\theta} - \theta)^2]}{\epsilon^2} = \frac{MSE(\hat{\theta})}{\epsilon^2} \to 0$$

- Unbiased ⇒ asymptotically unbiased (common sense)

KHINCHINE'S WEAK LAW OF LARGE NUMBERS

If $X_1, \ldots, X_n$ are random samples (i.i.d) with finite mean μ,

$$\operatorname{plim}(\bar{X}) = \operatorname{plim} \frac{1}{n}\sum_{i=1}^{n} X_i = E(X_i) \equiv \mu$$

no need to assume $\operatorname{Var}(X_i)$ is finite, or derive $\operatorname{Var}(\bar{X})$

- Sample averages converge to their population counterparts
- Can use to prove others, $\operatorname{plim} \frac{1}{n}\sum X_i^2 = E(X_i^2)$ (X_i iid ⇒ X_i^2 iid; if it exists), $\operatorname{plim} \frac{1}{n}\sum X_i \varepsilon_i = E(X_i \varepsilon_i) = 0$ (X_i, ε_i iid ⇒ $X_i \varepsilon_i$ iid) with Slutsky theorem
- Alt. LLNs exist for X_i, ε_i not iid
- Approach: Write equations of estimators as summation, multiply & divide by n and apply WLLN

SLUTSKY THEOREM

$\operatorname{plim} g(Y_n) = g(\operatorname{plim}(Y_n))$ $\forall$ continuous $g(\cdot)$

addition
multiplication
matrix multiplication
inverse
into elements of a matrix

OTHER TOOLS

$\operatorname{plim}(c) = c$

- If something isn't a function of n, it is equal to its plim

$\operatorname{plim}_{n\to\infty}(X_n) = \lim_{n\to\infty}(X_n)$ if X_n is not random

- e.g. $\operatorname{plim} \frac{n}{n-k} = \lim_{n\to\infty} \frac{n}{n-k}$ if k is not changing

PROOFS OF CONSISTENCY using Slutsky, WLLN

SAMPLE MEAN IS CONSISTENT If $X_1,\dots,X_n$ is a random sample (i.i.d), $X_i \sim (\mu,\sigma^2)$, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a consistent estimator of μ

Sufficient conditions
Method ① $E(\bar{X}) = \mu$, $Var(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow \lim_{n\to\infty} E(\bar{X}) = \mu$, $\lim_{n\to\infty} Var(\bar{X}) = 0 \Rightarrow \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mu$ assumes σ^2 is finite

WLLN
Method ② $E(X_i)$ is finite, so apply WLLN directly: $\text{plim}\,\bar{X} = E(X_i) \equiv \mu$. Consistent! No need $Var(X_i) = \sigma^2 < \infty$. No need to derive $Var(\bar{X})$

$\bar{X}^2$ IS CONSISTENT $X_1,\dots,X_n$ is a random sample (i.i.d) with finite mean μ

Sufficient conditions
Method ① $Var(\bar{X}) = E(\bar{X}^2) - (E(\bar{X}))^2 \Rightarrow E(\bar{X}^2) = \underbrace{\frac{\sigma^2}{n} + \mu^2}_{\text{biased}} \xrightarrow[n\to\infty]{\text{but asymptotically unbiased}} \mu^2$
$Var(\bar{X}^2) = \dots$? too hard.

WLLN, Slutsky
Method ② $\xrightarrow{WLLN}$ $\text{plim}\,\bar{X} = \mu$ $\xRightarrow{Slutsky}$ $\text{plim}\,\bar{X} = \mu^2$ ✓

$\widehat{Cov}(x_i,\varepsilon_i)$ IS CONSISTENT $\text{plim}\,\widehat{Cov}(x_i,\varepsilon_i) = \text{plim}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i - \bar{x}\bar{\varepsilon}\right) \overset{Slutsky}{=} \text{plim}\,\overline{x_i\varepsilon_i} - \text{plim}\,\bar{x}\,\text{plim}\,\bar{\varepsilon} \overset{WLLN}{=} E(x_i\varepsilon_i) - E(x_i)E(\varepsilon_i) = Cov(x_i,\varepsilon_i)$

$\hat{\beta}$ IS CONSISTENT - BIVARIATE $\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Sufficient conditions
Method ① $E(\hat{\beta}) = \beta$ and $Var(\hat{\beta}|X) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\frac{\sigma^2}{n} \to 0 \text{ as } n\to\infty}{\left(\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}\right) \not\to 0}$ excitation condition ✓

WLLN, Slutsky
Method ② $\text{plim}(\hat{\beta}) = \text{plim}\left(\beta + \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}\right) = \beta + \frac{\text{plim}\,\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})}{\text{plim}\,\frac{1}{n}\sum(x_i-\bar{x})^2} = \beta + \frac{\text{plim}\,\widehat{Cov}(x_i,\varepsilon_i)}{\text{plim}\,\widehat{Var}(x_i)} \overset{WLLN}{=} \beta + \frac{Cov(x_i,\varepsilon_i)}{Var(x_i)} \overset{\text{assume } (x_i,\varepsilon_i) \text{ iid}}{=} \beta$

$\hat{\beta}$ IS CONSISTENT - MULTIVARIATE $\hat{\beta} = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'\varepsilon = \beta + \left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}$ write in terms of averages

In bivariate case, $\frac{X'X}{n} = \begin{pmatrix} 1 & \sum_{i=1}^{n}\frac{x_i}{n} \\ \sum_{i=1}^{n}\frac{x_i}{n} & \sum_{i=1}^{n}\frac{x_i^2}{n} \end{pmatrix}$ PS1 Q3

* $E(x_i^2)$ must exist ($E(x_i)$ exists by existence of moments)

$$\text{plim}\left(\frac{X'X}{n}\right) \overset{Slutsky}{=} \begin{pmatrix} 1 & \text{plim}\frac{1}{n}\sum x_i \\ \text{plim}\frac{1}{n}\sum x_i & \text{plim}\frac{1}{n}\sum x_i^2 \end{pmatrix} \overset{WLLN}{=} \begin{pmatrix} 1 & E(x_i) \\ E(x_i) & E(x_i^2) \end{pmatrix} = D \quad (\text{assume } D^{-1} \text{ exists})$$

$$\text{plim}\left(\frac{X'\varepsilon}{n}\right) = \text{plim}\,\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i \overset{WLLN}{=} E(x_i\varepsilon_i) \overset{LIE}{=} E[E(x_i\varepsilon_i|X)] = E[x_i E(\varepsilon_i|X)] \overset{A3}{=} 0$$

$$\text{plim}\,\hat{\beta} \overset{Slutsky}{=} \beta + (\text{plim}\,\frac{X'X}{n})^{-1}\,\text{plim}\,\frac{X'\varepsilon}{n} = \beta$$

s^2 IS CONSISTENT $s^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} = \frac{\varepsilon' M \varepsilon}{n-k} = \frac{n}{n-k}\frac{\varepsilon' M \varepsilon}{n} = \frac{n}{n-k}\left[\frac{\varepsilon'\varepsilon}{n} - \frac{\varepsilon' X}{n}\left(\frac{X'X}{n}\right)^{-1}\frac{X'\varepsilon}{n}\right] \overset{Slutsky}{\underset{WLLN}{=}} \sigma^2$ PS5 extra

For proving biased/inconsistent due to heteroscedasticity, see 2018 2c

ASYMPTOTIC DISTRIBUTION

The asymptotic property of convergence in distribution enables hypothesis testing of an estimator without knowing its finite sample distribution — gives us a distribution we can use to approximate the estimator's distribution arbitrarily well for sufficiently large sample.

Formal definition: The sequence Z_n with CDF $F_{Z_n}(z)$ converges in distribution to a random variable Z with CDF $F_Z(z)$ if $\lim_{n\to\infty} |F_{Z_n}(z) - F_Z(z)| = 0$ at all points of continuity of $F_Z(z)$

Intuitively: Distribution of $Z_n \to$ Distribution of Z as $n \to \infty$

Example: $t_{n-k} \xrightarrow{n\to\infty} N(0,1)$

CENTRAL LIMIT THEOREM

If $X_1, \dots, X_n$ is a random sample (i.i.d) with finite mean μ and finite variance σ^2,

$$\underbrace{\sqrt{n}(\bar{X} - \mu)}_{\text{standardised}} \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - E(X_i)) \xrightarrow{d} N(0, \sigma^2)$$

"$\sqrt{n}(\bar{X} - \mu)$ has a $N(0,\sigma^2)$ limiting distribution"
↳ does not depend on sample size n

- If X_i normally distributed, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ exact
- If X_i unknown distribution, $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0,\sigma^2)$ by CLT
  - so $\bar{X} \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n})$ approximately

OLS ESTIMATOR

Even if we don't have A5 ($\varepsilon_i \sim N(0,\sigma^2)$), we can still rely on A1-4 and CLT to conclude $\hat{\beta} | X \overset{a}{\sim} N(\beta, \sigma^2(X'X)^{-1})$ if ε_i is iid, $E(\varepsilon_i) = 0$, $Var(\varepsilon_i) = \sigma^2$ finite

APPLICATION TO H-TEST

SINGLE LINEAR RESTRICTION $H_0: c'\beta = \gamma$

Under finite sample test, $t = \frac{(c'\hat{\beta} - \gamma)}{SE(c'\hat{\beta} - \gamma)} \sim t_{n-k}$ by A1-5 under H_0

Under asymptotic test, $Z = \frac{(c'\hat{\beta} - \gamma)}{SE(c'\hat{\beta} - \gamma)} \overset{a}{\sim} N(0,1)$ by A1-4 and CLT under H_0

- Doesn't matter if we know σ^2 or not — s^2 is a consistent estimator of σ^2

MULTIPLE LINEAR RESTRICTION $H_0: R\beta = c$ $\quad d = R\hat{\beta} - c$

Under finite sample test, $\frac{d'(\widehat{Var}(d))^{-1}d}{J} \sim F_{J,n-k}$ by A1-5 under H_0, $d \sim N(0, \sigma^2 R(X'X)^{-1}R')$

Under asymptotic test, $d'(\widehat{Var}(d))^{-1}d \overset{a}{\sim} \chi^2_J$ by A1-4 and CLT under H_0, $d \overset{a}{\sim} N(0, \sigma^2 R(X'X)^{-1}R')$

- Doesn't matter if we know σ^2 or not — s^2 is a consistent estimator of σ^2
- Asymptotically, no need to account for imprecision of s^2 with F-dist
- t and F test are exact tests that rely on the assumption of normality
- z and χ^2- test don't — use CLT

TIME-SERIES DATA

alt. title: Why A3 doesn't work

$\{(y_t, x_{t1}, \dots, x_{tk}) : t = 1, \dots, T\}$ where T is sample size

OLS estimator biased! Need large samples

Two assumptions needed for LLN, CLT and standard statistical inference to work
- Stationarity
- Weak dependence

