1. LINEAR REGRESSION UNDER IDEAL CONDITIONS

[1] What is "Regression Model"?

Example:

- Suppose you are interested in the average relationship between income (y) and education (x).
- For the people with 12 years of schooling (x =12), what is the average income (E(y|x=12))?
- For the people with x years of schooling, what is the average income (E(y|x))?
- Regression model:

$$y = E(y \mid x) + \varepsilon ,$$

  where ε is a disturbance (error) term with $E(\varepsilon \mid x) = 0$.

- Regression analysis is aimed to estimate $E(y \mid x)$.

Digression to Probability Theory

(1) Bivariate Distributions

- Consider two random variables (RV), X and Y with a joint probability density function (pdf): $f(x, y) = Pr(X=x, Y=y)$.

- **Marginal (unconditional) pdf**:

  $f_x(x) = \Sigma_y f(x,y) = Pr(X = x)$ regardless of Y;

  $f_y(y) = \Sigma_x f(x,y) = Pr(Y = y)$ regardless of X.

- **Conditional pdf**:

  $f(y|x) = Pr(Y = y, \text{given } X = x) = f(x,y)/f_x(x)$.

- **Stochastic independence**:
  - X and Y are stochastically independent iff $f(x,y) = f_x(x)f_y(y)$, for all x,y.
  - Under this condition, $f(y|x) = f(x,y)/f_x(x) = [f_x(x)f_y(y)]/f_x(x) = f_y(y)$.

EX:

- Toss two coins, A and B.
- X = 1 if head from A; = 0 if tail from A.

  Y = 1 if head from B; = 0 if tail from B.

  $f(x,y) = 1/4$ for any $x,y = 0, 1$. (4 possible cases)

- Marginal pdf of x:

  $f_x(0) = Pr(X=0)$ regardless of $y = f(0,1) + f(0,0) = 1/4 + 1/4 = 1/2$.

  $f_x(1) = Pr(X=1)$ regardless of $y = f(1,1) + f(1,0) = 1/4 + 1/4 = 1/2$.

  $f_x(x) = 1/2$, for $x = 0, 1$.

  Similarly, $f_y(y) = 1/2$, for $y = 0, 1$.

- Conditional pdf:

  $f(y = 1| x = 1) = f(1,1)/f_x(1) = (1/4)/(1/2) = 1/2$;

  $f(y = 0| x = 1) = f(0,1)/f_x(1) = 1/2$.

  $\rightarrow$ $f(y| x=1) = 1/2$, for $y = 0, 1$.

- Find $f(y|x=0)$ by yourself.

- Stochastic independence:

  $f_x(x) = f_y(y) = 1/2$; $f_x(x)f_y(y) = 1/4 = f(x,y)$, for any x and y.

  Thus, x and y are stochastically independent.

Expectation:

$$E[g(x,y)] = \Sigma_x\Sigma_y g(x,y)f(x,y) \ [\text{or} \int_\Omega g(x,y)\, f(x,y)dxdy\,].$$

Means:

$$\mu_x = E(x) = \Sigma_x\Sigma_y x f(x,y) = \Sigma_x x f_x(x).$$

$$\mu_y = E(y) = \Sigma_x\Sigma_y y f(x,y) = \Sigma_y y f_y(y).$$

Variances:

$$\begin{aligned}\sigma_x^2 &= E[(x-\mu_x)^2] = \Sigma_x\Sigma_y (x-\mu_x)^2 f(x,y) = \Sigma_x (x-\mu_x)^2 f_x(x) \\ &= E(x^2) - [E(x)]^2 = \Sigma_x x^2 f_x(x) - \mu_x^2\end{aligned}.$$

$$\begin{aligned}\sigma_y^2 &= \Sigma_x\Sigma_y (y-\mu_y)^2 f(x,y) = \Sigma_y (y-\mu_y)^2 f_y(y) \\ &= E(y^2) - [E(y)]^2 = \Sigma_y y^2 f_y(y) - \mu_y^2\end{aligned}$$

Covariance:

$$\begin{aligned}\sigma_{xy} = \text{cov}(x,y) = E[(x-\mu_x)(y-\mu_y)] &= \Sigma_x\Sigma_y (x-\mu_x)(y-\mu_y) f(x,y) \\ = E(xy) - \mu_x\mu_y &= \Sigma_x\Sigma_y xyf(x,y) - \mu_x\mu_y\end{aligned}$$

Note: $\sigma_{xy} > 0 \rightarrow$ positively linearly related;

$\sigma_{xy} < 0 \rightarrow$ negatively linearly related;

$\sigma_{xy} = 0 \rightarrow$ no linear relation.

EX: $x, y = 1, 0$, with $f(x,y) = 1/4$.

$E(xy) = \Sigma_x\Sigma_y xyf(x,y)$

$= 0\times0\times(1/4) + 0\times1\times(1/4)+ 1\times0\times(1/4) + 1\times1\times(1/4) = 1/4.$

Correlation Coefficient:

The correlation coefficient between x and y is defined by:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}.$$

Theorem:

$-1 \le \rho_{xy} \le 1$.

Note: $\rho_{xy} \to 1$: highly positively linearly related;

$\rho_{xy} \to -1$; highly negatively linearly related;

$\rho_{xy} \to 0$: no linear relation.

Theorem:

If X and Y are stochastically independent, then, $\sigma_{xy} = 0$. But, not vice versa.

Conditioning in a Bivariate Distribution:

- X,Y: RVs with f(x,y). (e.g., Y = income, X = education)
- Population of billions and billions: $\{(x^{(1)},y^{(1)}), ... (x^{(b)},y^{(b)})\}$.
- Average of $y^{(j)}$ = E(y).
- For the people earning a specific education level x, what is the average of y?

Conditional Mean and Variance:

- $E(y \mid x) = E(y \mid X = x) = \Sigma_y y f(y \mid x)$.
- $\text{var}(y \mid x) = E[(y - E(y \mid x))^2 \mid x] = \Sigma_y \left(y - E(y \mid x)\right)^2 f(y \mid x)$.

Regression model:

- Let ε = y - E(y|x) (deviation from conditional mean).
- y = E(y|x) + y - E(y| x) = E(y|x) + ε (regression model).
- E(y|x) = explained part of y by x.

  ε = unexplained part of y (called disturbance term).

  E(ε|x) = 0 and var(ε| x) = var(y|x).

Note:

- E(y|x) may vary with x, i.e., E(y|x) is a function of x.
- Thus, we can define $E_x[E(y|x)]$, where $E_x(\bullet)$ is the expectation over x = $\Sigma_x \bullet f_x(x)$ or $\int_\Omega \bullet f_x(x)dx$.

Theorem: (Law of Iterative Expectations)

$E(y)$ [unconditional mean] $= E_x[E(y|x)]$.

Proof:

$E(y) = \Sigma_x\Sigma_y yf(x,y) = \Sigma_x\Sigma_y yf(y|x)f_x(x) = \Sigma_x[\Sigma_y yf(y|x)]f_x(x)$.

Note:

For discrete RV, X with $x = x_1, ...,$

$E(y) = \Sigma_x E(y|x)f_x(x) = E(y|x=x_1)f_x(x_1) + E(y|x=x_2)f_x(x_2) + ...$.

Implication:

If you know the conditional mean of y and the marginal distribution of x, you can also find the unconditional mean of y, too.

EX 1: Suppose $E(y|x) = 0$, for all x. $E(y) = E_x[E(y|x)] = E_x(0) = 0$.

EX 2: $E(y|x) = \beta_1 + \beta_2 x$. $\rightarrow$ $E(y) = E_x(E(y|x)) = E_x(\beta_1 + \beta_2 x) = \beta_1 + \beta_2 E(x)$.

Question: When can $E(y|x)$ be linear? Answered later.

Definition:

We say that y is homoskedastic if $var(y|x)$ is constant.

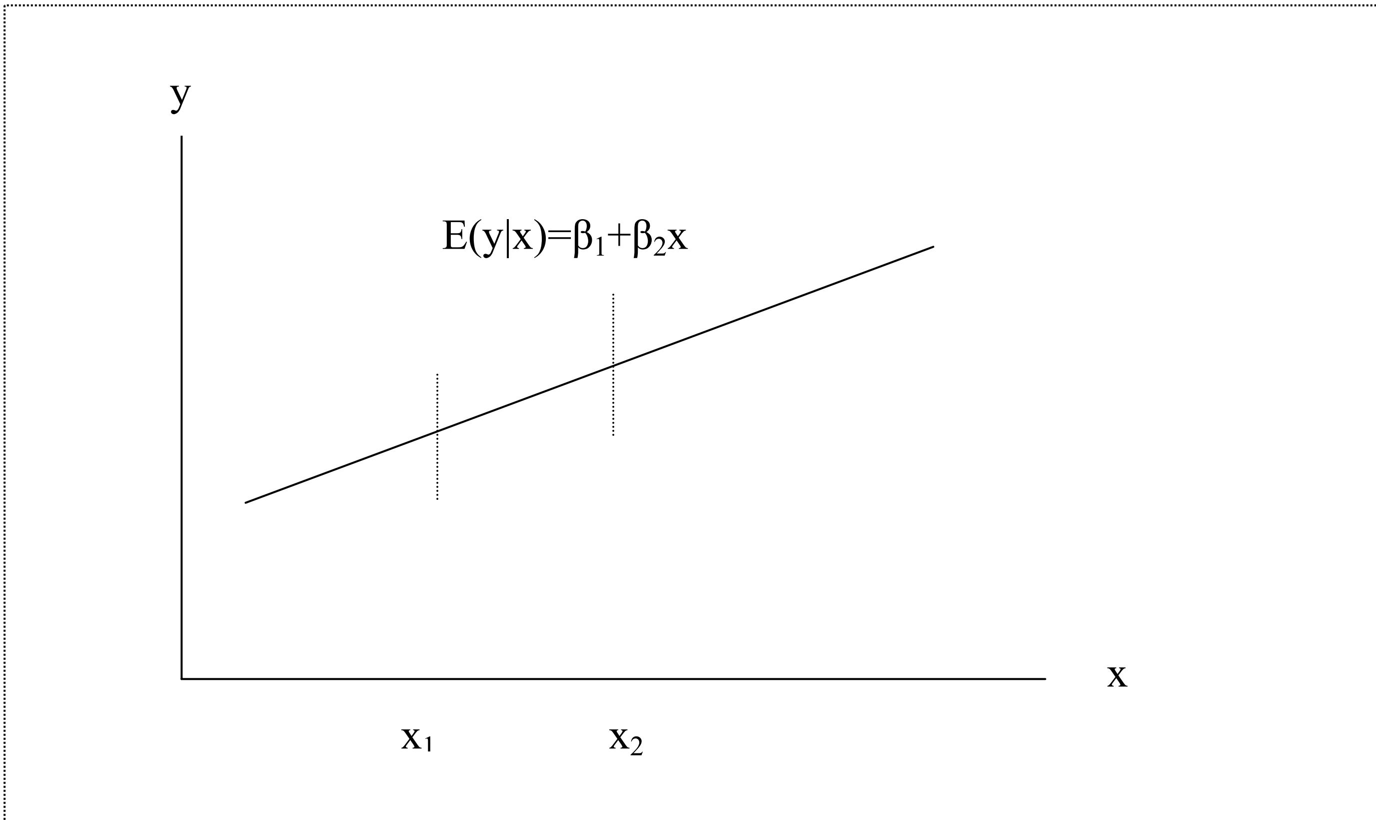
EX: $y = E(y|x) + \varepsilon$ with $var(\varepsilon|x) = \sigma^2$ for all x(constant).

$\rightarrow var(y|x) = var[E(y|x)+\varepsilon|x] = var(\varepsilon|x) = \sigma^2$, for all x.

$\rightarrow$ y is homoskedastic.

Graphical Interpretation of Conditional Means and Variances

- Consider the following population:



- $E(y|x=x_1)$ measures the average value of y for the group of $x = x_1$.
- $var(y|x=x_1)$ measures the dispersion of y given $x = x_1$.
- If $var(y|x=x_1) = var(y|x=x_2) = ...$, we say that y is homoskedastic.
- Law of iterative expectation:

  $E(y) = \Sigma_x E(y|x)f_x(x) = E(y|x=x_1)Pr(x=x_1) + E(y|x=x_2)Pr(x=x_2) + ...$

Question: It is worth finding $E(y|x)$?

Theorem: (Decomposition of Variance)

$var(y) = var_x[E(y|x)] + E_x[var(y|x)]$.

Note:

- $var_x[E(y|x)] \leq var(y)$, since $E_x[var(y|x)] \geq 0$.
- $var(y) = E[(y-E(y))^2]$
  $=$ total variation of y.
  $var_x[E(y|x)] = E_x[(E(y|x)-E(y))^2]$
  $=$ a part of variation in y due to variation in $E(y|x)$
  $=$ variation in y explained by $E(y|x)$.