MODELS USING TS DATA

STATIC

$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t$, $t = 1, \dots, T$ e.g. Philips curve

- Models a contemporaneous relationship between y and x (y_t doesn't depend on y_{t-1})
- Assumes a change in x_t at time t has immediate effect on y_t
- If A1-4, OLS is BLUE
  - But there is often autocorrelation – $\text{Cov}(\varepsilon_t, \varepsilon_s) \neq 0$ for some $t \neq s$
  - But as long as A3 $E(\varepsilon|X) = 0$, OLS still unbiased, and we can use RSE

FINITE DISTRIBUTED LAG MODELS

$y_t = \alpha + x_t\gamma_1 + x_{t-1}\gamma_2 + \varepsilon_t$ e.g. min W and UN^t

γ_1, γ_2 are distributed lag coefficients
- γ_1 is the contemporaneous effect – impact propensity
- $\gamma_1 + \gamma_2$ is the long run propensity – when x is changed permanently to a fixed value

$X = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots \\ x_t & x_{t-1} & x_{t-2} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$ (columns: income 2020, income 2019, income 2018)

If x_t changes slowly over time (x_t highly correlated to x_{t-1}), near multicollinearity may make estimates imprecise, but we can still estimate LR effect precisely

WEAKER EXOGENEITY

But for OLS to be unbiased, must assume strict exogeneity $E(\varepsilon_t|X) = E(\varepsilon_t|x_1, \dots, x_T) = 0$
- ε_t uncorrelated with all past, current and future x-values
- In reality x_{t+1} may respond to ε_t e.g. min W responding to past employment shocks · PS6 Q1
- More reasonable to assume $E(\varepsilon_t|x_t, x_{t-1}, \dots) = 0$
- Makes OLS of FDL/static models biased, but consistent under stationarity and weak dependence

Aim: $\text{plim}\,\hat{\beta} = \beta + \left[\text{plim}\left(\frac{X'X}{T}\right)\right]^{-1} \text{plim}\left(\frac{X'\varepsilon}{T}\right) = \beta$

Assume $\text{plim}\left(\frac{X'X}{T}\right) = D$ and D^{-1} exists. PS6

$$\text{plim}\,\frac{X'\varepsilon}{T} = \begin{pmatrix} \text{plim}\frac{1}{T}\sum \varepsilon_t \\ \text{plim}\frac{1}{T}\sum \varepsilon_t x_t \\ \text{plim}\frac{1}{T}\sum \varepsilon_t x_{t-1} \\ \text{plim}\frac{1}{T}\sum \varepsilon_t x_{t-2} \end{pmatrix} \overset{\text{WLLN}}{=} 0 \quad \text{if } E(\varepsilon_t|x_t, x_{t-1}, \dots) = 0$$

□

DYNAMIC: AUTOREGRESSIVE DISTRIBUTED LAG MODELS

$y_t = \alpha + \phi y_{t-1} + x_t\gamma_1 + x_{t-1}\gamma_2 + \varepsilon_t$, $|\phi| < 1$, $E(\varepsilon_t|\underbrace{x_t, x_{t-1}, \dots, y_{t-1}, y_{t-2}, \dots}_{\text{all avail info at time } t}) = 0$ (contemporaneous exogeneity) *"predeterminedness"

- Omitting y_{t-1} if y_{t-1} is correlated with x_t causes OVB (γ_1 captures additionally the effects of y_{t-1}); control for it!!
- γ_1 is the contemporaneous effect
- $\frac{\gamma_1 + \gamma_2}{1 - \phi}$ is the long run propensity in equilibrium (x^*, y^*), $\varepsilon_t = E(\varepsilon_t) = 0$ so $y^* = \alpha + \phi y^* + (\gamma_1 + \gamma_2)x^* \Rightarrow y^* = \frac{\alpha}{1-\phi} + \frac{\gamma_1+\gamma_2}{1-\phi}x^*$

$X = \begin{pmatrix} \vdots & \vdots & \vdots & \vdots \\ y_{t-1} & x_t & x_{t-1} & x_{t-2} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$ (columns: UN^2 2020, income 2020, income 2019, income 2018) $\quad \beta = \begin{pmatrix} \alpha \\ \phi \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$

$E(\varepsilon|X) = E(\varepsilon_t|\underbrace{x_1, \dots, x_T, y_1, \dots, y_t, \dots, y_{T-1}}_{\text{strict exogeneity}}) = 0$ cannot be true

Since $\text{Cov}(\varepsilon_t, y_t) \neq 0$ by definition. $\hat{\beta}_{ols}$ biased! PS6 Q2

- With contemporaneous exogeneity, $\hat{\beta}_{ols}$ consistent under stationarity and weak dependence (same proof as above)
- Not consistent under ε_t autocorrelation

STATIONARITY

replaces identically distributed

Joint dist of $(x_{t_1},\dots,x_{t_m})$ identical to that of $(x_{t_1+h},\dots,x_{t_m+h})$ $\forall h$

- Actually stronger than identically distributed (id $\not\Rightarrow$ Cov stat)

DEFINITION A stochastic process $\{x_t : t=1,2,\dots\}$ is covariance stationary if

1. $E(x_t)$ is finite and constant (i.e. same $\forall t$)
2. $Var(x_t)$ is finite and constant (i.e. same $\forall t$)
3. $Cov(x_t, x_{t+h})$ is finite and depends only on distance in time, h (i.e. same $\forall t$ holding h constant)
   - sometimes $t-h$ (for eqn manipulation)

if you "know" it is cov stat, let $h=0$ first to derive the cov (veri

- Implies the first two moments must exist and not change over time
- Can allow for deterministic (non-stochastic) trending behaviour that do not affect dependence of process over time
  - Just detrend it first!

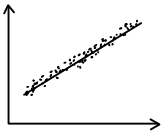
DETERMINISTIC TREND

$y_t = \alpha+\beta t+\varepsilon_t,\ t=1,\dots,T$

$E(y_t) = \alpha+\beta t$ changes with t

$Var(y_t) = Var(\varepsilon_t)$

$Cov(y_t, y_{t+h}) = Cov(\varepsilon_t, \varepsilon_{t+h})$



If ε_t is covariance stationary, the detrended $y_t-\beta t=\alpha+\varepsilon_t$ is covariance stationary (FWL)

WEAK DEPENDENCE

replaces independence in TS data

Weak dependence requires $Corr(x_t, x_{t+h}) = \dfrac{Cov(x_t,x_{t+h})}{\sqrt{Var(x_t)Var(x_{t+h})}} \to 0$ as $h\to\infty$ (asymptotic independence)

AUTOCORRELATION FUNCTION $\rho(h) = Corr(x_t, x_{t+h})$ (Sample analogue is the correlogram)

- $\rho(0) = Corr(x_t,x_t)=1$

COMMON DEPENDENCE PROCESSES

AUTOREGRESSIVE PROCESS of order 1 AR(1)

$y_t = \phi y_{t-1}+\varepsilon_t,\ \varepsilon_t \sim i.i.d(0,\sigma^2)$ (white noise)

Stationarity requires $|\phi|<1$ (>0: no finite mean var; =0: random walk)

- By recursive substitution, $y_t = \varepsilon_t+\phi\varepsilon_{t-1}+\phi^2\varepsilon_{t-2}+\dots$ as $\phi^s\to 0$
- $Var(y_t) = Var(\varepsilon_t)+\phi^2Var(\varepsilon_{t-1})+\phi^4Var(\varepsilon_{t-2})+\dots = (1+\phi^2+\phi^4+\phi^6+\dots)\sigma^2$ converges iff $|\phi|<1$
  - if an intercept is added, mean will diverge too
- $\rho(h)=\phi^h$: a shock affects all future obs with decreasing (geometric) effect
- $E(y_t)=0$, $Var(y_t)=\frac{\sigma^2}{1-\phi^2}$, $Cov(y_t,y_{t+h})=\phi$

MOVING AVERAGE PROCESS of order 1 MA(1)

$y_t = \varepsilon_t+\theta\varepsilon_{t-1},\ \varepsilon_t\sim i.i.d(0,\sigma^2)$

- Always stationary
- $\rho(0)=1$, $\rho(1)=\frac{\theta}{1+\theta^2}$, $\rho(h)=0\ \forall h=2,3,\dots$ Shock only affects y_t in 2 periods
- $E(y_t)=0$, $Var(y_t)=\sigma^2(1+\theta^2)$, $Cov(y_{t+1},y_t)=\theta\sigma$, $Cov(y_{t+h},y_t)=0\ \forall h>1$ PS6 extra Q1

AUTOREGRESSIVE MOVING AVERAGE PROCESS ARMA(1,1)

$y_t = \phi y_{t-1}+\varepsilon_t+\theta\varepsilon_{t-1},\ \varepsilon_t\sim i.i.d(0,\sigma^2)$

HETEROSCEDASTICITY & AUTOCORRELATION

DEFINITION

A1 $y = x\beta + \varepsilon$, $E(\varepsilon) = 0$

A2 No perfect multicollinearity

A3 $E(\varepsilon|X) = 0$

A4* General covariance matrix $Var(\varepsilon|X) \overset{A3}{=} E(\varepsilon\varepsilon'|X) = \Sigma$

Σ is a symmetric positive semidefinite matrix

often written as $\Sigma = \sigma^2\Omega$ (σ^2: unknown scaling parameter)

HETEROSCEDASTICITY $Cov(\varepsilon_i, \varepsilon_j|X) = \begin{cases} \sigma_i^2 & \text{if } i=j \\ 0 & \text{if } i \neq j \end{cases}$ $\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \sigma_n^2 \end{pmatrix}$

AUTOCORRELATION $Cov(\varepsilon_t, \varepsilon_s|X) \neq 0$ for some $t \neq s$

e.g. AR(1): if $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$, $|\rho| < 1$, $v_t \sim iid(0, \sigma^2)$

$\Sigma = \frac{\sigma^2}{1-\rho^2}\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \cdots & \rho^{n-1} \\ \rho & 1 & \cdots & & & \rho^{n-2} \\ \rho^{n-1} & \rho^{n-2} & & & \ddots & 1 \end{pmatrix}$ "fading memory" — homoscedastic, but autocorrelated (by cov stationarity)

ISSUE $\hat{\beta}_{OLS}$ not BLUE!

A1-3 Unbiased ✓ Consistent ✓ Linear ✓ But efficiency loss ✗

$Var(\hat{\beta}|X) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$ given A.4*

$= \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$ Standard t- and F-tests invalid

SOLUTION

① Stick with $\hat{\beta}_{OLS}$ but use correct (RSE) estimator of covariance matrix (heteroscedasticity-robust inference)

- ✓ Don't have to specify form of heteroscedasticity/serial correlation
- ✗ Lose efficiency

② Derive alternative estimator $\hat{\beta}_{GLS}$ that is BLUE

- ✓ Regain efficiency (asymptotically)
- ✗ Must specify form of heteroscedasticity/serial correlation

APPROACHES

Heteroscedasticity only
$Var(\varepsilon_i|X) = \sigma_i^2 \neq \sigma^2$, $Cov(\varepsilon_i, \varepsilon_j|X) = 0$

- know form of heteroscedasticity (BLUE estimator)
  - $\{\sigma_i^2\}_{i=1}^n$ known → WLS, t-stats exact ($\sim N, \chi^2$)
  - $\{\sigma_i^2\}_{i=1}^n$ unknown → FWLS (consistent $\hat{\sigma}_i^2$), t-stats approximate ($\sim t, F$)
- don't know form (OLS consistent) → OLS with White SEs, t-stats asymptotically distributed ($\overset{a}{\sim} N, \chi^2$)

Autocorrelation too — OLS consistent → OLS with HAC/Newey-West SEs, t-stats asymptotically distributed ($\overset{a}{\sim} N, \chi^2$)

Note: Should only use RSEs when there's evidence of heteroscedasticity/autocorrelation

HETEROSCEDASTICITY-ROBUST INFERENCE

When there is heteroscedasticity but no autocorrelation,

$$\text{Var}(\hat{\beta}|X) = \left(\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\sum_{i=1}^n \sigma_i^2 x_i x_i'\right)\left(\sum_{i=1}^n x_i x_i'\right)^{-1} \quad \blacksquare \quad x_j' = (x_{j1} \cdots x_{jk})$$

When $\{\sigma_i^2\}_{i=1}^n$ is unknown

$$\widehat{\text{Var}}(\hat{\beta}|X) = \left(\sum_{i=1}^n x_i x_i'\right)^{-1}\left(\sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i'\right)\left(\sum_{i=1}^n x_i x_i'\right)^{-1} \quad \text{HCSE/ White SE}$$

- ✡ Consistent but only valid asymptotically
- ✡ Does not imply $\hat{\varepsilon}_i^2$ is a consistent estimator of σ_i^2 but that $\text{plim}\left(\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon}_i^2 x_i x_i' - \frac{1}{n}\sum_{i=1}^n \sigma_i^2 x_i x_i'\right)$ (averages)

WEIGHTED LEAST SQUARES

With weights $r_i = \frac{1}{\sigma_i^2}$

Satisfies A1-4 $\frac{y_i}{\sigma_i} = \left(\frac{x_i}{\sigma_i}\right)'\beta + \frac{\varepsilon_i}{\sigma_i}$ same as having $R = \begin{pmatrix} \frac{1}{\sigma_1} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sigma_n} \end{pmatrix}$

e.g. A3: $\text{Var}\left(\frac{\varepsilon_i}{\sigma_i}|X\right) = \frac{1}{\sigma_i^2}\text{Var}(\varepsilon_i|X) = 1$ (constant)

$$\hat{\beta} = \left[\sum_{i=1}^n \frac{x_i x_i'}{\sigma_i^2}\right]^{-1}\sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2}$$

can also replace σ_i with weights ω_i s.t. $\sigma_i^2 = \sigma^2\omega_i$
if we know σ_i^2, just do standard t/F-tests

FEASIBLE WLS

$= \text{Var}(\varepsilon_i|x_i) \overset{A3}{=} E(\varepsilon_i^2|x_i)$

Don't know σ_i^2? Parameterise σ_i^2 to get $\hat{\sigma}_i^2$

$$\hat{\beta}_{FGLS} = \left[\sum_{i=1}^n \frac{x_i x_i'}{\hat{\sigma}_i^2}\right]^{-1}\sum_{i=1}^n \frac{x_i y_i}{\hat{\sigma}_i^2}$$

One way to parameterise σ_i^2

LINEAR HETEROSCEDASTICITY $\sigma_i^2 = \delta_0 + z_i'\delta_i = E(\varepsilon_i^2|x_i)$ z_i is some function of x_i

✡ σ_i^2 can be negative !!

$\hat{\varepsilon}_i^2$ is a good guess of ε_i^2, which on average is σ_i^2

Then we find consistent $\hat{\delta}_0, \hat{\delta}_1$

DERIVING $\varepsilon_i^2 = \delta_0 + z_i'\delta_1 + v_i$

$E(\varepsilon_i^2|x_i) = \delta_0 + z_i'\delta_1 + \underbrace{E(v_i|x_i)}_{=0}$

$E(v_i|x_i) = E(v_i|z_i) = 0$

$\Rightarrow$ assume GM $\Rightarrow$ Regress $\hat{\varepsilon}$ on Z to get BLUE $\hat{\beta}_0, \hat{\beta}_1$

TESTING $H_0: \delta_0 = \delta_1 = 0$ vs $H_1: \neg H_0$ asymptotic bc we're using $\hat{\varepsilon}_i^2$, not ε_i^2

z_i is some function of x_i

EXPONENTIAL HETEROSCEDASTICITY $\sigma_i^2 = e^{\delta_0 + z_i'\delta_1} = \sigma^2 e^{z_i'\delta_1}$ $\sigma^2 = e^{\delta_0}$ is constant of proportionality

Regress $\log\hat{\varepsilon}_i = e^{\delta_0 + z_i'\delta_0} + e_i$

GENERALISED LINEAR MODEL

LM but relaxes A4 $Var(\varepsilon|X)=\sigma^2 I$

Find square, invertible matrix R s.t. $R'R=\Omega^{-1}$. Such an R exist 

Transform model $Ry = RX\beta + R\varepsilon$ $(y^*=X^*\beta+\varepsilon^*)$

Then A1-4 are satisfied $\Rightarrow$ OLS on transformed model is BLUE 

MINIMISE GENERALISED SUM OF SQUARES

$\hat{\beta}_{OLS} = \arg\min_b S(b)$, $S(b) = (y-Xb)'\Omega^{-1}(y-Xb)$

quadratic form suitably weighted by precision

can also use $(y-Xb)'\Sigma^{-1}(y-Xb) = \varepsilon' Var(\varepsilon|X)^{-1}\varepsilon$

Accounts for the fact that some observations are associated with higher variability than others by giving their discrepancies (residuals) less weight

When S(b) is at minimum,

Proof

$$\frac{\partial S(b)}{\partial b} = \begin{pmatrix} \frac{\partial S(b)}{\partial b_1} \\ \vdots \\ \frac{\partial S(b)}{\partial b_k} \end{pmatrix} = 2y'\Omega^{-1}X - 2X'\Omega^{-1}Xb = 0 \Rightarrow \hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$$

but typically infeasible as we don't know Ω replace with $\hat{\Omega}$ (feasible GLS)

or $\hat{\beta}_{GLS} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$

PROPERTIES OF $\hat{\beta}_{GLS}$

Derive from transformed model $y^* = X\beta + \varepsilon^*$, $\begin{cases} E(\varepsilon^*|X^*)=0 \\ Var(\varepsilon^*|X^*)=0 \end{cases}$

$E(\hat{\beta}_{GLS}) = \beta$ PST

$Var(\hat{\beta}_{GLS}|X) = \sigma^2(X^{*\prime}X^*)^{-1} = \sigma^2(X'\Omega^{-1}X)^{-1}$ PST

PROPERTIES OF $\hat{S}^2_{GLS}$

$$\hat{S}^2_{GLS} = \frac{\hat{\varepsilon}^{*\prime}\hat{\varepsilon}^*}{N-k} = \frac{(y-X\hat{\beta}_{GLS})'\Omega^{-1}(y-X\hat{\beta}_{GLS})}{N-k}$$

Unbiased: $\hat{S}^2_{GLS} = \frac{\varepsilon^{*\prime}M\varepsilon^*}{N-k}$, $M = I - X^*(X^{*\prime}X^*)^{-1}X^{*\prime}$, proceed

If A5: $\varepsilon|X \sim N(0,\sigma^2\Omega)$, then $\hat{\beta}_{GLS}|X \sim N(\beta, \sigma^2(X'\Omega^{-1}X)^{-1})$ and $\hat{\beta}_{GLS} \perp\!\!\!\perp S^2_{GLS}$

FEASIBLE GLS

$$\hat{\beta}_{FGLS} = (X'\hat{\Omega}^{-1}X)^{-1}X'\hat{\Omega}^{-1}y$$

e.g. AR1 $\Omega(\phi) = \frac{\sigma^2}{1-\phi^2}\begin{pmatrix} 1 & \phi & \phi^2 & \phi^3 & \cdots & \phi^{n-1} \\ \phi & 1 & \cdots & & & \phi^{n-2} \\ \vdots & & \ddots & & & \\ \phi^{n-1} & \phi^{n-2} & \cdots & & & 1 \end{pmatrix}$

To derive a consistent $\hat{\Omega}$, we must parameterise (put structure on) Ω in terms of a finite-dimensional parameter vector θ $\Omega = \Omega(\theta)$. Then use $\hat{\varepsilon} = y - X\hat{\beta}$ to derive $\hat{\theta}$, s.t. $plim(\hat{\theta}) = \theta$, so $plim\,\hat{\Omega} = \Omega(plim\,\hat{\theta}) = \Omega(\theta)$

e.g. if Ω is diagonal, $\Omega_{ii} = z_i'\gamma$ ← unobserved

1. Regress y on X to get $\hat{\varepsilon} = y - X\hat{\beta}_{OLS}$
2. Regress $\hat{\varepsilon}_i^2$ on z_i to get $\hat{\gamma}$ (since $E(\varepsilon_i^2) = Var(\varepsilon_i) + [E(\varepsilon_i)]^2 = Var(\varepsilon_i) = z_i\gamma$)
3. Thus $\hat{\Omega}_{ii} = z_i'\hat{\gamma}$, $\hat{\Omega} = \begin{pmatrix} \hat{\Omega}^2_{11} & & \\ & \ddots & \\ & & \hat{\Omega}^2_{nn} \end{pmatrix}$

$\hat{\beta}_{FGLS}$ ASYPTOTIC PROPERTIES

$\hat{\beta}_{FGLS}$ is usually biased, and isn't even linear ($\hat{\varepsilon}$, which is a function of y, is used to derive $\hat{\Omega}$, so y enters > once) $\hat{\Omega}$ so $\hat{\beta}_{FGLS}$ is not BLUE; shouldn't be used with small N too

Asymptotic normal: $\hat{\beta}_{FGLS}|X \overset{a}{\sim} N(\beta, \sigma^2(X'\Omega^{-1}X)^{-1})$

- We have to use this property even if we assume/know $\varepsilon|X \sim N(0,\sigma^2\Omega)$ because we're also using $\hat{\Omega}$
- Only asymptotically does $\hat{\beta}_{FGLS}$ inherit the desirability of $\hat{\beta}_{GLS}$

TEST HETEROSCEDASTICITY

WALD TEST (FGLS) SETTING

Test for heteroscedasticity of specific form

$\begin{cases} H_0: \sigma_i^2 = \sigma^2 \ \forall i \\ H_A: \sigma_i^2 = \delta_0 + z_i \delta_1 \end{cases} \Leftrightarrow \begin{cases} H_0: \delta_1 = 0 \\ H_A: \text{at least one } \delta_1 \text{ is non-zero} \end{cases}$

or $\sigma_i^2 = e^{\delta_0} e^{z_i \delta_1}$

$\Leftrightarrow$ Test significance of the regression $\hat{\varepsilon}_i^2 = \delta_0 + z_i' \delta_1 + v_i$

or $\ln(\hat{\varepsilon}_i^2) = \delta_0 + z_i' \delta_1 + v_i$

$\Leftrightarrow$ Read F-stat, or (more accurately since we don't have ε_i^2)

calculate $\hat{\delta}_1' (\widehat{Var}(\hat{\delta}_1))^{-1} \hat{\delta}_1 \overset{a}{\sim} \chi_p^2$, where $p = \dim(\delta_1)$

↑ asymptotically!