Coefficient of Determination:

$\boldsymbol{R}^2 = var_x[E(y|x)]/var(y)$.

$\rightarrow$ Measure of worthiness of knowing $E(y|x)$.

$\rightarrow$ $0 \leq \boldsymbol{R}^2 \leq 1$.

Note:

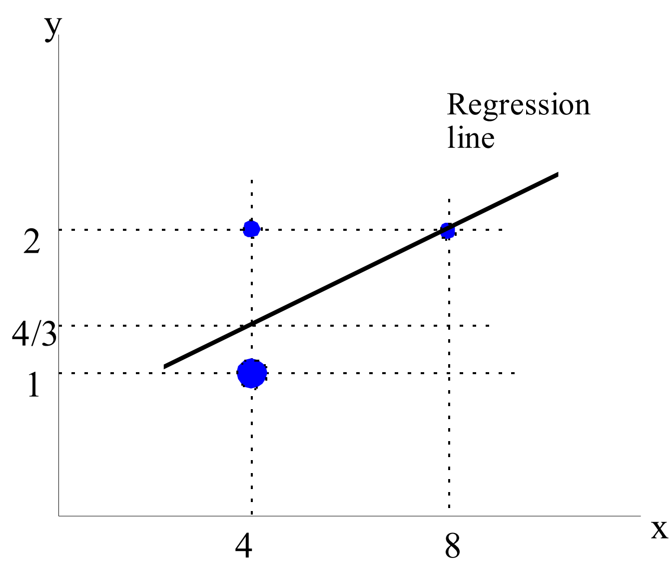
- $\boldsymbol{R}^2$ = variation in y explained by $E(y|x)$/total variation of y.
- Wish $\boldsymbol{R}^2$ close to 1.

Summarizing Exercise:

- A population with X (income=\$10,000) and Y (consumption=\$10,000).
- Joint pdf:

Y\X	4	8
1	1/2	0
2	1/4	1/4

- Graph for this population:



- Marginal pdf:

Y\X	4	8	$f_y(y)$
1	1/2	0	1/2
2	1/4	1/4	1/2
$f_x(x)$	3/4	1/4	

- Means of X and Y:
    - $E(x) = \mu_x = \Sigma_x x f_x(x) = 4\times f_x(4) + 8\times f_x(8) = 4\times(3/4) + 8\times(1/4) = 5.$
    - $E(y) = \mu_y = \Sigma_y y f_y(y) = 1.5.$
- Variances of X and Y:
    - $\text{var}(x) = \sigma_x^2 = \Sigma_x (x-\mu_x)^2 f_x(x)$

      $= (4-5)^2 f_x(4) + (8-5)^2 f_x(8) = 1\times(3/4) + 9\times(1/4) = 3.$
    - $\text{var}(y) = \sigma_y^2 = 1/4.$
- Covariance between X and Y:
    - $\sigma_{xy} = E[(x-\mu_x)(y-\mu_y)] = E(xy) - \mu_x\mu_y = \Sigma_x\Sigma_y xyf(x,y) - \mu_x\mu_y$

      $= 4\times1\times f(4,1)+4\times2\times f(4,2)+8\times1\times f(8,1)+8\times2\times f(8,2)-5\times1.5 = 0.5.$
    - $\rho_{xy} = \dfrac{\sigma_{xy}}{\sigma_x\sigma_y} = \dfrac{0.5}{\sqrt{3}\sqrt{1/4}} \cong 0.58.$

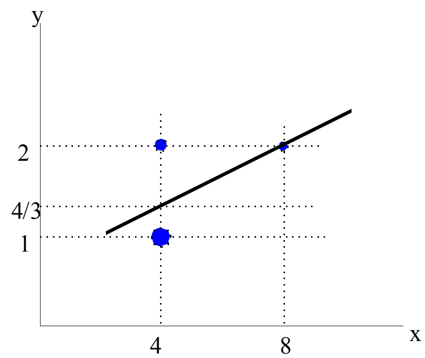
- Conditional Probabilities

Y\X	4	8	$f_y(y)$
1	1/2	0	1/2
2	1/4	1/4	1/2
$f_x(x)$	3/4	1/4	

- f(y|x):

Y\X	4	8
1	2/3	0
2	1/3	1

- Conditional mean:
  - $E(y|x=4) = \Sigma_y y f(y|x=4) \quad = 1\times f(y=1|x=4) + 2\times f(y=2|x=4)$
    $= 1\times(2/3) + 2\times(1/3) = 4/3.$
  - $E(y|x=8) = 2.$



- Conditional variance of Y:
  - $var(y|x=4) = \Sigma_y[y-E(y|x=4)]^2 f(y|x=4) = 6/27.$
  - $var(y|x=8) = 0.$

- Law of iterative expectation:
  - $E_x[E(y|x)] = \Sigma_x E(y|x)f_x(x) = E(y|x=4)f_x(4) + E(y|x=8)f_x(8)$
    $= (4/3)\times(3/4) + 2\times(1/4) = 1.5 = E(y)$!!!

(2) Bivariate Normal Distribution

Definition:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \rho_{xy}\sigma_x\sigma_y \\ \rho_{xy}\sigma_x\sigma_y & \sigma_y^2 \end{pmatrix} \right).$$

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho_{xy}^2}}$$

$$\times \exp\left(-\frac{1}{2(1-\rho_{xy}^2)} \left\{ \frac{(x-\mu_x)^2}{\sigma_x^{\ 2}} - 2\rho_{xy}\frac{x-\mu_x}{\sigma_x}\frac{y-\mu_y}{\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^{\ 2}} \right\} \right),$$

where x, y $\in \Re$.

Facts:

- $f_x(x) \sim N(\mu_x, \sigma_x^2)$ and $f_y(y) \sim N(\mu_y, \sigma_y^2)$.
- E(y|x) = $\beta_1 + \beta_2$x and var(y|x) is constant (see Greene).

  $\rightarrow$ E(y|x) is linear in x and y is homoskedastic.
- If $\sigma_{xy} = 0$ (or $\rho_{xy} = 0$), x and y are stochastically independent.

(3) Multivariate Distributions

Definition: (Mean vector and covariance matrix)

$X_1, \ldots, X_n$: random variables.

Let $x = [x_1, \ldots, x_n]'$ ($n\times 1$ vector). Then,

$$E(x) = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_n) \end{pmatrix}; \quad Cov(x) = \begin{bmatrix} \mathrm{var}(x_1) & \mathrm{cov}(x_1,x_2) & \ldots & \mathrm{cov}(x_1,x_n) \\ \mathrm{cov}(x_2,x_1) & \mathrm{var}(x_2) & \ldots & \mathrm{cov}(x_2,x_n) \\ \vdots & \vdots & & \vdots \\ \mathrm{cov}(x_n,x_1) & \mathrm{cov}(x_n,x_1) & \ldots & \mathrm{var}(x_n) \end{bmatrix}.$$

$\rightarrow$ Cov(x) is symmetric.

EX: If x is scalar, $Cov(x) = E[(x-\mu)^2] = var(x)$.

EX: $x = [x_1,x_2]'$; $E(x) = \mu = [\mu_1, \mu_2]'$

$x - \mu = [x_1-\mu_1, x_2-\mu_2]'$

$$\rightarrow \quad (x-\mu)(x-\mu)' = \begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \begin{pmatrix} x_1-\mu_1 & x_2-\mu_2 \end{pmatrix}$$

$$= \begin{pmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 \end{pmatrix}.$$

$\rightarrow$ $E[(x-\mu)(x-\mu)'] = Cov(x)$.

Theorem: $Cov(x) = E[(x-\mu)(x-\mu)'] = E(xx') - \mu\mu'$.

Proof: See Greene.

Note: **In Greene, Cov(x) is denoted by Var(x).**

Definition: Covariance Matrix between Two Random Vectors

$X = (X_1, X_2, ..., X_n)'$ and $Y = (Y_1, Y_2, ..., Y_m)'$ are random vectors. Then,

$$Cov(x,y) = \begin{pmatrix} \text{cov}(x_1,y_1) & \text{cov}(x_1,y_2) & \dots & \text{cov}(x_1,y_m) \\ \text{cov}(x_2,y_1) & \text{cov}(x_2,y_2) & & \text{cov}(x_2,y_m) \\ \vdots & \vdots & & \vdots \\ \text{cov}(x_n,y_1) & \text{cov}(x_n,y_2) & \dots & \text{cov}(x_n,y_m) \end{pmatrix}.$$

Definition: (Expectation of random matrix)

Suppose that B_{ij} are RVs. Then,

$$B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1q} \\ B_{21} & B_{22} & \dots & B_{2q} \\ \vdots & \vdots & & \vdots \\ B_{p1} & B_{p2} & \dots & B_{pq} \end{bmatrix} \Rightarrow E(B) = \begin{bmatrix} E(B_{11}) & E(B_{12}) & \dots & E(B_{1q}) \\ E(B_{21}) & E(B_{22}) & \dots & E(B_{2q}) \\ \vdots & \vdots & & \vdots \\ E(B_{p1}) & E(B_{p2}) & \dots & E(B_{pq}) \end{bmatrix}$$

(4) Multivariate Normal distribution

Definition:

$X = [X_1, \dots, X_n]'$ is a normal vector, i.e., each of the x_j's is normal.

Let $E(x) = \mu = [\mu_1, \dots, \mu_n]'$ and $Cov(x) = \Sigma = [\Sigma_{ij}]_{n\times n}$. Then,

$$x \sim N(\mu,\Sigma).$$

Pdf of x:

$f(x) = f(x_1, \dots, x_n) = (2\pi)^{-n/2} |\Sigma|^{-1/2} \exp[-(1/2)(x-\mu)'\Sigma^{-1}(x-\mu)]$,

where $|\Sigma| = \det(\Sigma)$.

EX:

Let X be a single RV with $N(\mu_x, \sigma_x^2)$. Then,

$$f(x) = (2\pi)^{-1/2}(\sigma_x^2)^{-1/2}\exp[-(1/2)(x-\mu_x)(\sigma_x^2)^{-1}(x-\mu_x)]$$

$$= \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left[-\frac{(x-\mu_x)^2}{2\sigma_x^{\,2}}\right].$$

EX:

Assume that all the X_i ($i = 1, \ldots, n$) are iid with $N(\mu_x, \sigma_x^2)$. Then,

(1) $\mu = E(x) = [\mu_x, \ldots, \mu_x]'$;

(2) $\Sigma = Cov(x) = diag(\sigma_x^2, \sigma_x^2, \ldots, \sigma_x^2) = \sigma_x^2 I_n$.

Using (1) and (2), we can show that $f(x) = f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$,

where $f(x_i) = \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left[-\frac{(x_i-\mu_x)^2}{2\sigma_x^{\,2}}\right]$.

Theorem: Conditional normal distribution

$[y, x_2, \ldots, x_k]'$ is a normal vector. Then,

$$E(y|\, x_2,\ldots,x_k) = \beta_1 + \beta_2 x_2 + \ldots + \beta_k x_k = x^{*\prime}; \; var(y|x^*) = \sigma^2.$$

where $x^* = (1, x_2, \ldots, x_k)'$ and $\beta = (\beta_1, \ldots, \beta_k)'$]. That is, the regression of y on $x_1, \ldots, x_k$ is linear & homoskedastic.

Proof: See Greene.

(5) Properties of the Covariance Matrix of a Random Vector

Definition:

Let $X = [X_1, \ldots, X_n]'$ be a random vector and let $c = [c_1, \ldots, c_n]'$ be a $n\times1$ vector of fixed constants. Then,

$$c'x = x'c = c_1x_1 + \ldots + c_nx_n = \Sigma_j c_j x_j \text{ (scalar)}.$$

Theorem:

(1) $E(c'x) = c'E(x)$;

(2) $var(c'x) = c'Cov(x)c$.

Proof:

(1) $E(c'x) = E(\Sigma_j c_j x_j) = E(c_1x_1 + \ldots + c_nx_n)$

$= c_1E(x_1) + \ldots + c_nE(x_n) = \Sigma_j c_j E(x_j) = c'E(x)$.

(2) $var(c'x) = E[(c'x - E(c'x))^2] = E[\{c'x - c'E(x)\}^2]$

$= E[\{c'(x-E(x))\}^2] = E[\{c'(x-E(x))\}\{c'(x-E(x))\}]$

$= E[\{c'(x-E(x))\}\{(x-E(x))'c\}]$

$= E[c'(x-E(x))(x-E(x))'c] = c'E[(x-E(x))(x-E(x))']c = c'Cov(x)c$.

Remark:

(2) implies that Cov(x) is always positive semidefinite.