BREUSH-PAGAN TEST

An example of a Lagrange Multiplier test

$\begin{cases} H_0: \sigma_i^2 = \sigma^2 \ \forall i \\ H_A: \sigma^2 h(\delta_0 + z_i' \delta_0) \end{cases}$ unknown, continuously differentiable

sample size

$\mathcal{L} = nR^2 \overset{a}{\sim} \chi_p^2$ under H_0 where $p = \dim(\delta_1)$

Reject H_0 if $nR^2 > \chi_{p,\alpha}^2$

R^2 is goodness-of-fit from auxiliary regression $\hat{\varepsilon}_i^2 = \delta_0^* + z_i' \delta_1^* + v_i$ (OLS residual)

- High $R^2 \Rightarrow z_i$ and $\hat{\varepsilon}_i$ very related $\Rightarrow$ heteroscedasticity
- Cannot estimate δ_1 and do Wald $\hat{\delta}_1' (\widehat{Var}(\hat{\delta}_1))^{-1} \hat{\delta}_1$ since form is not explicit
- Also cannot do F test on δ_i^* since form is not explicit

WHITE TEST

No idea of potential form of heteroscedasticity

$\begin{cases} H_0: \sigma_i^2 = \sigma^2 \ \forall i \\ H_A: \neg H_0 \end{cases} \Leftrightarrow \begin{cases} H_0: E(\varepsilon_i^2 | x_i) = \sigma^2 \ \forall i \\ H_A: \neg H_0 \end{cases}$ if homoscedastic, then x_i/functions of x_i shouldn't explain $E(\varepsilon_i^2 | x_i)$

Run auxiliary regression $\hat{\varepsilon}_i^2 = \gamma_0 + z_i' \gamma_1 + v_i$

Test $\gamma_1 = 0$, compute R^2. $nR^2 \overset{a}{\sim} \chi_p^2$ where $p = \dim(\gamma_1)$

Reject H_0 if $nR^2 > \chi_{p,\alpha}^2$

- z_i may include some/all variables in x_i and other variables that depends on x_i
- White: "use x_i, and set of all unique squares and cross products of variables in x_i"
  - e.g. $x_{i2}, x_{i3}, x_{i2}^2, x_{i3}^2, x_{i2}x_{i3}$
  - But cannot use x_{i2}^2 if x_{i2} is a dummy

AUTOCORRELATION-ROBUST INFERENCE

$Var(\hat{\beta}|X) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$

USE OLS WITH RSE Only when OLS is consistent

- At least contemporaneous exogeneity $E(\varepsilon_t|x_t) = 0$
- Not useful with lagged y-values – use Newey-West/ HAC RSE instead (consistent)
  - When $Cov(\varepsilon_t, \varepsilon_{t+s}|X) = \gamma_s$, with $\gamma_s = 0$ when $s > lag$ Woolridge (v technical)
  - HAC also allows lag to ↑ as T↑
  - Stata: newey y x1 - xk, lag (insert no here)
    ↑ R can auto-det best lag

TEST FOR AUTOCORRELATION

Must specify autocor. model, e.g. AR(1) $\varepsilon_t = \rho\varepsilon_{t-1} + v_t$

$H_0: \rho = 0$ vs $H_A: \rho \neq 0$ or $\rho > 0$

Use $\hat{\varepsilon}_t$ since we can't observe $\{\varepsilon_t\}$

UNDER STRICT EXOGENEITY

① Estimate $y_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt} + u_t$ by OLS, save $\{\hat{u}_t\}$
② Run AR(1) $\hat{u}_t$ on $\hat{u}_{t-1}$
③ $\frac{\hat{\rho}}{SE(\hat{\rho})} \overset{a}{\sim} N(0,1)$ Asymptotic t-test as we use $\hat{u}_t$, not u_t
★ Can also add lags – regress $\hat{u}_t$ on $\hat{u}_{t-1}, \hat{u}_{t-2}$, test for joint significance (AR(2), AR(3))

UNDER CONTEMPORANEOUS EXOGENEITY

Same as above but for ②
Run OLS of $\hat{u}_t$ on $\hat{u}_{t-1}, x_{t1}, \dots, x_{tk}$ ← controls
- Account for the fact that u_{t-1} might be correlated with $x_{t1}, \dots, x_{tk}$ might respond to past shocks not strictly exogenous
- Use if x_{tj}'s contain lagged y

NOTE Not tested

We note that $\begin{cases} y_t = \alpha_0 + \beta x_t + \varepsilon_t \\ \varepsilon_t = \rho\varepsilon_{t-1} + u_t \text{ AR1} \end{cases}$ $|\rho| < 1$, $u_t \sim (0, \sigma^2)$ i.i.d is GM error

Transform model to regain efficiency: $y_t - \rho y_{t-1} = (1-\rho)\alpha + \beta x_t - \rho\beta x_{t-1} + \varepsilon_t - \rho\varepsilon_{t-1}$

$= (1-\rho)\alpha + \beta(x_t - \rho x_{t-1}) + u_t$ is GM!!
(new regressor)

- We dk ρ, so make it feasible by regressing on residuals
- Will lose first obs

ENDOGENEITY

Correlation between errors and regressors
$E(\varepsilon|X) \neq 0$

OVB
lagged dep var when errors exhibit dep
regressor measurement error
simultaneity

CONSEQUENCE FOR OLS

biased $E(\hat{\beta}|X) = \beta + (X'X)^{-1}X'E(\varepsilon|X) \neq \beta$

inconsistent $\text{plim}\,\hat{\beta} = \beta + \text{plim}\left(\frac{X'X}{n}\right)^{-1} \text{plim}\left(\frac{X'\varepsilon}{n}\right) \neq \beta$

$\text{plim}\left(\frac{X'\varepsilon}{n}\right) = E(x_i\varepsilon_i)$ by LLN

Not surprising as FOC of OLS require $X'\hat{\varepsilon} = \sum_{i=1}^{n} x_i\hat{\varepsilon}_i = 0$, which is proportional to $\frac{1}{n}\sum_{i=1}^{n} x_i\hat{\varepsilon}_i = 0$, sample analogue of $E(x_i\varepsilon_i) = 0$. If $E(x_i\varepsilon_i) \neq 0$ then imposing orthogonality is inappropriate

LAGGED DEP VARS & SERIALLY CORREL ERRORS

$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t, \quad |\beta_3| < 1, \quad t = 1, \ldots, T$
$\varepsilon_t = \rho\varepsilon_{t-1} + v_t, \quad |\rho| < 1 \quad v_t \overset{iid}{\sim} (0, \sigma_v^2)$

$\text{Cov}(y_{t-1}, \varepsilon_t) = E(y_{t-1}\varepsilon_t) \neq 0$ so y_{t-1} is endogenous

$$\begin{aligned}\text{Cov}(y_{t-1}, \varepsilon_t) &= \text{Cov}(\rho\varepsilon_{t-1} + v_t,\ \beta_1 + \beta_2 x_{t-1} + \beta_3 y_{t-2} + \varepsilon_{t-1}) \\ &= \rho\beta_2\text{Cov}(\varepsilon_{t-1}, x_{t-1}) + \rho\beta_3\text{Cov}(\varepsilon_{t-1}, y_{t-2}) + \rho\text{Var}(\varepsilon_{t-1}) \\ &= \rho\beta_3\text{Cov}(\varepsilon_t, y_{t-1}) + \rho\sigma^2 \quad (\text{Cov stat})\end{aligned}$$

$$\Rightarrow \text{Cov}(\varepsilon_t, y_{t-1}) = \frac{\rho\sigma^2}{1-\rho\beta_3} \neq 0 \quad \text{assuming covariance stationarity } \rho\beta_3 \neq 1$$

$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ inconsistent: let $X = \begin{pmatrix} \top & \top & \top \\ \perp & \perp & \perp \\ 1 & x_t & y_{t-1} \end{pmatrix}$ then $\text{plim}\,\frac{X'\varepsilon}{T} = \text{plim}\begin{pmatrix} \frac{1}{T}\sum\varepsilon_t \\ \frac{1}{T}\sum x_t\varepsilon_t \\ \frac{1}{T}\sum y_{t-1}\varepsilon_t \end{pmatrix} \overset{LLN}{=} \begin{pmatrix} E(\varepsilon_{t-1}) \\ E(\varepsilon_t x_t) \\ E(\varepsilon_t y_t) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \text{Cov}(y_{t-1}, \varepsilon_t) \end{pmatrix}$

MEASUREMENT ERROR

True model: $y_i = x_i^{*\prime}\beta + \varepsilon_i$
Mismeasurement: $x_i = x_i^* + u_i$
Mismeasured model: $y_i = x_i'\beta + \underbrace{v_i}_{\text{composite error}} = x_i^{*\prime}\beta + \varepsilon_i - u_i'\beta$

CLASSICAL MEASUREMENT ERROR ASSUMPTION

x_i^*: true value of regressor being measured

u_i is independent of x_i^* and ε_i^* $\quad (u_i \sim \text{i.i.d}(0, \sigma_u^2))$

$\hat{\beta}$ inconsistent as $\text{plim}\,\hat{\beta} \overset{\text{Slutsky}}{=} \beta + \left(\text{plim}\left(\frac{X'X}{n}\right)\right)^{-1}\text{plim}\left(\frac{X'v}{n}\right) \overset{WLLN}{=} \beta + E(x_i x_i')^{-1}E(x_i v_i) \neq \beta$ because

$$\begin{aligned}E(x_i v_i) &= E\left((x_i^* + u_i)(\varepsilon_i - u_i'\beta)\right) \\ &= E(x_i^*\varepsilon_i) + E(u_i\varepsilon_i) - E(x_i^* u_i')\beta - E(u_i u_i')\beta \\ &= -E(u_i u_i')\beta \neq 0 \quad \text{by CM / assume } x_i^* \text{ fixed}\end{aligned}$$

$E(x_i x_i') = E(x_i^* x_i^{*\prime}) + E(u_i u_i')$

In general all parameters will be affected by measurement error in one/more regressors

Attenuation bias: when just one regressor has CME, estimated parameters will be closer to 0 PS8 Q1

SIMULTANEITY X jointly determined with y in the same economic model

STRUCTURAL FORM gives relationships of interest (behavioural relations)

$$\begin{cases} \text{Demand} & q_t = \alpha_1 + \alpha_2 p_t + \alpha_3 m_t + u_{1t} \\ \text{Supply} & q_t = \beta_1 + \beta_2 p_t + \beta_3 c_t + u_{2t} \end{cases}$$

m_t: exogenous/predetermined — independent of ε_t

p_t: usually correlated

If u_{1t}, u_{2t} correlated then $\text{Cov}(u_{1t}, u_{2t}) = \sigma_{12}$

$$\begin{cases} C_t = \alpha + \beta Y_t + \varepsilon_t, \quad E(\varepsilon_t) = 0 \\ Y_t = C_t + I_t \end{cases}$$

OLS (α, β) inconsistent since $\text{Cov}(Y_t, \varepsilon_t) = \frac{1}{1-\beta}\sigma_\varepsilon^2 \neq 0$

REDUCED FORM expresses endo var in terms of exo vars and errors

$$Y_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} I_t + \frac{1}{1-\beta}\varepsilon_t \quad , \beta \neq 1$$

$$\begin{cases} p_t = \pi_{11} + \pi_{12} m_t + \pi_{13} c_t + v_{1t} \\ q_t = \pi_{21} + \pi_{22} m_t + \pi_{23} c_t + v_{2t} \end{cases}$$

v_{1t} & v_{2t}: fn of u_{1t} & u_{2t}

IV ESTIMATION

$y = X\beta + \varepsilon$ $\quad \text{Cov}(X, \varepsilon) \neq 0$

ASSUMPTIONS

Validity: $\text{Cov}(z_t, \varepsilon_t) = E(z_t \varepsilon_t) = 0$ not correlated with ε. Implied by $E(\varepsilon_t | z_t) = 0$

Relevance: $E(z_t' x_t)$ has rank $k \Rightarrow$ if square, invertible. Z is correlated with X

- # instruments ≥ # regressors (order condition), if not rank < k
- Instruments cannot have a direct effect on y for full rank

Our IV estimator $\hat{\beta}_{IV}$ is given by the sample analogue of the moment condition $E(z_i \varepsilon_i) = 0$ (validity)

$$\frac{1}{n} Z' \hat{\varepsilon}^{IV} = \frac{1}{n} Z'(y - X\hat{\beta}_{IV}) = 0$$

If $Z'X$ is square and invertible (just-identified)

$$\hat{\beta}_{IV} = (Z'X)^{-1} Z'y$$

if all regressors are exogenous, just take $Z = X$

need to have enough instruments (order condition)

(Calculus - bivariate) $\hat{\beta}_{1,IV} = \beta_1 + \sum_{i=1}^{n} d_i u_i$, $d_i = \frac{z_i - \bar{z}}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$ ← $\neq 0$ by relevance

BIASED BUT CONSISTENT

CALCULUS

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $E(\varepsilon_i) = 0$ but $E(\varepsilon_i x_i) \neq 0$

$\hat{\beta}_{1,IV} = \beta_1 + \sum_{i=1}^{n} d_i \varepsilon_i$,

$d_i = \frac{z_i - \bar{z}}{\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})}$ ($\neq 0$ by relevance)

LIE $E(\hat{\beta}_{1,IV} | x, z) = \beta_1 + \sum_{i=1}^{n} d_i E(\varepsilon_i | x, z) \neq \beta$ (correlated..) is biased

$\text{plim}\, \hat{\beta}_{IV} = \beta + \frac{\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(\varepsilon_i - \bar{\varepsilon})\right)}{\text{plim}\left(\frac{1}{n}\sum_{i=1}^{n}(z_i - \bar{z})(x_i - \bar{x})\right)} \overset{LLN}{=} \beta + \frac{\text{Cov}(z_i, \varepsilon_i)}{\text{Cov}(z_i, x_i)} = \beta$ ($=0$ validity, $\neq 0$ relevance) is consistent

$\text{Var}(\hat{\beta}_{1,IV} | x, z) = \text{Var}\left(\beta_1 + \sum_{i=1}^{n} d_i \varepsilon_i | x, z\right) = \sum_{i=1}^{n} d_i^2 \text{Var}(\varepsilon_i | x, z)$ assuming independence

$= \sigma^2 \sum_{i=1}^{n} d_i^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \frac{1}{R^2_{zx}} = \frac{\sigma^2}{n\widehat{\text{Var}}(x_i)} \frac{1}{R^2_{zx}}$ assuming conditional heteroscedasticity, no autocorrelation

For more precision: $\uparrow n$ $\downarrow \sigma^2$ $\uparrow$ regressor variability $\uparrow$ strong first stage effect

Use exogenous variables as instruments for themselves!

LINEAR ALGEBRA

$\hat{\beta}_{IV} = (Z'X)^{-1} Z'y$

LIE $E(\hat{\beta}_{IV} | X, Z) = \beta + (Z'X)^{-1} Z' E(\varepsilon | X, Z) \neq \beta$ (correlated..) is biased

$\text{plim}\, \hat{\beta}_{IV} = \beta + (\text{plim}\, \frac{1}{n} Z'X)^{-1} \text{plim}\, \frac{1}{n} Z'\varepsilon = \beta$ is consistent

- $\text{plim}\, \frac{1}{n} Z'X \overset{\text{iid (LLN)}}{=} M_{ZX} = E(z_i x_i')$ invertible by relevance and exclusion
- $\text{plim}\, \frac{1}{n} Z'\varepsilon \overset{LLN}{=} E(z_i \varepsilon_i) = 0$ by validity

$\text{Var}(\hat{\beta}_{IV}) = \sigma^2 (Z'X)^{-1} Z'Z (X'Z)^{-1}$

- $\frac{X'Z}{n}$ could be viewed as the matrix of sample covariances between Z and X
- Weak first stage will give $(X'Z)^{-1}$ huge elements and give $\hat{\beta}_{IV}$ MMse

By CLT, $\tilde{\beta}_{IV} \overset{a}{\sim} N(\beta, \sigma^2 (Z'X)^{-1} Z'Z (X'Z)^{-1})$

Assuming conditional A4, $\text{Var}(\varepsilon | Z) = \sigma^2 I$

$S^2_{IV} = \frac{RSS}{n-k} = \frac{(y - X\hat{\beta}_{IV})'(y - X\hat{\beta}_{IV})}{n-k}$

EXCLUSION RESTRICTION

If d_t has a direct effect on y_t in the true model: $y_t = \alpha + \beta w_t + \gamma d_t + \varepsilon_t$, $\text{Cov}(w_t, \varepsilon_t) = E(w_t \varepsilon_t) \neq 0$

Let $x_t = \begin{pmatrix} 1 \\ d_t \\ w_t \end{pmatrix}$. If we use the instrument $z_t = \begin{pmatrix} 1 \\ d_t \\ d_t \end{pmatrix}$, then $E(z_t x_t') = \begin{pmatrix} 1 & E(d_t) & E(d_t) \\ E(d_t) & E(d_t^2) & E(d_t^2) \\ E(w_t) & E(w_t d_t) & E(w_t d_t) \end{pmatrix}$ rank < 3

- Thus instruments are variables that don't enter the regression equation itself
- In simultaneous equation models, instruments often come from other equations in the system
- Lags of the included exogenous variables may provide suitable lagged dependent variables

OPTIMAL INSTRUMENT

When overidentified (# instrument > # regressors)

Best choice Z^{opt} is obtained by using the fitted values of an OLS regression of the columns of X on all instruments $Z^{opt} = \hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$ (projects all endo vars to all exo vars)

- Lin. comb. of all instruments (✓valid) with the highest correlation (✓relevant)

☆ if var in X is exogenous (i.e. it is its own inst) then $\hat{X}$ perfectly predicts (just returns the var)

- Basically regress $X = Z\delta + w$ to get $\hat{X} = Z\hat{\delta}$.

Then $\hat{\beta}_{IV} = (Z^{opt\prime}X)^{-1} Z^{opt\prime} y = (X'P_Z X)^{-1} X'P_Z y \overset{PS8}{=} (\hat{X}'X)^{-1}\hat{X}'y$

IV FOR LAGGED Y WITH AUTOCORREL

$y_t = \beta_1 + \beta_2 x_t + \beta_3 y_{t-1} + \varepsilon_t, \quad |\beta_3| < 1, \quad t = 1, \dots, T$
$\varepsilon_t = \rho \varepsilon_{t-1} + v_t, \quad |\rho| < 1 \quad v_t \overset{iid}{\sim} (0, \sigma_v^2)$

Instrument $\begin{pmatrix} 1 \\ x_t \\ y_{t-1} \end{pmatrix}$ with $\begin{pmatrix} 1 \\ x_t \\ x_{t-1} \end{pmatrix}$ $\quad \hat{X} = \begin{pmatrix} 1 \\ x_t \\ \hat{y}_{t-1} \end{pmatrix}$

Can also use more lagged values $\begin{pmatrix} 1 \\ x_t \\ x_{t-1} \\ x_{t-2} \\ \vdots \end{pmatrix}$ to ↑ correlation, but also ↑ bias

---

Note: What if supply/demand shifters don't exist

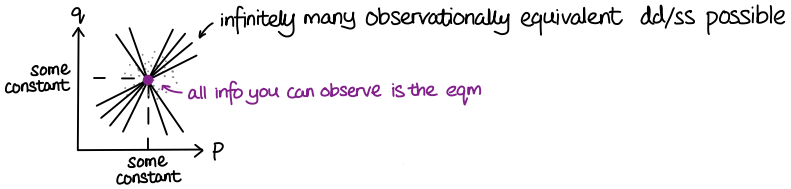
$$\begin{cases} \text{Demand} & q_t = \alpha_1 + \alpha_2 p_t + \varepsilon_{1t} \\ \text{Supply} & q_t = \beta_1 + \beta_2 p_t + \varepsilon_{2t} \end{cases} \quad \alpha_2 \neq \beta_2$$

Observable data: $\{(p_t, q_t)\}_{t=1}^{T}$

When D=S,

$$p_t = \underbrace{\frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2}}_{\text{a constant}} + \underbrace{\frac{\varepsilon_{2t} - \varepsilon_{1t}}{\alpha_2 - \beta_2}}_{\text{just noise}}$$

$$q_t = \underbrace{\alpha_1 + \alpha_2 \frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2}}_{\text{a constant}} + \underbrace{\alpha_1 \left(\frac{\varepsilon_{2t} - \varepsilon_{1t}}{\alpha_2 - \beta_2} \right) + \varepsilon_{1t}}_{\text{just noise}}$$



Parameters under-identified! Can't estimate them

2SLS in over-identification case

1. Compute $\hat{X}$, the j^{th} column of which contains the fitted values of an OLS regression of the j^{th} column (endogenous regressor) of X on all instruments Z. $\quad \hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$
   i.e. regress reduced form of X
2. Perform OLS of y on $\hat{X}$. $\hat{\beta}_{IV} = (\hat{X}'\hat{X})^{-1}\hat{X}'y$

☆ Correct the SEs (Stata will use an inconsistent $s^2 = \frac{(y - X\hat{\beta}_{IV})'(y - X\hat{\beta}_{IV})}{n-k}$ instead of $s^2 = \frac{(y - X\hat{\beta}_{IV})'(y - X\hat{\beta}_{IV})}{n-k}$)

TEST RELEVANCE OF INSTRUMENTS

Obtain reduced form and do t-test (one restriction) or F-test (joint) on coefficient estimates of instruments

TEST FOR ENDOGENEITY

Given $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 x_1 + u_1$, $E(u_1) = E(x_1 u_1) = 0$

Question: is y_2 and u_1 correlated or not?