$\rightarrow c'Cov(x)c \geq 0$, for any nonzero vector c.

Proof:

For any nonzero vector c, $c'Cov(x)c = var(c'x) \geq 0$.

Remark:

- Cov(x) is symmetric and positive **semidefinite (what does it mean?)**.
- Usually, Cov(x) is positive definite, that is, $c'\mathrm{Cov}(x)c > 0$, for any nonzero vector c.

Definition:

Let $B = [b_{ij}]_{n \times n}$ be a symmetric matrix, and $c = [c_1, \ldots, c_n]'$. Then, a scalar $c'Bc$ is called a quadratic form of B.

Definition:

- If $c'Bc > (<) 0$ for any nonzero vector c, B is called positive (negative) definite.
- If $c'Bc \geq (\leq) 0$ for any nonzero c, B is called positive (negative) semidefinite.

Theorem:

Let B be a symmetric and square matrix given by:

$$B = \begin{bmatrix} b_{11} & b_{12} & \dots & b_{1n} \\ b_{21} & b_{22} & \dots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{n1} & b_{n2} & \dots & b_{nn} \end{bmatrix}.$$

Define the principal minors by:

$$|B_1| = b_{11}; |B_2| = \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix}; |B_3| = \begin{vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{vmatrix}; \dots.$$

B is positive definite iff $|B_1|$, $|B_2|$, ... , $|B_n|$ are all positive. B is negative definite iff $|B_1| < 0$, $|B_2| > 0$, $|B_3| < 0$, ...

EX:

Show that B is positive definite:

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

End of Digression

[2] Classical Linear Regression (CLR) Model

Example:

- Wish to find important determinants of individuals' earnings and estimate the size of the effect of each determinant.
- Data: (WAGE2.WF1 or WAGE2.TXT)

# of observations (T): 935

1. wage	monthly earnings
2. hours	average weekly hours
3. IQ	IQ score
4. KWW	knowledge of world work score
5. educ	years of education
6. exper	years of work experience
7. tenure	years with current employer
8. age	age in years
9. married	=1 if married
10. black	=1 if black
11. south	=1 if live in south
12. urban	=1 if live in SMSA
13. sibs	number of siblings
14. brthord	birth order
15. meduc	mother's education
16. feduc	father's education
17. lwage	natural log of wage

What variables would be important determinants of log(wage)?
From now on, we use both "log" and "ln" to refer to natural log.

Mincerian Wage Equation:

- Set y (dependent variable) = log(wage).
- Set x. (vector of independent variables) = [1, educ, exper, exper^2]′.
  - $x_{\bullet}$ = vector of independent variables (or explanatory variables, or regressors).
- Use subscript "o" for "true value".
- Assume $E(y \mid x_{\bullet}) = \beta_{1,o} + \beta_{2,o} educ + \beta_{3,o} \exp er + \beta_{4,o} \exp er^2$
- y = $E(y \mid x_{\bullet}) + \varepsilon = \beta_{1,o} + \beta_{2,o} educ + \beta_{3,o} \exp er + \beta_{4,o} \exp er^2 + \varepsilon$
- $y = x_{\bullet}'\beta_o + \varepsilon$, where $\beta_o = (\beta_{1,o}, \beta_{2,o}, \beta_{3,o}, \beta_{4,o})'$
- Here,
  - $\beta_{2,o} \times 100$ = %Δ in wage by one more year of education.
  - $(\beta_{3,o} + 2\beta_{4,o}\text{exper}) \times 100$ = %Δ by one more year of exper.

- Issues:
  - How to estimate β_o's?
  - Estimated β's would not be equal to the true values of β (β_o). How close would our estimates to the true values?

Basic Assumptions for CLR

(I call these assumptions **Strong Ideal Conditions (SIC)**.)

To understand SIC better; imagine a population of T-groups with the following properties.

- For each group t = 1, 2,..., T, y_t denotes the dependent variable and $x_{t\bullet}$ = $(x_{t1},x_{t2},...,x_{tk})'$ denotes the vector of regressors.
- The T-groups are assumed to be independent.
- Your sample consists of T observations, each of which comes from each different group.

As you may find, the above assumptions are unrealistic. But under the assumptions, more intuitive discussions about the statistical properties of OLS can be made. The statistical properties of OLS discussed later still hold even under more realistic assumptions.

Notation:

- $E(x_{t1})$ is the group population mean of x_1 for group t, while $E(x_1)$ is the population mean of x_1 for the whole population.

We now discuss each of SIC in detail:

(SIC.1) The conditional mean of y_t (dependent variable) given $x_{t\bullet}$ (vector of explanatory variables) is linear:

$$y_t = E(y_t | x_{t\bullet}) + \varepsilon_t = x_{t\bullet}'\beta_o + \varepsilon_t = \beta_{1,o}x_{t1} + \beta_{2,o}x_{t2} + ... + \beta_{k,o}x_{tk} + \varepsilon_t, \qquad \textbf{(1)}$$

where $x_{t\bullet} = (x_{t1}, x_{t2}, ..., x_{tk})'$ and $\beta_o = (\beta_{1,o}, ..., \beta_{k,o})'$.

Comment:

- Usually, $x_{t1} = 1$ for all t. That is, β_1 is an overall intercept term.
- $E(\varepsilon_t | x_{t\bullet}) = 0$.
- $E(x_{t\bullet}\varepsilon_t) = E_{x_{t\bullet}}[E(x_{t\bullet}\varepsilon_t | x_{t\bullet})] = E_{x_{t\bullet}}[x_{t\bullet}E(\varepsilon_t | x_{t\bullet})] = E_{x_{t\bullet}}(0) = 0$.

(SIC.2) $\beta_o = (\beta_{1,o}, ..., \beta_{k,o})'$ is unique.

Comment:

- No other β_* such that $E(y_t | x_{t\bullet}) = x_{t\bullet}'\beta_o = x_{t\bullet}'\beta_*$ for all t.
- The uniqueness assumption of β_o is called "identification" condition.
- Rules out perfect multicollinearity (perfect linear relationship among the regressors):
  - Suppose $\beta = (\beta_1, \beta_2, \beta_3)'$ and $x_{t3} = x_{t1} + x_{t2}$ for all t.
  - Set $\beta_{1,*} = \beta_{1,o} + a$; $\beta_{2,*} = \beta_{2,o} + a$; $\beta_{3,*} = \beta_{3,o} - a$ for an arbitrary $a \in \Re$.

    $x_{t\bullet}'\beta_* = x_{t1}\beta_{1,*} + x_{t2}\beta_{2,*} + x_{t3}\beta_{3,*}$
  - $= x_{t1}\beta_{1,o} + x_{t2}\beta_{2,o} + x_{t3}\beta_{3,o} + a(x_{t1} + x_{t2} - x_{t3})$

    $= x_{t\bullet}'\beta_o$
  - (SIC.2) rules out this possibility.

(SIC.3) The variables, y_t, x_{t1}, … , x_{tk}, have finite moments up to fourth order.

Comment:

- $E(y_t^2 x_{t2}^2)$, $E(x_{t3}x_{t4}^3)$, $E(x_{t3}^4)$, etc, exist.
- Rules out extreme outliers.
- We need this assumption for consistency and asymptotic normality of the OLS estimator.
- SIC implies the Weak Ideal Conditions (WIC) that will be discussed later.
- Violated if $x_{t2} = t$ or $x_{t2} = x_{t-1,2} + v_{t2}$.

(SIC.4) A random sample $\left\{ (y_t, x_{t1}, x_{t2}, ..., x_{tk})' \right\}_{t=1,...,T}$ is available and $T \geq k$.

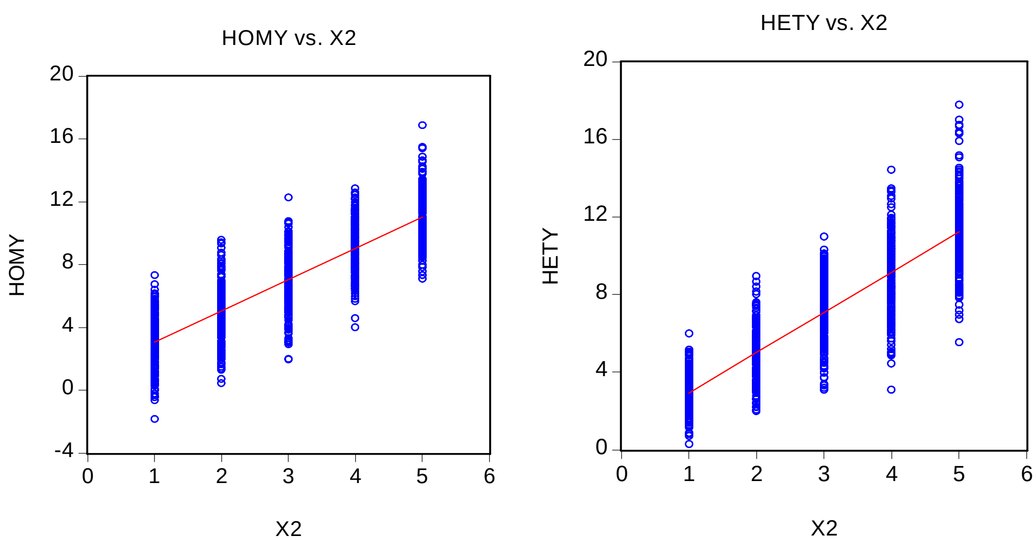
Comment:

- $(y_t, x_{t1}, x_{t2}, ..., x_{tk})'$ are iid (independently and identically distributed):
    - T groups which are iid with
    $$E\left(\begin{pmatrix} y_t \\ x_{t\bullet} \end{pmatrix}\right) = E\left(\begin{pmatrix} y \\ x \end{pmatrix}\right) \text{ and } Cov\left(\begin{pmatrix} y_t \\ x_{t\bullet} \end{pmatrix}\right) = Cov\left(\begin{pmatrix} y \\ x \end{pmatrix}\right).$$
    - One observation is drawn from each of the T group.
- Could be appropriate for cross-section data.
- Violated if time series data are used. That is why we add "strong" for the name of the conditions.
- If $T < k$, there are infinitely many β_* such that $x_{t\bullet}'\beta_o = x_{t\bullet}'\beta_*$ for all t. For this case, the sample cannot identify β.
- Implies no autocorrelation: $\mathrm{cov}(\varepsilon_t, \varepsilon_s) = 0$ for all $t \neq s$.

(SIC.5) $\text{var}(\varepsilon_t \mid x_{t\bullet}) = \sigma_o^2$, for all $x_{t\bullet}$. (Homoskedasticity Assumption).

Comment:

- Often violated when cross-section data are used.
- Consider the two different populations:
    - Population 1 (homoskedastic population):
        - homy = 1 + $2x_2$ + ε, where var(ε|x_2) = 9.
    - Population 2 (heteroskedastic population):
        - hety = 1 + $2x_2$ + ε, where var(ε|x_2) = x_2^2.
    - x_2 = 1, or 2, or 3, or 4, or 5, for both populations.



(SIC.6) The errors ε_t are normally distributed conditional on $x_{t\bullet}$.

(SIC.7) $x_{t1} = 1$, for all t = 1, ... , T.

Comment:

- Optional. Not critical.
- This condition implies that $\beta_{1,o}$ is an overall intercept term.
- Need this assumption for convenient interpretation of empirical R^2.

- Link between β_o and covariances:
  - Consider a simple regression model, $y_t = \beta_{1,o} + \beta_{2,o}x_{t2} + \varepsilon_t = x_{t\bullet}'\beta_o + \varepsilon_t$.
  - Assume (SIC.1) – (SIC.4) and (SIC.7).
  - $E(x_{t\bullet}y_t) = E\left(x_{t\bullet}(x_{t\bullet}'\beta_o + \varepsilon_t)\right) = E\left(x_{t\bullet}x_{t\bullet}'\right)\beta_o$

    $\rightarrow E(x_\bullet y) = E(x_\bullet x_\bullet')\beta_o$

    $\rightarrow \beta_o = [E(x_\bullet x_\bullet)]^{-1}E(x_\bullet y)$,

    where,

$$E(x_\bullet x_\bullet) = E\left(\begin{pmatrix}1\\ x_2\end{pmatrix}\begin{pmatrix}1 & x_2\end{pmatrix}\right) = E\begin{pmatrix}1 & x_2\\ x_2 & x_2^2\end{pmatrix} = \begin{pmatrix}1 & E(x_2)\\ E(x_2) & E(x_2^2)\end{pmatrix};$$

$$E(x_\bullet y) = E\left(\begin{pmatrix}1\\ x_2\end{pmatrix}y\right) = E\begin{pmatrix}y\\ x_2 y\end{pmatrix} = \begin{pmatrix}E(y)\\ E(x_2 y)\end{pmatrix}.$$

    $\rightarrow \beta_{2,o} = \dfrac{\operatorname{cov}(x_2, y)}{\operatorname{var}(x_2)}$; $\beta_{1,o} = E(y) - \beta_{2,o}E(x_2)$.

Theorem:

Let $y_t = x_{t\bullet}'\beta_o \equiv \beta_{1,o} + w_{t\bullet}'\beta_{w,o} + \varepsilon_t$, where $x_{t\bullet}' = (1, w_{t\bullet}')$, $\beta_o = (\beta_{1,o}, \beta_{w,o}')'$, $w_{t\bullet} = (x_{t2},..., x_{tk})'$ and $\beta_{w,o} = (\beta_{2,o},..., \beta_{k,o})'$. Suppose that this model satisfies (SIC.1)-(SIC4) and (SIC.7). Then,

$$\beta_o = \left(E(x_{t\bullet}x_{t\bullet}')\right)^{-1} E(x_{t\bullet}y_t) = \left(E(x_\bullet x_\bullet')\right)^{-1} E(x_\bullet y).$$

And,

$$\begin{aligned}\beta_{w,o} &= \left(E(w_\bullet w_\bullet') - E(w_\bullet)E(w_\bullet')\right)^{-1}\left(E(w_\bullet y) - E(w_\bullet)E(y)\right) \\ &= \left(Cov(w_\bullet)\right)^{-1} Cov(w_\bullet, y)\end{aligned};$$

Hint for proof:

$$\begin{aligned}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} &= \begin{pmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} A_{11}^{-1}A_{12} \\ -I \end{pmatrix}(A_{22} - A_{21}A_{11}^{-1}A_{12})\begin{pmatrix} A_{21}A_{11}^{-1} & -I \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & A_{22}^{-1} \end{pmatrix} + \begin{pmatrix} -I \\ -A_{22}^{-1}A_{21} \end{pmatrix}(A_{11} - A_{12}A_{22}^{-1}A_{21})\begin{pmatrix} -I & A_{12}A_{22}^{-1} \end{pmatrix}\end{aligned},$$

where 0's here are zero matrices.

Implications:

- The slopes, $\beta_{2,o}$, … , $\beta_{k,o}$, measure the correlations between regressors and dependent variables.
- $\beta_{2,o} \neq 0$ means non-zero correlation between y_t and x_{t2}. It does not mean that x_{t2} causes y_t. $\beta_{2,o} \neq 0$ could mean that y_t causes x_{t2}.
- SIC do not talk about causality. SIC may hold even if y_t determines $x_{t\bullet}$: It can be the case that $E(edu_t \mid wage_t) = \beta_{1,o} + \beta_{2,o} wage_t$.
- But, the regression model (1) is not meaningful if the x variables are not causal variables. We would like to know by how much hourly wage rate increases with one more of education. We would not be interested in how many more years of education an individual could have obtained if his/her current wage rate increased now by $1!

[3] Ordinary Least Squares (OLS)

Definition:

For a given sample $\left\{\left(y_t, x_{t1}, ..., x_{tk}\right)'\right\}_{t=1,...,T}$ without perfect multicollinearity among regressors $x_{t1}, ..., x_{tk}$, the OLS estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k)'$ minimizes:

$$S_T(\beta) \equiv \Sigma_t (y_t - x_{t1}\beta_1 - ... - x_{tk}\beta_k)^2$$
$$= \Sigma_t (y_t - x_{t\bullet}'\beta) = (y - X\beta)'(y - X\beta),$$

where $\Sigma_t = \Sigma_{t=1}^{T}$, and

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{pmatrix}; X = \begin{pmatrix} x_{1\bullet}' \\ x_{2\bullet}' \\ \vdots \\ x_{T\bullet}' \end{pmatrix}; x_{t\bullet}' = (x_{t1}, x_{t2}, ..., x_{tk}).$$

Comment on the assumption of no perfect multicollinearity.

- $rank(X'X) = rank(X) = k$. So, $X'X = \Sigma_t x_{t\bullet} x_{t\bullet}'$ is invertible.
- If perfect multicollinearity exists, $rank(X'X) = rank(X) < k$. So, $X'X = \Sigma_t x_{t\bullet} x_{t\bullet}'$ is not invertible.
- If $T < k$, $rank(X'X) = rank(X) \leq \min(T, k) < k$. So, $X'X = \Sigma_t x_{t\bullet} x_{t\bullet}'$ is not invertible. $T < k$ is a case of perfect multicollinearity.

EX: Simple Regression Model

- Wish to estimate $y_t = \beta_{1,o}x_{t1} + \beta_{2,o}x_{t2} + \varepsilon_t$:

$$S_T(\beta_1,\beta_2) = \Sigma_t(y_t - x_{t1}\beta_1 - x_{t2}\beta_2)^2 .$$

- The first order condition for minimization:

$$\partial S_T/\partial\beta_1 = \Sigma_t 2(y_t - x_{t1}\beta_1 - x_{t2}\beta_2)(-x_{t1}) = 0 \rightarrow \Sigma_t(x_{t1}y_t - x_{t1}^2\beta_1 - x_{t1}x_{t2}\beta_2) = 0$$
$$\partial S_T/\partial\beta_2 = \Sigma_t 2(y_t - x_{t1}\beta_1 - x_{t2}\beta_2)(-x_{t2}) = 0 \rightarrow \Sigma_t(x_{t2}y_t - x_{t1}x_{t2}\beta_1 - x_{t2}^2\beta_2) = 0$$

$\rightarrow$ $\Sigma_t x_{t1}y_t = (\Sigma_t x_{t1}^2)\beta_1 + (\Sigma_t x_{t1}x_{t2})\beta_2$

$\Sigma_t x_{t2}y_t = (\Sigma_t x_{t1}x_{t2})\beta_1 + (\Sigma_t x_{t2}^2)\beta_2$

$$\rightarrow \quad \begin{pmatrix} \Sigma_t x_{t1}y_t \\ \Sigma_t x_{t2}y_t \end{pmatrix} = \begin{pmatrix} \Sigma_t x_{t1}^2 & \Sigma_t x_{t1}x_{t2} \\ \Sigma_t x_{t2}x_{t1} & \Sigma_t x_{t2}^2 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}.$$

$\rightarrow$ But, this equation is equivalent to $X'y = X'X\hat{\beta}$.

$\rightarrow$ $\hat{\beta} = (X'X)^{-1}X'y$.

Derivation of the OLS estimator for general cases:

- $S_T(\beta) = (y' - \beta'X')(y - X\beta) = y'y - \beta'X'y - y'X\beta + \beta'X'X\beta$.
- Since $y'X\beta$ is a scalar, $y'X\beta = (y'X\beta)' = \beta'X'y$.
- Thus, $S_T(\beta) = y'y - 2\beta'X'y + \beta'X'X\beta$.
- FOC for minimization of $S_T(\beta)$: $\frac{\partial S_T(\beta)}{\partial\beta} \equiv \begin{pmatrix} \frac{\partial S_T(\beta)}{\partial\beta_1} \\ \frac{\partial S_T(\beta)}{\partial\beta_2} \\ \vdots \\ \frac{\partial S_T(\beta)}{\partial\beta_k} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} = 0_{k\times 1}$.

But,

$$\partial(\beta'X'y)/\partial\beta = X'y;$$

$$\partial(\beta'X'X\beta)/\partial\beta = 2X'X\beta.$$

[In fact, for any k×1 vector d, $\partial(\beta'd)/\partial\beta = d$; and, for any k×k symmetric matrix A, $\partial(\beta'A\beta)/\partial\beta = 2A\beta$.]

Thus, FOC implies

$$\frac{\partial S_T(\beta)}{\partial\beta} = -2X'y + 2X'X\beta = 0_{k\times 1}$$

→

$$X'y - X'X\beta = 0_{k\times 1} \qquad \textbf{(2)}$$

→ Solving (2), we have

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

SOC (second order condition) for minimization:

$$\frac{\partial^2 S_T(\beta)}{\partial\beta\partial\beta'} = \left[\frac{\partial^2 S_T(\beta)}{\partial\beta_i\partial\beta_j}\right]_{k\times k} = 2X'X,$$

which is a positive definite matrix for any value of β. That is, the function $S_T(\beta)$ is globally convex. This indicates that $\hat{\beta}$ indeed minimizes $S_T(\beta)$. [Here, we use the fact that $\partial(\beta'A\beta)/\partial\beta\partial\beta' = 2A$ for any symmetric matrix A.]

Theorem: $\hat{\beta} = (X'X)^{-1}X'y$.

Definition:

- t'th residual: $e_t = y_t - x_{t\bullet}'\hat{\beta}$ (can be viewed as an estimate of ε_t).
- Vector of residuals: $e = (e_1, ..., e_T)' = y - X\hat{\beta}$.

Theorem: $X'e = 0_{k\times 1}$

Proof:

From the proof of the previous theorem,

$$X'y - X'X\hat{\beta} = 0_{k\times 1} \rightarrow X'(y - X\hat{\beta}) = 0_{k\times 1} \rightarrow X'e = 0 .$$

Corollary:

If (SIC.7) holds ($x_{t1} = 1$ for all t: β_1 is the intercept), $\Sigma_t e_t = 0$.

Proof:

$$X'e = \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{T1} \\ x_{12} & x_{22} & \cdots & x_{T2} \\ \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{Tk} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_T \end{bmatrix} = \begin{bmatrix} \Sigma_t x_{t1} e_t \\ \Sigma_t x_{t2} e_t \\ \vdots \\ \Sigma_t x_{tk} e_t \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{k\times 1} .$$

$\rightarrow \Sigma_t x_{t1} e_t = 0 \rightarrow \Sigma_t e_t = 0$ (by SIC.7).

Question:

Consider the following two models:

(A) $y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + x_{t3}\beta_3 + \varepsilon_t$;

(B) $y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \varepsilon_t$.

Are the OLS estimates of β_1 and β_2 from (A) the same as those from (B)?

Digression to Matrix Algebra

Definition: Let A be a T×p matrix.

$P(A) = A(A'A)^{-1}A'$ (T×T matrix called "projection matrix");

$M(A) = I_T - P(A) = I_T - A(A'A)^{-1}A'$ (T×T matrix called "residual maker).

Facts:

1) P(A) and M(A) are both symmetric and idempotent:

$P(A)' = P(A)$, $M(A)' = M(A)$, $P(A)P(A) = P(A)$, $M(A)M(A) = M(A)$.

2) P(A) and M(A) are psd (positive semi-definite).

3) $P(A)M(A) = 0_{T\times T}$ (orthogonal).

4) $P(A)A = [A(A'A)^{-1}A']A = A$.

5) $M(A)A = [I_T-P(A)]A = A - P(A)A = A - A = 0_{T\times T}$.

End of Digression

Theorem: $e = M(X)y$.

<Proof> $e = y - X\hat{\beta} = I_T y - X(X'X)^{-1}X'y = [I_T - X(X'X)^{-1}X']y = M(X)y$.

Frisch-Waugh Theorem:

Partition X into $[X_A, X_B]$ and $\beta = (\beta_A', \beta_B')'$. Let $\hat{\beta}_A$ be the OLS estimate of β_A from a regression of the model $y = X\beta + \varepsilon = X_A\beta_A + X_B\beta_B + \varepsilon$. Then,

$$\hat{\beta}_A = [X_A'M(X_B)X_A]^{-1}X_A'M(X_B)y.$$

That is, $\hat{\beta}_A$ is obtained by regressing $M(X_B)y$ on $M(X_B)X_A$.

Comment:

$\hat{\beta}_A$ is different from the OLS estimate of β_A from a regression of y on X_A.

Theorem:

Consider the following models:

(A) $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \text{error}$

(B) $y_t = \alpha_1 + \alpha_2 x_{t2} + \text{error}$

(C) $x_{t3} = \delta_1 + \delta_2 x_{t2} + \text{error}$

Then, $\hat{\alpha}_2 = \hat{\beta}_2 + \hat{\delta}_2\hat{\beta}_3$.

Theorem:

Consider the following two models:

(A) $y_t = \beta_1 + \beta_2 x_{t2} + ... + \beta_k x_{tk} + \varepsilon_t$;

(B) $y_t - \overline{y} = \beta_2(x_{t2} - \overline{x}_2) + ... + \beta_k(x_{tk} - \overline{x}_k) + error$.

Then, the OLS estimates of β_2, ... , β_k from the regression of (A) are the same as the OLS estimates of β_2, ..., β_k from the regression of (B).