H_0: $Cov(y_2, u_1) = 0$ (y_2 exogenous)
- $\hat{\alpha}_{OLS}$ consistent, efficient (BLUE)
- $\hat{\alpha}_{IV}$ consistent too (but not needed)

H_1: $Cov(y_2, u_1) \neq 0$ (y_2 endogenous)
- only $\hat{\alpha}_{IV}$ consistent

HAUSMAN TEST

Under H_0, $\text{plim}(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) = 0$ while under H_1, $\text{plim}(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \neq 0$
Furthermore, $Var(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) = Var(\hat{\beta}_{IV}) - Var(\hat{\beta}_{OLS})$

positive definite as $\hat{\beta}_{OLS}$ is efficient (has minimum variance) under H_0

Thus, we should compare if the difference is close to 0
In the scalar setting, $SE(\hat{\beta}_{IV} - \hat{\beta}_{OLS}) = \sqrt{\widehat{Var}(\hat{\beta}_{IV}) - \widehat{Var}(\hat{\beta}_{OLS})} = \sqrt{SE(\hat{\beta}_{IV})^2 - SE(\hat{\beta}_{OLS})^2}$

Thus, t-stat is

scalar $t = \dfrac{\hat{\beta}_{IV} - \hat{\beta}_{OLS}}{SE(\hat{\beta}_{IV} - \hat{\beta}_{OLS})} \overset{a}{\sim} N(0,1)$. Reject if $|t| > z_{\frac{\alpha}{2}}$

vector $(\hat{\beta}_{IV} - \hat{\beta}_{OLS})' \left[\widehat{Var}(\hat{\beta}_{IV}) - \widehat{Var}(\hat{\beta}_{OLS})\right]^{-1} (\hat{\beta}_{IV} - \hat{\beta}_{OLS}) \overset{a}{\sim} \chi^2_{dim(\beta)}$ * χ^2_1 is quadratic form of $N(0,1)$

SIMPLE REGRESSION-BASED TEST

Obtain $\hat{y}_2$ from fitting $y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 x_1 + v_2$, where z_1, z_2 are instruments for y_2

- $\hat{y}_2$ is a "good" regressor where y_2's endogeneity is "washed out"
- $\hat{v}_2 = y_2 - \hat{y}_2$ must contain the "bad" component of y_2
  - $Cov(y_2, u_1) = Cov(\hat{y}_2 + \hat{v}_2, u_1) = Cov(\hat{y}_2, u_1) + Cov(\hat{v}_2, u_1)$
- Add $\hat{v}_2$ to original regression

$$y_1 = \alpha_0 + \alpha_1 y_2 + \alpha_2 x_1 + \delta \hat{v}_2 + \text{error}$$
Test H_0: $\delta = 0$
Test statistic: $\dfrac{\hat{\delta}}{SE(\hat{\delta})} \sim N(0,1)$

Reject H_0	Need to "control" for endogeneity of y_2 by including $\hat{v}_2$ Resulting parameter estimates $\hat{\alpha}$ will be same as when obtained through 2SLS
Don't reject H_0	Can just use original model w/o $\hat{v}_2$ to estimate α Then should just do OLS

Note: if ≥ 2 endogenous variables y_1, y_2, decompose both, put both residuals back in original regressor and do asymptotic F test on joint hypothesis ($H_0: \delta_1 = 0$ and $\delta_2 = 0$)

MLE

Assume random variables $y_1, \dots, y_n$ are distributed independently

$$L(\theta; y_1, \dots, y_n) = f(y_1, \dots, y_n; \theta) \overset{\text{independence}}{=} \prod_{i=1}^{n} f(y_i; \theta)$$

Take log-likelihood $\log L(\theta; y_1, \dots, y_n) = \sum_{i=1}^{n} \log L_i(\theta)$

$$\hat{\theta}_{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}} \log L(\theta; y_1, \dots, y_n) \equiv \underset{\theta \in \Theta}{\operatorname{argmax}} \log L(\theta; y)$$

FOC: $\left. \frac{\partial \log L(\theta; y)}{\partial \theta} \right|_{\hat{\theta}_{MLE}} = 0$

SOC (to verify maximum): $\left. \frac{\partial^2 \log L(\theta; y)}{\partial \theta \partial \theta'} \right|_{\hat{\theta}_{MLE}}$ is negative definite matrix

Usually hard to analytically determine (solve for) $\hat{\theta}_{MLE}$
Need to do gradient ascent to find local (global if globally concave) maximum

TERMS

Score vector/gradient is a $k \times 1$ vector of first derivatives $s(\theta) = \sum_{i=1}^{n} \frac{\partial \log L_i(\theta)}{\partial \theta} = \sum_{i=1}^{n} s_i(\theta)$
(evaluated at the true θ)

Hessian H is $k \times k$ matrix of second derivatives $H(\theta) = \sum_{i=1}^{n} \frac{\partial^2 \log L_i(\theta)}{\partial \theta \partial \theta'} = \sum_{i=1}^{n} H_i(\theta) = \begin{pmatrix} \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial \theta_1 \partial \theta_k} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 \log L(\theta)}{\partial \theta_k \partial \theta_1} & \cdots & \frac{\partial^2 \log L(\theta)}{\partial \theta_k \partial \theta_k} \end{pmatrix}$ is symmetric
↑ negative definite at $\hat{\theta}_{MLE}$

Information matrix $(k \times k)$ $I(\theta) = -E(H(\theta))$

PROPERTIES OF MLE ESTIMATOR

Assuming certain regularity conditions are met

Consistent but may be biased
Asymptotically Normal enables hypothesis testing
- $\hat{\theta} \overset{a}{\sim} N(\theta, I(\theta)^{-1})$, where $I(\theta) = -E(H(\theta))$

Asymptotically Efficient, so estimates are precise and tests using them are powerful
- $I(\theta)^{-1}$ provides a lower bound on the asymptotic covariance matrix for any consistent, asymptotically normal estimator for θ (Cramer-Rao lower bound)

☆ The asymptotic variance can be estimated by $\widehat{Avar}(\hat{\theta}) = -H(\hat{\theta})^{-1}$ or $\widehat{Avar}(\hat{\theta}) = \left(\sum_{i=1}^{n} s_i(\hat{\theta}) s_i(\hat{\theta})' \right)^{-1}$
($H(\hat{\theta})$ is symmetric)
↑ better properties in small samples
↑ outer product of first derivatives easier than computing H

MLE: BINARY CHOICE

$$f(y) = \begin{cases} \theta^y (1-\theta)^{1-y} & y = 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

$L(\theta; Y) = f(Y_1, \dots, Y_n; \theta) = \prod_{i=1}^{n} f(Y_i)$ assuming independence $= \theta^{\sum_{i=1}^{n} Y_i} (1-\theta)^{n - \sum_{i=1}^{n} Y_i}$

$\log L(\theta; Y) = \sum_{i=1}^{n} Y_i \log(\theta) + (n - \sum_{i=1}^{n} Y_i) \log(1-\theta)$

FOC: $\frac{\partial \log L}{\partial \theta} = \frac{\sum_{i=1}^{n} Y_i}{\theta} = \frac{n - \sum_{i=1}^{n} Y_i}{1-\theta} = 0 \Rightarrow \hat{\theta}_{MLE} = \frac{\sum_{i=1}^{n} Y_i}{n} = \bar{Y}$

Consistent: $\frac{1}{n} \sum_{i=1}^{n} Y_i \overset{LLN}{=} E(Y_i) = \theta$

Unbiased: $E(\hat{\theta}_{MLE}) = \frac{1}{n} \sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\theta = \theta$

Asymptotically normal $\hat{\theta}_{MLE} \overset{a}{\sim} N(\theta, \frac{\theta(1-\theta)}{n})$ so $SE(\hat{\theta}_{MLE}) = \sqrt{\frac{\hat{\theta}_{MLE}(1-\hat{\theta}_{MLE})}{n}}$

To show: obtain $I(\theta)$, or look at properties of $\bar{Y}$

MLE: LINEAR REGRESSION

$$y_i = x_i'\beta + u_i, \quad u_i | x_i \sim \text{iid } N(0,\sigma^2), \quad i = 1,\dots,n$$
$$y_i \sim N(x_i'\beta, \sigma^2)$$

assume X fixed (if not use conditional)

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i - x_i'\beta)^2}$$

$$L(\beta,\sigma^2) = f(y_1,\dots,y_n) \overset{\text{indep}}{=} \prod_{i=1}^{n} f(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y - x_i'\beta)^2}$$

$$\log L(\beta,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)$$

$$\text{FOCs}: \begin{cases} \frac{\partial \log L}{\partial \beta} = -\frac{1}{2\sigma^2}(-2X'y + 2X'X\beta) \\ \frac{\partial \log L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y - X\beta)'(y - X\beta) \end{cases}$$

$$\hat{\beta}_{MLE} = (X'X)^{-1}X'y$$
$$\hat{\sigma}^2_{MLE} = \frac{\hat{u}'\hat{u}}{n}, \quad \hat{u} = y - X\hat{\beta}_{MLE}$$

$$\begin{pmatrix} \hat{\beta}_{MLE} \\ \hat{\sigma}^2_{MLE} \end{pmatrix} \overset{a}{\sim} N\left(\begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}, \begin{pmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{pmatrix}\right)$$

Calculate Hessian (W10 slides) to get $I \Rightarrow I^{-1}$

Zero covariance + normality $\Rightarrow$ Independence

$$\hat{\beta}_{MLE} \overset{a}{\sim} N(\beta, \sigma^2(X'X)^{-1})$$

OLS derivation tells us this is in fact true for any n

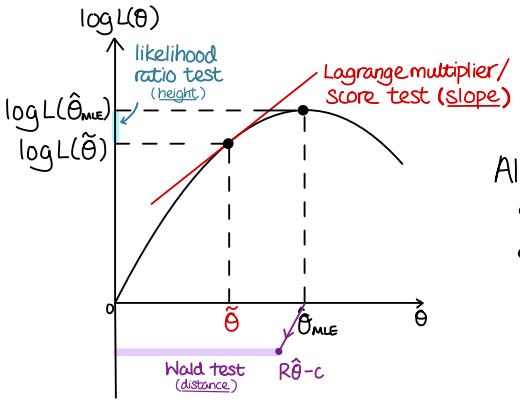
$$\hat{\sigma}^2_{MLE} \overset{a}{\sim} N\left(\sigma^2, \frac{2\sigma^4}{n}\right)$$

TRINITY OF CLASSICAL TESTING

$$H_0: R\theta = c \quad , \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_k \end{pmatrix}$$

$\hat{\theta}$ is the unrestricted (in this case MLE) estimator of θ

$\tilde{\theta}$ is the restricted (in this case MLE) estimator of θ, where $R\tilde{\theta} = c$



All three tests are asymptotically equivalent
- Often chosen based on computational ease
- May have different finite (small) sample properties

WALD TEST

Is the discrepancy $d = R\hat{\beta} - c$ close to 0?