Proof :

Model (A) can be written as

$$y = X\beta = 1_T\beta_1 + X_*\beta_* + \varepsilon,$$

where 1_T is the T×1 vector of ones and $\beta_* = (\beta_2,...,\beta_k)'$. Then,

$$\hat{\beta}_* = \left(X_*'M(1_T)X_*\right)^{-1} X_*'M(1_T)y.$$

Observe that:

$$M(1_T)y = \left(y_1 - \overline{y} \quad y_2 - \overline{y} \quad ... \quad y_T - \overline{y}\right)'.$$

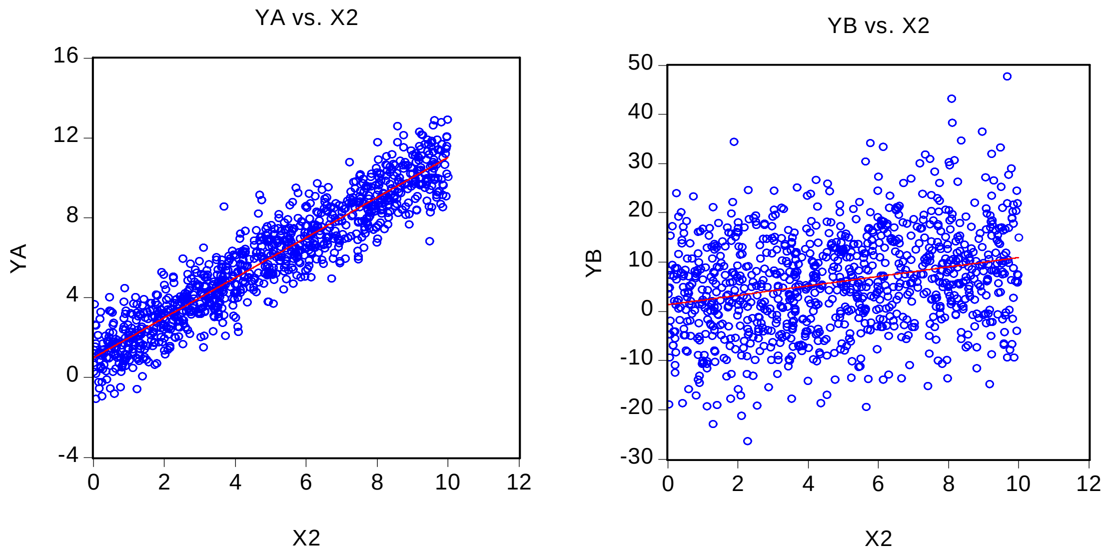
Now, complete the proof by yourself.

[4] Goodness of Fit

Question: How well does your regression explain y_t?

Example:

- A simple regression model: $y_t = \beta_1 + \beta_2 x_{t2} + \varepsilon_t$, with $\beta_{1,o} = \beta_{2,o} = 1$.
- For population A, $\sigma_o^2 = 1$. For population B, $\sigma_o^2 = 10$.



- Clearly, the regression line $E(y_t \mid x_{t\bullet})$ explains Population A better.
- How can we measure the goodness of fit of $E(y_t \mid x_{t\bullet})$?

Definition:

- "Fitted value" of y_t: $\hat{y}_t = x_{t\bullet}'\hat{\beta}$ (an estimate of $E(y_t \mid x_{t\bullet})$).
- Vector of fitted values: $\hat{y} = X\hat{\beta}$.

Definition:

SSE $= e'e = (y - X\hat{\beta})'(y - X\hat{\beta}) = (y - \hat{y})'(y - \hat{y}) = \Sigma_t (y_t - \hat{y}_t)^2$.

(Unexplained sum of squares)

$\rightarrow$ Measures unexplained variation of y_t.

$\rightarrow$ SSE/T is an estimate of $E_x[var(y|x)]$.

SSR $= \Sigma_t(\hat{y}_t - \bar{y})^2$, where $\bar{y} = T^{-1}\Sigma_t y_t$ (Explained sum of squares).

$\rightarrow$ Measures variation of y_t explained by regression.

$\rightarrow$ SSR/T is an estimate of $var_x[E(y|x)]$.

SST $= \Sigma_t(y_t - \bar{y})^2$ (Total sum of squares)

$\rightarrow$ SST/T measures total variation of y_t.

Theorem: SSE $= \Sigma_t e_t^2 = y'y - \hat{\beta}'X'y$.

Proof:

SSE $= (y - X\hat{\beta})'(y - X\hat{\beta}) = y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta}$

$= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X(X'X)^{-1}X'y = y'y - \hat{\beta}'X'y$.

Theorem:

SST $= \Sigma_t(y_t - \bar{y})^2 = \Sigma_t y_t^2 - T\bar{y}^2$,

SSR $= \hat{\beta}'X'y - T\bar{y}^2$ [if (SIC.7) holds].

Proof: For SSR, see Schmidt.

Theorem:

Suppose that $x_{t1} = 1$, for all t (that is, (SIC.7) holds). Then, SST = SSE + SSR.

Proof: Obvious.

Implication:

Total variation of y_t equals sum of explained and unexplained variations of y_t.

Definition: [Measure of goodness of fit]

$R^2 = 1 - (SSE/SST) = (SST-SSE)/SST$.

Theorem:

Suppose that $x_{t1} = 1$, for all t (SIC.7). Then, $R^2 = SSR/SST$ and $0 \leq R^2 \leq 1$.

Note:

1) If (SIC.7) holds, then, $R^2 = 1 - (SSE/SST) = SSR/SST$.

2) If (SIC.7) does not hold, then, $1 - (SSE/SST) \neq SSR/SST$.

3) $1 - (SSE/SST)$ can never be greater than 1, but it could be negative.

SSR/SST can never be negative, but it could be greater than 1.

Definition:

R_u^2 (uncentered R^2) $= \hat{y}'\hat{y} / y'y = \Sigma_t \hat{y}_t^2 / \Sigma_t y_t^2$.

Note:

- Some people use R_u^2, when the model has no intercept term.
- $0 \le R_u^2 \le 1$, since $e'e + \hat{y}'\hat{y} = y'y$. [Why? Try it at home.]

  $\rightarrow$ This holds even if (SIC.7) does not hold.
- If $\bar{y} = 0$, then, $R_u^2 = R^2$.

Definition:

An estimator of covariance between y_t and $\hat{y}_t$ (which be viewed as an estimate of $E(y_t \mid x_{t\bullet})$) is defined by:

$$e\,\mathrm{cov}(y_t, \hat{y}_t) = \frac{1}{T-1}\Sigma_t (y_t - \bar{y})(\hat{y}_t - \tilde{y}),$$

where $\tilde{y} = T^{-1}\Sigma_t \hat{y}_t$. Similarly, the estimators of var(y_t) and var($\hat{y}_t$) are defined by:

$$e\,\mathrm{var}(y_t) = \frac{1}{T-1}\Sigma_t (y_t - \bar{y})^2; e\,\mathrm{var}(\hat{y}_t) = \frac{1}{T-1}\Sigma_t (\hat{y}_t - \tilde{y})^2.$$

Then, the estimated correlation coefficient between y_t and $\hat{y}_t$ is defined by:

$$\hat{\rho} = \frac{e\,\mathrm{cov}(y_t, \hat{y}_t)}{\sqrt{e\,\mathrm{var}(y_t)}\sqrt{e\,\mathrm{var}(\hat{y}_t)}}.$$

Note:

1) $0 \leq \hat{\rho}^2 \leq 1$, whether (SIC. 7) holds or not.

2) If (SIC.7) holds, $\tilde{y} = \bar{y}$.

3) If (SIC.7) holds, 1-(SSE/SST) = SSR/SST = $\hat{\rho}^2$.

Remark for the case where (SIC.7) holds:

1) If $R^2 = 1$, y_t and $\hat{y}_t$ are perfectly correlated (perfect fit).

2) If $R^2 = 0$, y_t and $\hat{y}_t$ have no correlation.

$\rightarrow$ Regression may not be much useful.

3) Does a high R^2 always mean that your regression is good?

[Answer]

No. If you use more regressors, then, you will get higher R^2. In particular, if $k = T$, $R^2 = 1$.

4) R^2 tends to exaggerate goodness of fit when T is small.

Definition: [Adjusted R^2, Theil (1971)]

$$\overline{R}^2 = 1 - \frac{SSE/(T-k)}{SST/(T-1)}.$$

Comment:

- $\overline{R}^2 < R^2$ unless $k > 1$ and $R^2 < 1$.
- $\overline{R}^2$ could be negative.

[Proof for the fact that R^2 increases with k]

Theorem: Let $A = [A_1, A_2]$. Then,

$$M(A)A_j = 0;\ P(A)A_j = A_j,\ j = 1, 2;\ P(A) = P(A_1) + P[M(A_1)A_2].$$

Theorem: $\hat{y} = P(X)y$ and $e = M(X)y$.

Proof: Because $\hat{y} = X\hat{\beta} = X(X'X)^{-1}X'y = P(X)y$. And $e = y - P(X)y = M(X)y$.

Lemma: $SSE = y'M(X)y = y'y - y'P(X)y$.

Proof: $SSE = e'e = [M(X)y]'M(X)y = y'M(X)'M(X)y = y'M(X)y$.

Theorem:

When k increases, SSE never increases.

Proof:

Compare:

Model 1: $y = X\beta + \varepsilon$

Model 2: $y = X\beta + Z\gamma + \upsilon = W\xi + \upsilon$,

where $W = [X,Z]$ and $\xi = [\beta',\gamma']'$.

SSE_1 = SSE from M1 = $y'M(X)y = y'y - y'P(X)y$

SSE_2 = SSE from M2 = $y'M(W)y = y'y - y'P(W)y$

$= y'y - y'[P(X)+P\{M(X)Z\}]y$

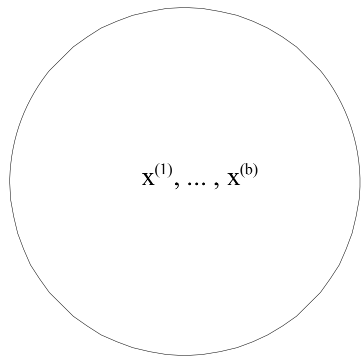
$= y'y - y'P(X)y - y'P\{M(X)Z\}y$

$SSE_1 - SSE_2 = y'P\{M(X)Z\}y \geq 0$.

[5] Statistical Properties of the OLS estimator

(1) Random Sample:

- A population (of billions and billions)



Here, the $x^{(j)}$ are the members of the population.

- θ: An unknown parameter of interest (e.g., population mean or population variance.)
    - If we know the pdf of this population, we could easily compute θ. But if you do not know the pdf?
- Need to estimate θ, using a random sample $\{x_1, \ldots , x_T\}$ of size T from the population.

- What do we mean by "random sample"?
  - A sample that represents the population well.
  - Divide the population into T groups such that the groups are stochastically independent and the pdf of each group is the same as the pdf of the whole population. Then, draw one from each group: Then, the $x_1, \ldots, x_T$ should be iid (independently and identically distributed).
  - "Random sample" means a sample obtained by this sampling strategy.
  - An example of nonrandom sampling:
    - Suppose you wish to estimate the % of supporters of the Republican Party in the Phoenix metropolitan area.
    - t is a zip-code area. Choose a person living in a street corner from each t.
    - If you do, your sample is not random. Because rich people are likely to live in corner houses! Republicans are over-sampled!

- Let $\hat{\theta}$ be an estimator of θ. What properties should $\hat{\theta}$ have?

(2) Criteria for "good" estimators

1) Unbiasedness.
2) Small variance.
3) Distributed following a known form of pdf (e.g., normal, or χ^2).

Definition: (Unbiasedness)

If $E(\hat{\theta}) = \theta_o$, then we say that $\hat{\theta}$ is an unbiased estimator of θ.

Comment:

- Consider the set of all possible random samples of size T:

$$
\begin{array}{lcc}
 & & \text{Estimate} \\
\text{Sample 1: } \{x_1^{[1]}, x_2^{[1]}, \ldots, x_T^{[1]}\} & \rightarrow & \hat{\theta}^{[1]} \\
\text{Sample 2: } \{x_1^{[2]}, x_2^{[2]}, \ldots, x_T^{[2]}\} & \rightarrow & \hat{\theta}^{[2]} \\
\text{Sample 3: } \{x_1^{[3]}, x_2^{[3]}, \ldots, x_T^{[3]}\} & \rightarrow & \hat{\theta}^{[3]} \\
\vdots & & \\
\text{Sample b}'\text{: } \{x_1^{[b']}, x_2^{[b']}, \ldots, x_T^{[b']}\} & \rightarrow & \hat{\theta}^{[b']}.
\end{array}
$$

- Consider the population of $S_\theta \equiv \{\hat{\theta}^{[1]}, \ldots, \hat{\theta}^{[b']}\}$.
- Unbiasedness of $\hat{\theta}$ means that $E(\hat{\theta})$ = population average of $S_\theta = \theta_o$.

Definition: (Relative Efficiency)

Let $\hat{\theta}$ and $\breve{\theta}$ be unbiased estimators of θ. If $\mathrm{var}(\hat{\theta}) < \mathrm{var}(\breve{\theta})$, we say that $\hat{\theta}$ is more efficient than $\breve{\theta}$.

Comment:

If $\hat{\theta}$ is more efficient than $\breve{\theta}$, it means that the value of $\hat{\theta}$ that I can obtain from a particular sample would be generally closer to the true value of θ (θ_o) than the value of $\breve{\theta}$.

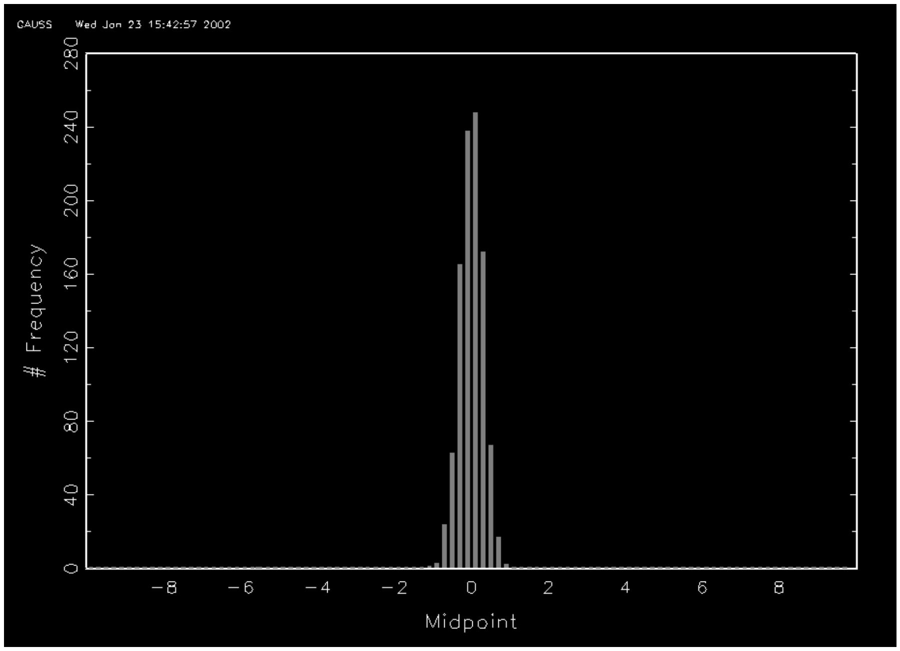
Example:

- A population is normally distributed with $N(\mu,\sigma^2)$, where $\mu_o = 0$ and $\sigma_o^2 = 9$.
- $\{x_1, x_2, \ldots, x_T\}$ is a random sample ($T = 100$):
- Two possible unbiased estimators of μ: $\bar{x} = \frac{1}{T}\Sigma_t x_t$ and $\hat{x} = x_1$.
    - $E(\bar{x}) = E\left(\frac{1}{T}\Sigma_t x_t\right) = \frac{1}{T}\Sigma_t E(x_t) = \frac{1}{T}\Sigma_t \mu_o = \mu_o$; $E(\hat{x}) = E(x_1) = \mu_o$.
- Which estimator is more efficient?
    - $\mathrm{var}(\bar{x}) = \mathrm{var}\left(\frac{1}{T}\Sigma_t x_t\right) = \left(\frac{1}{T}\right)^2 \Sigma_t \mathrm{var}(x_t) = \left(\frac{1}{T}\right)^2 \Sigma_t \sigma_o^2 = \frac{\sigma_o^2}{T}$;
    - $\mathrm{var}(\hat{x}) = \mathrm{var}(x_1) = \sigma_o^2$.
    - Thus, $\mathrm{var}(\bar{x}) = \frac{\sigma_o^2}{T} < \sigma_o^2 = \mathrm{var}(\hat{x})$, if $T > 1$.

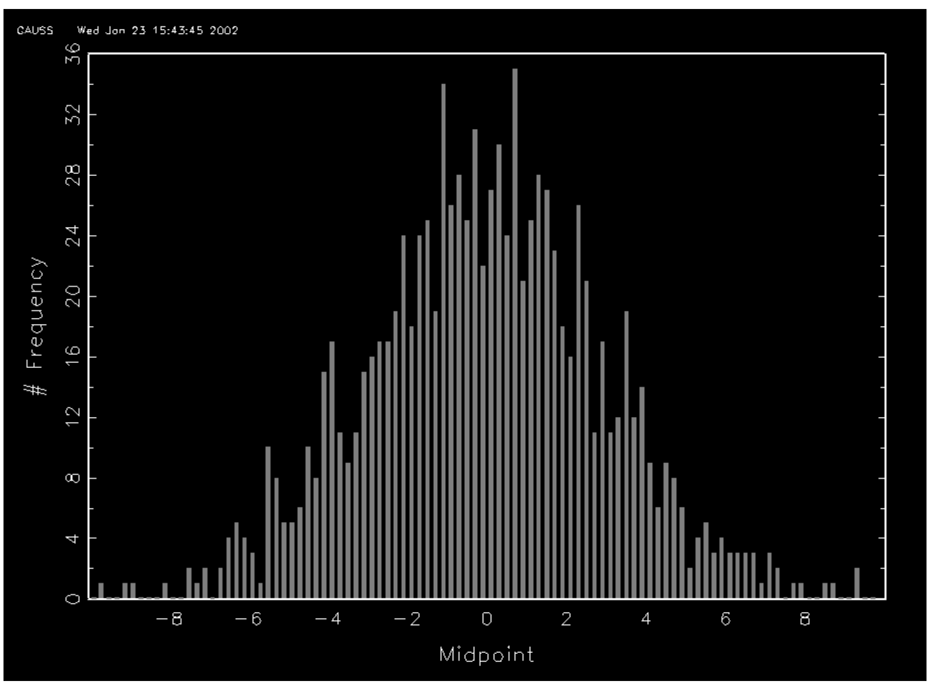
Gauss Exercise:

- From N(0,9), draw 1,000 random samples of size equal to T = 100.
- For each sample, compute $\bar{x}$ and $\hat{x}$.
- Draw a histogram for each estimator.
- Gauss program name: mmonte.prg.

```
/*
**  Monte Carlo Program for sample mean
*/

seed   = 1;
tt     = 100; @ # of observations @
iter   = 1000; @ # of sets of different data @

storem  = zeros(iter,1) ;
stores  = zeros(iter,1) ;

i = 1; do while i <= iter;

@ compute sample mean for each sample @

x = 3*rndns(tt,1,seed);
m = meanc(x);
storem[i,1]  = m;
stores[i,1]  = x[1,1];

i = i + 1; endo;

@ Reporting Monte Carlo results @

output file = mmonte.out reset;

format /rd 12,3;

"Monte Carlo results";
"-----------";
"Mean of x bar     ="  meanc(storem);
"mean of x rou     ="  meanc(stores);
library pgraph;
graphset;

v = seqa(-10, .2, 100);
{a1,a2,a3}=hist(storem,v);
@ {b1,b2,b3}=hist(stores,v);  @

output off ;
```


CAUSS Wed Jan 23 15:42:57 2002
Frequency
Midpoint

$\bar{x}$


CAUSS Wed Jan 23 15:43:45 2002
Frequency
Midpoint

$\hat{x}$

Extension to the Cases with Multiple Parameters:

- $\theta = (\theta_1,\theta_2, \ldots , \theta_p)'$ is a unknown parameter vector.

Definition: (Unbiasedness)

$\hat{\theta}$ is unbiased iff $E(\hat{\theta}) = \theta_o$:

$$E(\hat{\theta}) = \begin{bmatrix} E(\hat{\theta}_1) \\ E(\hat{\theta}_2) \\ \vdots \\ E(\hat{\theta}_p) \end{bmatrix} = \begin{bmatrix} \theta_{1,o} \\ \theta_{2,o} \\ \vdots \\ \theta_{p,o} \end{bmatrix} = \theta_o .$$

Definition: (Relative Efficiency)

Suppose that $\hat{\theta}$ and $\breve{\theta}$ are unbiased estimators. Let $c = (c_1,c_2, \ldots , c_p)'$ is a nonzero vector. $\hat{\theta}$ is said to be more efficient than $\breve{\theta}$, iff $\mathrm{var}(c'\breve{\theta}) \geq \mathrm{var}(c'\hat{\theta})$ for any nonzero vector c.

Remark:

$\mathrm{var}(c'\breve{\theta}) \geq \mathrm{var}(c'\hat{\theta})$.

$\leftrightarrow c'Cov(\breve{\theta})c - c'Cov(\hat{\theta})c \geq 0$, for any nonzero c.

$\leftrightarrow c'[Cov(\breve{\theta}) - Cov(\hat{\theta})]c \geq 0$, for any nonzero c.

$\leftrightarrow Cov(\breve{\theta}) - Cov(\hat{\theta})$ is positive semi-definite.

Comment:

- Let $\theta = (\theta_1,\theta_2)'$ and $c = (c_1,c_2)'$.
- Suppose you wish to estimate $c'\theta = c_1\theta_1 + c_2\theta_2$.
- If, for any nonzero c, $\text{var}(c'\tilde{\theta}) = \text{var}(c_1\tilde{\theta}_1+c_2\tilde{\theta}_2) \geq \text{var}(c_1\hat{\theta}_1+c_2\hat{\theta}_2) = \text{var}(c'\hat{\theta})$, we say that $\hat{\theta}$ is more efficient than $\tilde{\theta}$.

Example:

- Let $\theta = (\theta_1, \theta_2)'$. Suppose $Cov(\hat{\theta}) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}; Cov(\tilde{\theta}) = \begin{bmatrix} 1.5 & 1 \\ 1 & 1.5 \end{bmatrix}$.
- Note that:

$$\text{var}(\hat{\theta}_1) = 1 < 1.5 = \text{var}(\tilde{\theta}_1);\ \text{var}(\hat{\theta}_2) = 1 < 1.5 = \text{var}(\tilde{\theta}_2).$$

- But,

$$Cov(\tilde{\theta}) - Cov(\hat{\theta}) = \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix} \equiv A \ \rightarrow |A_1| = 0.5 > 0; |A_2| = -0.75 < 0.$$

- A is neither positive nor negative semi-definite.
- $\hat{\theta}$ is not necessarily more efficient than $\tilde{\theta}$.
- For example, suppose you wish to estimate $\theta_1-\theta_2 = c'\theta$ (where $c = (1,-1)'$):
  - $\text{var}(c'\hat{\theta}) = c'Cov(\hat{\theta})c = 2$; $\text{var}(c'\tilde{\theta}) = c'Cov(\tilde{\theta})c = 1$.
  - That is, for the given $c = (1,-1)'$, $c'\tilde{\theta}$ is more efficient than $c'\hat{\theta}$.
  - This example is a case where relative efficiency of estimators depends on c. For such cases, we can't claim that one estimator is superior to others.

Theorem:

If $\hat{\theta}$ is more efficient than $\tilde{\theta}$, $\text{var}(\hat{\theta}_j) \leq \text{var}(\tilde{\theta}_j)$, for all j = 1, ..., p. But not vice versa.

Proof:

Choose c = (1,0,...,0)′. Then, you can show $\text{var}(\hat{\theta}_1) \leq \text{var}(\tilde{\theta}_1)$. Now, choose c = (0,1,0,....,0)′. Then, we can show $\text{var}(\hat{\theta}_2) \leq \text{var}(\tilde{\theta}_2)$. Keep doing this until j = p.

(3) Population Projection

- Suppose you have data from all population members (say, t = 1, …., B).
- Assume that $E(x_{\bullet}x_{\bullet}') = \frac{1}{B}\Sigma_{t=1}^{B} x_{t\bullet}x_{t\bullet}'$ is pd, where $x_{t1} = 1$ for all t.
- Let $\beta_p = (\beta_{1,p},...,\beta_{2,p})'$ be the OLS estimator obtained using all population:

  Notice that β_p is a population parameter vector. Denote

$$\Pr oj(y_t \mid x_{t\bullet}) = x_{t\bullet}'\beta_p.$$

- Let $e_{p,t} = y_t - x_{t\bullet}'\beta_p$, where t = 1, … , B.
- Population projection model:

$$y_t = \Pr oj(y_t \mid x_{t\bullet}) + \varepsilon_t = x_{t\bullet}'\beta_p + e_{p,t}.$$

- By definition, β_p always exists. Notice that (SIC.1) *assumes* that the conditional mean of y_t is linear in $x_{t\bullet}$: $E(y_t \mid x_{t\bullet}) = x_{t\bullet}'\beta_o$. In contrast, the population projection of y_t is always linear.