- Uses only $\hat{\theta}$, not $\tilde{\theta}$
- Compared to Wald test in lin. reg. case:
  - Need to assume MLE regularity conditions
  - Need to use $\hat{\beta}_{MLE} \overset{a}{\sim} N(\beta, V)$, and V obtained differently, using $(I(\beta))^{-1}$
  - Test statistic valid only asymptotically, as we don't have AI-5

TEST STATISTIC

Single linear restriction: $\boxed{Z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \xrightarrow{d} N(0,1)}$ under H_0

Called z-test (asymptotic) rather than t-test (finite) since $\hat{\theta}_{MLE}$ are usually asymptotically normal

Joint linear restriction: $\boxed{W = d'(\widehat{Var}(d))^{-1} d \xrightarrow{d} \chi^2_r}$ under H_0 (r: no. of restrictions)

$Var(d)$ or $\widehat{var}(d)$ don't matter. Both have same limiting distribution

- Since d is a vector, we want to check length of vector $d'd$
- We don't know how $d'd$ is distributed, but we know how the Mahalanobis distance $d'(Var(d))'d$ is distributed

LIKELIHOOD RATIO TEST

Does imposing restriction on estimator lead to a significant loss of fit (similar to RRSS/URSS)

- Is $\log L(\hat{\theta}) - \log(\tilde{\theta})$ significantly different from 0
- Easy to apply, but need both $\hat{\theta}, \tilde{\theta}$

TEST STATISTIC

$$\boxed{2(\log L(\hat{\theta}) - \log L(\tilde{\theta})) \xrightarrow{d} \chi^2_r} \text{ under } H_0$$

(unrestricted: $\hat{\theta}$; restricted: $\tilde{\theta}$; r: no. of restrictions)

If H_0 is true, likelihood ratio $\frac{\log L(\tilde{\theta})}{\log L(\hat{\theta})} \approx 1 \Leftrightarrow \log L(\tilde{\theta}) - \log L(\hat{\theta}) \approx 0$

LAGRANGE MULTIPLIER TEST

imposes restriction!

$$\max \log L(\theta) \text{ s.t. } R\theta = c$$

$$\mathcal{L} = \log L(\theta) + \lambda'(R\theta - c)$$

- λ are shadow prices: how expensive it is to impose restrictions
- Reject restrictions (H_0) if shadow prices are too high (far from 0)
- Uses only $\tilde{\theta}$, not $\hat{\theta}$

TEST STATISTIC

$$\tilde{\lambda}'(\widehat{Var}(\tilde{\lambda}))^{-1}\tilde{\lambda} \xrightarrow{d} \chi^2_r \quad \text{under } H_0$$

or equivalently,

$$s(\tilde{\theta})'\left(\widehat{Var}(s(\tilde{\theta}))\right)^{-1}s(\tilde{\theta}) \xrightarrow{d} \chi^2_r \quad \text{under } H_0$$

$$s(\tilde{\theta}) = \frac{\partial \log L(\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\partial \log L_i(\theta)}{\partial \theta}\bigg|_{\tilde{\theta}} = \sum_{i=1}^{n} s_i(\tilde{\theta})$$

the latter can be rewritten as $\left[\sum_{i=1}^{n} s_i(\tilde{\theta})\right]' \left[\sum_{i=1}^{n} s_i(\tilde{\theta}) s_i(\tilde{\theta})'\right]^{-1} \left[\sum_{i=1}^{n} s_i(\tilde{\theta})\right]$ because

$$Var\left(\frac{\partial \log L(\theta)}{\partial \theta}\right) = E\left\{\left[\frac{\partial \log L}{\partial \theta} - E\left(\frac{\partial \log L}{\partial \theta}\right)\right]\left[\frac{\partial \log L}{\partial \theta} - E\left(\frac{\partial \log L}{\partial \theta}\right)\right]'\right\} = E\left[\left(\frac{\partial \log L}{\partial \theta}\right)\left(\frac{\partial \log L}{\partial \theta}\right)'\right]$$

($= 0$ at $\tilde{\theta}_{MLE}$; $= 0$)

which can be estimated by $\sum_{i=1}^{n} s_i(\tilde{\theta}) s_i(\tilde{\theta})'$

notation is different from MLE chapter

This can be obtained by computing the nR^2 of regressing $1_i = s_i(\tilde{\theta})'\gamma + v_i$, $i = 1, \dots, n$ where $s_i(\tilde{\theta}) = \begin{pmatrix} \frac{\partial \log L_i}{\partial \theta_1}\big|_{\tilde{\theta}} \\ \vdots \\ \frac{\partial \log L_i}{\partial \theta_k}\big|_{\tilde{\theta}} \end{pmatrix}$

or equivalently $\underset{n\times 1}{\iota} = \underset{n\times k}{S}\underset{k\times 1}{\gamma} + v$ in matrix form, where $S = \begin{pmatrix} \vdash s_1'(\tilde{\theta}) \dashv \\ \vdots \\ \vdash s_n'(\tilde{\theta}) \dashv \end{pmatrix}$. This gives $\hat{\gamma} = (S'S)^{-1}S'\iota$.

(vector of ones)

It works because $nR^2 = n\frac{ESS}{TSS} = n\frac{\sum_{i=1}^{n}\hat{y}_i^2}{\sum_{i=1}^{n} y_i^2} = n\frac{\hat{y}'\hat{y}}{n} = (S\hat{\gamma})'(S\hat{\gamma}) = \hat{\gamma}'S'S\hat{\gamma} = \iota'S(S'S)^{-1}S'\iota$

instead of the usual $\sum_{i=1}^{n}(y_i - \bar{y})^2$ which equals 0 i.e. R^2 is uncentred

is equivalent to $\left[\sum_{i=1}^{n} s_i(\tilde{\theta})\right]' \left[\sum_{i=1}^{n} s_i(\tilde{\theta}) s_i(\tilde{\theta})'\right]^{-1} \left[\sum_{i=1}^{n} s_i(\tilde{\theta})\right]$

BINARY CHOICE: LPM/NLLS

LINEAR PROBABILITY MODEL $y_i = x_i'\beta + \varepsilon_i$, $E(\varepsilon_i|x_i) = 0$ where $y_i \in \{0,1\}$

Benefits:

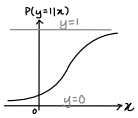
- Since $E(\varepsilon_i|x_i) = 0$ and y_i binary, $E(y_i|x_i) = 1 \times Pr(y_i=1|x_i) + 0 \times Pr(y_i=0|x_i) = Pr(y_i=1|x_i) = x_i'\beta$
  - $Pr(y_i=1|x_i)$ linear in the parameters; easy to interpret β_j
  - β_j: How does the j^{th} explanatory variable affect probability that $y=1$, cp
  - Constant marginal effects

Drawback:

- $x_i\hat{\beta}$ is not necessarily constrained to lie in $[0,1]$
- Error term $\varepsilon_i = y_i - x_i\beta$ can only take 2 values $1 - x_i'\beta$ or $-x_i'\beta$; highly non-normal
- Heteroscedasticity. $Var(\varepsilon_i|x_i) = Var(y_i|x_i) = Pr(y_i=1|x_i)(1 - Pr(y_i=1|x_i)) = x_i'\beta(1-x_i'\beta)$ is not constant
  - Must use RSE

$Var(y) = E(y^2) - [E(y)]^2$
$E(y) = E(y^2) = Pr(y=1)$

Thus, we may decide to model $P(y=1|x) = F(x'\beta)$, where $F(\cdot)$ is the CDF



Usually $F(x_i'\beta) = \Phi(x_i'\beta) = \int_{-\infty}^{x_i'\beta} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du$ probit or $\Lambda(x_i'\beta) = \frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$ logit

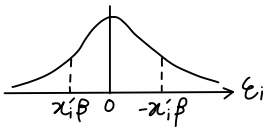
CDF of N(0,1)

CDF of logistic RV

WHY USE CDF? Intuitive explanation using behavioural model

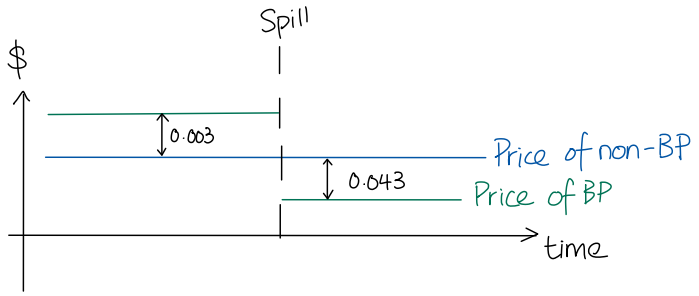
Let w_i be the excess utility of working (vs not working) s.t. $w_i = x_i'\beta + \varepsilon_i$.
We don't observe w_i but $y_i = 1$ if $w_i \geq 0$, $y_i = 0$ if $w_i < 0$ therefore

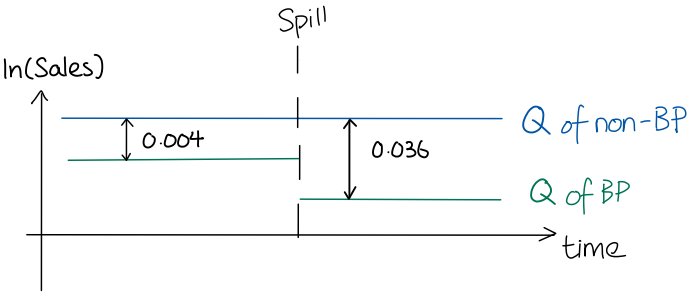
$$\begin{aligned} Pr(y_i=1|x_i) &= Pr(w_i > 0|x_i) = Pr(x_i'\beta + \varepsilon_i > 0|x_i) \\ &= Pr(\varepsilon_i > -x_i'\beta|x_i) \\ &= Pr(\varepsilon_i < x_i'\beta|x_i) \text{ given } \varepsilon_i\text{'s distribution (if symmetric)} \\ &= F_\varepsilon(x_i'\beta) \text{ depends on pdf of } \varepsilon \end{aligned}$$