Theorem:

$E(x_j e_p) = 0$ for all j = 1, … , k. That is, $E(x_{\bullet}e_p) = 0_{k\times 1}$.

Proof:

Recall $X'e = 0_{k\times 1} \rightarrow \Sigma_{t=1}^{T} x_{t\bullet}e_t = 0_{k\times 1}$. That is, $E(x_{\bullet}e_p) = \frac{1}{B}\Sigma_{t=1}^{B} x_{t\bullet}e_{p,t} = 0$.

Comment:

- $E(x_{\bullet}e_p) = 0 \rightarrow E(e_p \mid x_{\bullet}) = 0$, although the latter implies the former.

Theorem:

$$\beta_p = \left(E(x_\bullet x_\bullet')\right)^{-1} E(x_\bullet y).$$

Proof:

$$\beta_p = \left(\Sigma_{t=1}^{B} x_{t\bullet} x_{t\bullet}'\right)^{-1} \Sigma_{t=1}^{B} x_{t\bullet} y_t = \left(\frac{1}{B}\Sigma_{t=1}^{B} x_{t\bullet} x_{t\bullet}'\right)^{-1} \frac{1}{B}\Sigma_{t=1}^{B} x_{t\bullet} y_t .$$

Comment:

- Intuitively, the OLS estimator is a consistent estimator of β_p.
- Notice that under (SIC.1)-(SIC.4), $\beta_o = \beta_p$!
- Under (SIC.1)-(SIC.4), $E(y_t \mid x_{t\bullet}) = \Pr oj(y_t \mid x_{t\bullet}) = x_{t\bullet}'\beta_o$.
- Thus, under (SIC.1)-(SIC.4), the OLS estimator is a consistent estimator of β_o.

(4) The Stochastic Properties of the OLS Estimator.

(SIC.8) The regressor $x_{t1}, \ldots, x_{tk}$ ($x_{t\bullet}$) are nonstochastic.

Comment:

- The whole population consists of T groups, and each group has fixed $x_{t\bullet}$. We draw y_t from each group. The value of y_t would change over different trials, but the value of $x_{t\bullet}$ remains the same.
- Can be replaced by the assumption that $E(\varepsilon_t \mid x_{1\bullet}, \ldots, x_{T\bullet}) = 0$ for all t (assumption of strictly exogenous regressors). This assumption holds as long as (SIC.1) - (SIC.4) hold. If you do not use (SIC.8), the distributions of $\hat{\beta}$ and s^2 obtained below the conditional ones conditional on $x_{1\bullet}, x_{2\bullet}, \ldots, x_{T\bullet}$.

Theorem:

Assume (SIC.1)-(SIC.6) and (SIC.8). Then,

- $E(\hat{\beta}) = \beta_o$ (unbiased)
- $Cov(\hat{\beta}) = \sigma_o^2 (X'X)^{-1}$
- $E(s^2) = \sigma_o^2$, where $s^2 = SSE/(T-k) = \Sigma_t e_t^2/(T-k) = e'e/(T-k)$

  [even if the ε_t are not normal, that is, (SIC.6) does not hold]
- $\hat{\beta} \sim N(\beta_o, \sigma_o^2 (X'X)^{-1})$.
- $\hat{\beta}$ and SSE (so s^2) are stochastically independent.
- $SSE/\sigma_o^2 \sim \sigma^2(T-k)$ [if (SIC.6) holds.]

Comment:

- As discussed later, we need to estimate $Cov(\hat{\beta}) = \sigma_o^2 (X'X)^{-1}$.
- We can use s^2 to estimate $Cov(\hat{\beta})$.

Numerical Exercise:

- $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3} + \varepsilon_t$, T = 5:

$$y = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}; X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix}.$$

- Then,

$$X'X = \begin{bmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{bmatrix}; \ X'y = \begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix}; \ y'y = 11; \ \bar{y} = 1.$$

1) Compute $\hat{\beta}$:

$$(X'X)^{-1} = \begin{bmatrix} 17/35 & 0 & -1/7 \\ 0 & 1/10 & 0 \\ -1/7 & 0 & 1/14 \end{bmatrix}.$$

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} 17/35 & 0 & -1/7 \\ 0 & 1/10 & 0 \\ -1/7 & 0 & 1/14 \end{bmatrix} \begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix} = \begin{bmatrix} 0.571 \\ 0.7 \\ 0.214 \end{bmatrix}.$$

2) Compute s^2:

SSE = $y'y - y'X\hat{\beta}$ = 0.46

$\rightarrow s^2$ = SSE/(T-k) = 0.46/(5-3) = 0.23

3) Estimate $Cov(\hat{\beta})$:

$$s^2(X'X)^{-1} = 0.23\begin{bmatrix} 17/35 & 0 & -1/7 \\ 0 & 1/10 & 0 \\ -1/7 & 0 & 1/14 \end{bmatrix} = \begin{bmatrix} 0.112 & 0 & -0.032 \\ 0 & 0.023 & 0 \\ -0.032 & 0 & 0.016 \end{bmatrix}.$$

4) Compute SSE, SSR and SST:

- SST = $y'y - T\bar{y}^2$ = 11 - 5×(1)^2 = 6;
- SSE = $y'y - \hat{\beta}'X'y$ = 11 - $\begin{pmatrix} 0.571 & 0.7 & 0.214 \end{pmatrix}\begin{pmatrix} 5 \\ 7 \\ 13 \end{pmatrix}$ = 0.46
- SSR = SST – SSE = 5.54.

5) Compute R^2 and $\overline{R}^2$.

- R^2 = SSR/SST =5.54/6 = 0.923
- $\overline{R}^2 = 1 - \frac{T-1}{T-k}(1-R^2) = 1 - \frac{5-1}{5-3}(1-0.923) = 0.846.$

[Proofs of the General Results under SIC]

1) Some useful results:

a) Let $\varepsilon = (\varepsilon_1,...,\varepsilon_T)'$. Then, the model $y_t = x_{t\bullet}'\beta_o + \varepsilon_t$ (t = 1, … , T) can be written as $y = X\beta_o + \varepsilon$. [Be careful that ε is a vector from now on!]

b) $E(\varepsilon) = 0_{T\times 1}$, because $E(\varepsilon_t) = E_{x_{t\bullet}}[E(\varepsilon_t | x_{t\bullet})] = E_{x_{t\bullet}}(0) = 0$ for all t.

c) $E(\varepsilon\varepsilon') = E(\varepsilon\varepsilon') - E(\varepsilon)E(\varepsilon') = Cov(\varepsilon) = \sigma_o^2 I_T$, because $\text{cov}(\varepsilon_t, \varepsilon_s) = 0$ by (SIC.4) and $\text{var}(\varepsilon_t) = \sigma_o^2$ by (SIC.5).

d) Under (SIC.8), $E(X'\varepsilon) = X'E(\varepsilon) = 0_{k\times 1}$.

2) Show that $E(\hat{\beta}) = \beta_o$ and $Cov(\hat{\beta}) = \sigma_o^2 (X'X)^{-1}$.

Lemma D.1:

$$\hat{\beta} = \beta_o + (X'X)^{-1}X'\varepsilon.$$

Proof:

$$y = X\beta_o + \varepsilon.$$

$$\hat{\beta} = (X'X)^{-1}X'y = (X'X)^{-1}(X\beta_o + \varepsilon) = \beta_o + (X'X)^{-1}X'\varepsilon.$$

Theorem: (Unbiasedness)

$$E(\hat{\beta}) = \beta_o.$$

Proof:

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

Theorem:

$$Cov(\hat{\beta}) = \sigma_o^2 (X'X)^{-1}.$$

Proof:

$$\text{Cov}(\hat{\beta}) = Cov[\beta_o + (X'X)^{-1} X'\varepsilon]$$

$$= Cov[(X'X)^{-1} X'\varepsilon] = (X'X)^{-1} X' Cov(\varepsilon)[(X'X)^{-1} X']'$$

$$= (X'X)^{-1} X'(\sigma_o^2 I_T) X (X'X)^{-1} = \sigma_o^2 (X'X)^{-1} X' I_T X (X'X)^{-1}$$

$$= \sigma_o^2 (X'X)^{-1} X'X (X'X)^{-1} = \sigma_o^2 (X'X)^{-1}.$$

3) Show $E(s^2) = \sigma_o^2$.

Lemma D.2:

SSE = $e'e = y'M(X)y = \varepsilon' M(X)\varepsilon$.

Proof:

SSE $= y'M(X)y = (X\beta+\varepsilon)'M(X)(X\beta + \varepsilon) = (\beta'X'+\varepsilon')M(X)\varepsilon = \varepsilon'M(X)\varepsilon$.

Theorem:

$$E(SSE) = (T-k)\sigma_o^2.$$

Digression to Matrix Algebra:

Definition: (trace of a matrix)

$B = [b_{ij}]_{n \times n} \rightarrow tr(B) = \Sigma_{i=1}^{n} b_{ii}$ = sum of diagonals.

Lemma D.3:

For $A_{m\times n}$ and $B_{n\times m}$, tr(AB) = tr(BA).

Lemma D.4:

If B is an idempotent n×n matrix,

rank(B) = tr(B).

[Comment]

- For Lemma D.4, many econometrics books assume B to be also symmetric. But the matrix B does not have to be.
- An idempotent matrix does not have to be symmetric: For example,

$$\begin{pmatrix} 1/2 & 1 \\ 1/4 & 1/2 \end{pmatrix}; \begin{pmatrix} 1 & a \\ 0 & 0 \end{pmatrix}$$

- Theorem DA.1:

  The eigenvalues of an idempotent matrix, say B, are ones or zeros.

  <Proof> $\lambda\xi = B\xi = B^2\xi = B\lambda\xi = \lambda^2\xi$.

- Theorem DA.2:

  tr(B) = sum of the eigenvalues of B, where B is n×n.

  <Proof> $\det(\lambda I - B) = (\lambda - \lambda_1)...(\lambda - \lambda_n)$

  $$\rightarrow (b_{11} + b_{22} + ... + b_{nn})\lambda^{n-1} = (\lambda_1 + ... + \lambda_n)\lambda^{n-1}.$$

- Theorem DA.3:

  rank (B) = # of non-zero eigenvalues of B [See Greene.]

- Lemma D.4 is implied by Theorems DA.1-3.

Example:

Let A be T×k (T > k). Show that rank$[I_T\text{-}A(A'A)^{-1}A'] = T - k$.

[Solution]

rank$[I_T\text{-}A(A'A)^{-1}A']$

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

$$= tr(I_T) - tr[A(A'A)^{-1}A'] = T - tr[(A'A)^{-1}A'A]$$

$$= T - tr(I_k) = T - k.$$

End of Digression.

3) Show $E(s^2) = \sigma_o^2$:

$$E(SSE) = E(\varepsilon' M(X)\varepsilon) = E[tr\{\varepsilon' M(X)\varepsilon\}] = E[tr\{M(X)\varepsilon\varepsilon'\}]$$
$$= tr[M(X)E(\varepsilon\varepsilon')] = tr[M(X)\sigma_o^2 I_T] = \sigma_o^2 tr[M(X)]$$
$$= \sigma_o^2 tr[I_T - X(X'X)^{-1}X'] = \sigma_o^2(T-k)$$

$$\rightarrow E(s^2) = E(SSE/(T-k)) = E(SSE)/(T-k) = [\sigma_o^2(T-k)]/(T-k) = \sigma_o^2.$$

4) Show the normality of $\hat{\beta}$.

Lemma D. 5:

Let $z_{T\times 1} \sim N(\mu_{T\times 1}, \Omega_{T\times T})$. Suppose that A is a $k\times T$ nonstochastic matrix. Then, $b + Az \sim N(b + A\mu, A\Omega A')$.

Theorem: $\hat{\beta} \sim N(\beta_o, \sigma_o^2(X'X)^{-1})$

Proof:

$$\hat{\beta} = \beta_o + (X'X)^{-1}X'\varepsilon$$

$$\rightarrow \hat{\beta} \sim N(\beta_o + (X'X)^{-1}X'E(\varepsilon), (X'X)^{-1}X'Cov(\varepsilon)X(X'X)^{-1})$$
$$= N(\beta_o, \sigma_o^2(X'X)^{-1}).$$

5) Show that $\hat{\beta}$ and SSE are stochastically independent.

Lemma D.6:

Let Q be a T×T (nonstochastic) symmetric and idempotent matrix. Suppose $\varepsilon \sim N(0_{T\times 1}, \sigma_o^2 I_T)$. Then,

$$\frac{\varepsilon' Q \varepsilon}{\sigma_o^2} \sim \chi^2(r), \text{ r = tr(Q)}.$$

Proof: See Schmidt.