NON-LINEAR REGRESSION

- We can run $y_i = F(x_i'\beta) + \varepsilon_i$ to get a non-linear least squares estimator $\hat{\beta}_{NLLS} = \underset{b}{\text{argmin}} \sum_{i=1}^{n} (y - F(x_i'b))^2$ (usually no explicit form)
- But doesn't solve heteroscedasticity (use RSEs). So next use MLE: asymptotically efficient





$\Delta P = P_{after} - P_{before} = -0.046$

$\Delta Q = e^{\ln Q_{after}} - e^{\ln Q_{before}} =$

$d \ln Q = \frac{1}{Q}$

$\frac{\Delta \ln Q}{\Delta \ln P} = \beta$

BINARY CHOICE: MLE

y_i: Binary / Bernoulli RV $\quad f(y_i) = \begin{cases} \theta_i^{y_i}(1-\theta_i)^{1-y_i} \\ 0 \end{cases}$

$\theta_i = Pr(y_i=1|x_i)$ should differ from individual to individual

We can specify (model) a relation between θ_i and x_i using a CDF

$\theta_i = F(x_i'\beta) = \Phi(x_i'\beta)$ probit or $\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}$ logit

PROBIT

$$L(\beta) = \prod_{i=1}^{n} \Phi(x_i'\beta)^{y_i}(1-\Phi(x_i'\beta))^{1-y_i}$$

$$\log L(\beta) = \sum_{i=1}^{n} y_i \underbrace{\log \Phi(x_i'\beta)}_{Pr(y_i=1|x_i)} + (1-y_i)\underbrace{\log[1-\Phi(x_i'\beta)]}_{Pr(y_i=0|x_i)}$$

intuitively choosing $\hat{\beta}_{MLE}$ s.t. for individuals with $y_i=1$, $\widehat{Pr(y_i=1|x_i)}$ is high; individuals with $y_i=0$, $\widehat{Pr(y_i=0|x_i)}$ is high

FOC: $\left.\frac{\partial \ln L(\beta)}{\partial \beta}\right|_{\hat{\beta}_{MLE}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^{g} x_i = 0$ PS9Q3

where generalised residuals $\hat{\varepsilon}_i^{g} = \frac{y_i - \Phi(x_i'\hat{\beta}_{ML})}{\Phi(x_i'\hat{\beta}_{ML})[1-\Phi(x_i'\hat{\beta}_{ML})]} \overset{PDF}{\downarrow} \phi(x_i'\hat{\beta}_{MLE})$

However, FOC gives no explicit form for $\hat{\beta}_{MLE}$ (may need gradient ascent)

↑ "the parameter iteratively converging... globally concave obj fn..."

INTERPRET

Predicted probability: $\widehat{Pr(y_i=1|x_i)} = \Phi(x_i'\hat{\beta}_{MLE})$

Intuitively, $\hat{\beta}_{MLE}$ is chosen s.t. individuals with $y_i=1$ have high $\widehat{Pr(y_i=1|x_i)} = \Phi(x_i'\hat{\beta}_{MLE})$; individuals with $y_i=0$ have high $\widehat{Pr(y_i=0|x_i)} = 1-\Phi(x_i'\hat{\beta}_{MLE})$

Marginal effect (if x_j is continuous): $\frac{\partial \Phi(x'\beta)}{x_j} = \underbrace{\phi(x'\beta)}_{\geq 0}\beta_j$ not β_j (unlike OLS)

- Need "ceteris paribus"
- Not constant. Depends on individual characteristics $x = \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix}$, though sign of β_j still informative
- May want to evaluate at the means $\phi(\bar{x}'\beta)\beta_m$. (partial effect on average)
  - That said, individual $\bar{x}$ may not exist, esp for dummy variable ("half female")
  - Stata command: margins, dydx (*) atmeans
- May want to average over all individuals $\frac{1}{n}\sum_{i=1}^{n}\phi(x_i'\beta)\beta_j$ (partial average treatment effect)
  - "On average, increasing x_j by 1 increases probability of $y_i=1$ by..., ceteris paribus"
  - Stata command: margins, dydx (*)

Marginal effect (if x_j is a dummy): $\frac{\Delta Pr(y=1|x)}{\Delta x_1} = \Phi(\beta_0+\beta_1\times 1+\beta_2 x_2+\ldots) - \Phi(\beta_0+\beta_1\times 0+\beta_2 x_2+\ldots)$

$= Pr(y_i=1|x_1=1, x_2, \ldots) - Pr(y_i=1|x_1=0, x_2, \ldots)$

- Need "ceteris paribus"

$\alpha + \varepsilon$

BINARY CHOICE: MLE TEST

Same as generic cases but with additional pointers for MLE binary choice

example $H_0: \beta_j = 0$

WALD TEST

$z = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)} \overset{a}{\sim} N(0,1)$ under H_0, reject if $|z| > z_{\frac{\alpha}{2}}$

- $SE(\hat{\beta}_j)$ calculated by taking square root of the j,j^{th} element of $I(\hat{\beta})^{-1} = \left(-\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta'}\Big|_{\hat{\beta}_{MLE}}\right)^{-1}$ (might be $j+1, j+1$ if intercept β_0)
- Recall: Under suitable regularity conditions, $\hat{\beta}_{MLE} \overset{a}{\sim} N(\beta, I(\beta)^{-1})$ so $\hat{\beta}_j \overset{a}{\sim} N(\beta_j, [I(\hat{\beta})^{-1}]_{jj})$
- Not using t-dist to account for imprecision as this test is only valid asymptotically and asymptotically it doesn't matter if we know the true SE or use a consistent estimator for it

LIKELIHOOD RATIO TEST

$2(\log L^U - \log L^R) \xrightarrow{d} \chi^2_r$ under H_0, reject if $> \chi^2_{r,\alpha}$

(L^U: unconstrained probit/logit; r: no. of restrictions; "difference in dimensionality" (df of unrestricted model − df of restricted))

Stata commands (for testing $\beta_j = 1$):
constr 1 variable_j=1 set 1 restriction
profit depvar variable_1 variable_2 variable_j ..., constr(1) estimate probit w constraint

LAGRANGE MULTIPLIER TEST

$nR^2 \sim \chi^2_r$ under H_0 reject if $> \chi^2_{r,\alpha}$

Uncentred R^2 obtained from regressing $1_i = (\hat{\varepsilon}_i^c)\gamma_1 + (\hat{\varepsilon}_i^c x_{1i})\gamma_2 + \ldots + v_i$ (score vector in probit)

$nR^2 = n\frac{ESS}{TSS} = n\frac{\sum_{i=1}^n \hat{y}_i^2}{\sum_{i=1}^n y_i^2} = n\frac{\hat{y}'\hat{y}}{n} = (S\hat{\gamma})'(S\hat{\gamma}) = \hat{\gamma}'S'S\hat{\gamma} = i'S(S'S)^{-1}S'i$

In Stata probit output, joint significance of the slopes is automatically given as "LR chi2(k)" where k is the number of regressions in the top right hand corner
Test statistic will be small/derivative will be close to zero if the restrictions imposed were true. It is an indication that imposing the restrictions on the model is not unreasonable or costly.

CHOW TEST

"Are slopes/intercepts the same for a different dataset/segment of the population?" (rural/urban, male/female)

Conduct likelihood ratio test. $H_0: \beta_j^{rural} = \beta_j^{urban}\ \forall j$ vs H_1: at least one $\beta_j^{rural} \neq \beta_j^{urban}$

Restricted model:
probit y x
Then take the log likelihood directly from the Stata output to get $\log L^R$

Unrestricted model:
probit y x if urban == 1
probit y x if urban == 0
Then take the log likelihood from both outputs and add them up to get $\log L^U$

$2(\log L^U - \log L^R) \xrightarrow{d} \chi^2_r$

(r: number of variables (don't forget intercept); "difference in dimensionality" (df of unrestricted model − df of restricted = 2k − k = k))

COUNT VARIABLE

$y_i = 0, 1, 2, \ldots$ (nonnegative integer values)

$$f(y_i) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \quad y_i = 0, 1, 2, 3, \ldots$$

where $\lambda = E(y_i) = Var(y_i)$

↑ λ not constant · dep on x_i

- OLS has similar drawbacks as before (negative $E(y_i|x_i)$, heteroscedasticity)
- Weighted NLLS more efficient than NLLS due to heteroscedasticity; MLE even more efficient

We can specify (model) a relation between λ_i and x_i using an exponential (to ensure $\lambda_i \geq 0$)

$$\lambda_i = E(y_i|x_i) = e^{x_i'\beta}$$

MLE ESTIMATION

$$\ln L(\beta) \overset{indep}{=} \sum_{i=1}^{n} \ln\left(\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i}\right) = \sum_{i=1}^{n}\left[-\lambda_i + y_i \ln\lambda_i - \ln(y_i!)\right] = \sum_{i=1}^{n}\left[-x_i'\beta + y_i \ln x_i'\beta - \ln(y_i!)\right]$$

FOC: $\left.\frac{\partial \ln L(\beta)}{\partial \beta}\right|_{\hat{\beta}_{MLE}} = \sum_{i=1}^{n} \hat{\varepsilon}_i^R x_i = 0$ gives no explicit form for $\hat{\beta}_{MLE}$ (may need gradient ascent)

where residuals $\hat{\varepsilon}_i^R = y_i - e^{-x_i'\hat{\beta}_{MLE}}$

- $SE(\hat{\beta}_j)$ calculated by taking square root of the j,j^{th} / element of $I(\hat{\beta})^{-1} = \left(-\left.\frac{\partial^2 \ln L(\beta)}{\partial\beta\partial\beta}\right|_{\hat{\beta}_{MLE}}\right)^{-1}$ (might be $j+1, j+1$ if intercept β_0)
- Recall: Under suitable regularity conditions, $\hat{\beta}_{MLE} \overset{a}{\sim} N(\beta, I(\beta)^{-1})$ so $\hat{\beta}_j \overset{a}{\sim} N(\beta_j, [I(\hat{\beta})^{-1}]_{jj})$

Marginal effect: $\frac{\partial E(y|x)}{\partial x_k} = \frac{\partial e^{-x_i'\beta}}{\partial x_k} = \beta_k e^{x_i'\beta}$

" How does expected y_i change due to x_k, cp "