Lemma D.7:

Suppose that Q is a T×T (nonstochastic) symmetric and idempotent and B is a m×T nonstochastic matrix. If $\varepsilon \sim N(0_{T\times 1}, \sigma_o^2 I_T)$, Bε and ε′Qε are stochastically independent iff BQ = 0_{mxT}.

Proof: See Schmidt.

Theorem:

$$\frac{(T-k)s^2}{\sigma_o^2} = \frac{SSE}{\sigma_o^2} \sim \chi^2(T-k).$$

And, $\hat{\beta}$ and s^2 are stochastically independent.

Proof:

1) Note that $(T-k)s^2/\sigma_o^2 = SSE/\sigma_o^2 = \varepsilon' M(X)\varepsilon/\sigma_o^2$.

   Since M(X) is idempotent and symmetric and tr(M(X)) = T-k, by Lemma D.7, $\varepsilon' M(X)\varepsilon/\sigma_o^2 \sim \chi^2(T-k)$.

2) Note that $\hat{\beta} - \beta_o = (X'X)^{-1}X'\varepsilon$ (by Lemma D.1); $(T-k)s^2 = SSE = \varepsilon' M(X)\varepsilon$.

   Note that $(X'X)^{-1}X'M(X) = 0_{kxT}$. Therefore, Lemma D.7 applies, i.e., SSE and $\hat{\beta}$ are stochastically independent. So are s^2 and $\hat{\beta}$.

Theorem: $\text{var}(s^2) = 2\sigma_o^4/(T-k)$.

Proof:

Since $(T-k)s^2/\sigma_o^2 \sim \chi^2(T-k)$, $\text{var}[(T-k)s^2/\sigma_o^2] = 2(T-k)$ (since $\text{var}(\chi^2(r)) = 2r$), and $[(T-k)/\sigma_o^2]^2\,\text{var}(s^2) = 2(T-k)$ implies $\text{var}(s^2) = 2\sigma_o^4/(T-k)$.

Remark:

Let $\theta = \begin{pmatrix} \beta \\ \sigma^2 \end{pmatrix}$ and $\hat{\theta} = \begin{pmatrix} \hat{\beta} \\ s^2 \end{pmatrix}$. Then, $Cov(\hat{\theta}) = \begin{bmatrix} \sigma_o^2(X'X)^{-1} & 0_{k\times 1} \\ 0_{1\times k} & \dfrac{2\sigma_o^4}{T-k} \end{bmatrix}$.

[6] Efficiency of $\hat{\beta}$ and s^2

Question:

Are the OLS estimators, $\hat{\beta}$ and s^2, the best estimators among the unbiased estimators of β and σ^2?

Theorem: (Gauss-Markov)

Under (SIC.1) – (SIC.5) (ε may not be normal) and (SIC.8), $\hat{\beta}$ is the best linear unbiased estimator (BLUE) of β.

Comment:

Suppose that $\tilde{\beta}$ is an estimator which is linear in y; that is, there exists a $T\times k$ matrix C such that $\tilde{\beta} = C'y$. Let us assume that $E(\tilde{\beta}) = \beta_o$. Then, the above theorem means that $Cov(\tilde{\beta})$ - $Cov(\hat{\beta})$ is psd, for any $\tilde{\beta}$.

Proof of Gauss-Markov (A Sketch):

Let $\tilde{\beta}$ be an unbiased estimator linear in y: That is, there exists a T×k matrix C such that $\tilde{\beta}$ = C′y. Let $C' = (X'X)^{-1}X' + D'$. Then,

$$E(E(\tilde{\beta}) = E[(X'X)^{-1}X'y + D'y] = E(\hat{\beta} + D'y) = \beta_o + E(D'y).$$

Since $\tilde{\beta}$ is unbiased, it must be that:

$$E(D'y) = 0 \quad \rightarrow E[D'(X\beta+\varepsilon)] = 0 \rightarrow D'X\beta + D'E(\varepsilon) = 0$$

$$\rightarrow D'X\beta_o = 0.$$

Since this result must hold whatever β_o is, $D'X = 0_{k\times k}$. Then,

$$\tilde{\beta} = C'y = [(X'X)^{-1}X' + D']y = [(X'X)^{-1}X' + D'](X\beta_o + \varepsilon)$$

$$= \beta_o + [(X'X)^{-1}X' + D']\varepsilon$$

After some algebra, you can show that (do this by yourself):

$\mathrm{Cov}(\tilde{\beta}) = \mathrm{Cov}(\hat{\beta}) + \sigma_o^2 D'D$ [using the fact that $D'X = 0$].

Then, you can show:

$\mathrm{Cov}(\tilde{\beta}) - \mathrm{Cov}(\hat{\beta}) = \sigma_o^2 D'D$ is psd (by the theorem below)

Digression to Matrix Theory

Theorem:

Suppose A is p×q nonzero matrix. Then, A′A is psd. If rank(A) = q, then, A′A is pd.

End of Digression

Theorem:

Under (SIC.1) – (SIC.6) (ε should be normal) and (SIC.8), $\hat{\beta}$ and s^2 are the most efficient estimators of β and σ^2. [(SIC.7) does not have to hold.]

Digression to Mathematical Statistics

(1) Cases in which θ (unknown parameter) is scalar.

Definition: (Likelihood function)

- Let $\{x_1, \ldots , x_T\}$ be a sample from a population.
  - It does not have to be a random sample.
  - x_t is a scalar.
- Let $f(x_1,x_2, \ldots , x_T,\theta_o)$ be the joint density function of $x_1, \ldots , x_T$.
  - The functional form of f is known, but not θ_o.
- Then, $L_T(\theta) \equiv f(x_1, \ldots , x_T, \theta)$ is called "likelihood function".
  - $L_T(\theta)$ is a function of θ given $x_1, \ldots , x_T$.
  - The functional form of f is known, but not θ_o.

Definition: (log-likelihood function)

$l_T(\theta) = \ln[f(x_1, \ldots , x_T,\theta)]$.

Example:

- $\{x_1, \ldots, x_T\}$: a random sample from a population distributed with $f(x,\theta_o)$.
- $f(x_1, \ldots, x_T, \theta_o) = \prod_{t=1}^{T} f(x_t,\theta_o)$.

$\rightarrow$ $L_T(\theta) = f(x_1, \ldots, x_T, \theta) = \prod_{t=1}^{T} f(x_t,\theta)$.

$\rightarrow$ $l_T(\theta) = \ln\left(\prod_{t=1}^{T} f(x_t,\theta)\right) = \Sigma_t \ln f(x_t,\theta)$.

Definition: (Maximum Likelihood Estimator (MLE))

MLE $\hat{\theta}_{MLE}$ maximizes $l_T(\theta)$ given data points $x_1, \ldots, x_T$.

Theorem: (Minimum Variance Unbiased Estimator)

If $E(\hat{\theta}_{MLE}) = \theta_o$, then $\hat{\theta}_{MLE}$ is the MVUE. If $E(\hat{\theta}_{MLE}) \neq \theta_o$, but if there exists a function $g(\hat{\theta}_{MLE})$ such that $E[g(\hat{\theta}_{MLE})] = \theta_o$, then, $g(\hat{\theta}_{MLE})$ is the MVUE.

Example:

- $\{x_1, \ldots, x_T\}$ is a random sample from a population following a Poisson distribution [i.e., $f(x,\theta) = e^{-\theta}\theta^x/x!$ (suppressing subscript "o" from θ)].
- Note that $E(x) = var(x) = \theta_o$ for Poisson distribution.
- $l_T(\theta) = \Sigma_t \ln[f(x_t,\theta)] = -\theta T + (\ln(\theta))\Sigma_t x_t - \Sigma_t \ln(x_t!)$
- FOC of maximization: $\partial \ell_T / \partial\theta = -T + \frac{1}{\theta}\Sigma_t x_t = 0$.
- Solving this, $\hat{\theta}_{MLE} = \frac{\Sigma_t x_t}{T} = \bar{x}$.

(2) Extension to the Cases with Multiple Parameters.

Definition:

- $\theta = [\theta_1, \theta_2, \ldots, \theta_p]'$.
- $L_T(\theta) = f(x_1, \ldots, x_T, \theta) = f(x_1, \ldots, x_T, \theta_1, \ldots, \theta_p)$.
- $l_T(\theta) = \ln[f(x_1, \ldots, x_T, \theta) = \ln[f(x_1, \ldots, x_T, \theta_1, \ldots, \theta_p)]$.
    - x_t could be a vector.
    - If $\{x_1, \ldots, x_T\}$ is a random sample from a population with $f(x,\theta_o)$,

      $l_T(\theta) = \ln\left(\prod_{t=1}^{T} f(x_t,\theta)\right) = \Sigma_t \ln f(x_t,\theta)$.

Definition: (MLE)

MLE $\hat{\theta}_{MLE}$ maximizes $l_T(\theta)$ given data (vector) points $x_1, \ldots, x_T$. That is, $\hat{\theta}_{MLE}$ solves

$$\frac{\partial \ell_T(\theta)}{\partial \theta} = \begin{bmatrix} \partial \ell_T(\theta)/\partial \theta_1 \\ \partial \ell_T(\theta)/\partial \theta_2 \\ \vdots \\ \partial \ell_T(\theta)/\partial \theta_p \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{p\times 1}.$$

Theorem: (Minimum Variance Unbiased Estimator)

If $E(\hat{\theta}_{MLE}) = \theta_o$, then $\hat{\theta}_{MLE}$ is the MVUE. If $E(\hat{\theta}_{MLE}) \neq \theta_o$, but if there exists a function $g(\hat{\theta}_{MLE})$ such that $E[g(\hat{\theta}_{MLE})] = \theta_o$, then, $g(\hat{\theta}_{MLE})$ is the MVUE.

Comment:

Let $\widehat{\theta}$ be any unbiased estimator of θ_o. The above theorem implies that $[Cov(\widehat{\theta}) - Cov(\hat{\theta}_{MLE})]$ is psd.

Example:

- Let $\{x_1, \ldots, x_T\}$ be a random sample from $N(\mu_o, \sigma_o^2)$.
- Since $\{x_1, \ldots, x_T\}$ is a random sample, $E(x_t) = \mu_o$ and $\text{var}(x_t) = \sigma_o^2$.
- Let $\theta = (\mu, v)'$, where $v = \sigma^2$.
  - $f(x_t, \theta) = \dfrac{1}{\sqrt{2\pi v}} \exp\left[-\dfrac{(x_t-\mu)^2}{2v}\right] = (2\pi)^{-1/2}(v)^{-1/2} \exp\left[-\dfrac{(x_t-\mu)^2}{2v}\right]$.
  - $\ln[f(x_t,\theta)] = -\dfrac{1}{2}\ln(2\pi) - \dfrac{1}{2}\ln(v) - \dfrac{(x_t-\mu)^2}{2v}$.
  - $\ell_T(\theta) = -\dfrac{T}{2}\ln(2\pi) - \dfrac{T}{2}\ln(v) - \dfrac{\Sigma_t(x_t-\mu)^2}{2v}$.
  - MLE solves FOC:

    (1) $\dfrac{\partial \ell_T(\theta)}{\partial \mu} = -\dfrac{1}{2v}\Sigma_t 2(x_t - \mu)(-1) = \dfrac{\Sigma_t(x_t-\mu)}{v} = 0$;

    (2) $\dfrac{\partial \ell_T(\theta)}{\partial v} = -\dfrac{T}{2v} + \dfrac{\Sigma_t(x_t-\mu)^2}{2v^2} = 0$.

  - From (1):

    (3) $\Sigma_t(x_t - \mu) = 0 \rightarrow \Sigma_t x_t - T\mu = 0 \rightarrow \hat{\mu}_{MLE} = \dfrac{\Sigma_t x_t}{T} = \overline{x}$.

- Substituting (3) in to (2):

$$(4) \quad -Tv + \Sigma_t(x_t - \hat{\mu}_{MLE})^2 = 0 \rightarrow \hat{v}_{MLE} = \frac{1}{T}\Sigma_t(x_t - \bar{x})^2.$$

- Thus,

$$\hat{\theta}_{MLE} = \begin{pmatrix} \hat{\mu}_{MLE} \\ \hat{v}_{MLE} \end{pmatrix} = \begin{pmatrix} \bar{x} \\ \frac{1}{T}\Sigma_t(x_t - \bar{x})^2 \end{pmatrix}.$$

- Note that:
  - $E(\hat{\mu}_{MLE}) = E(\bar{x}) = E\left(\frac{1}{T}\Sigma_t x_t\right) = \frac{1}{T}\Sigma_t E(x_t) = \frac{1}{T}\Sigma_t \mu_o = \mu_o$.
  - $E(\hat{v}_{MLE}) = \frac{T-1}{T}\sigma_o^2$ (by the fact that $E\left[\frac{1}{T-1}\Sigma_t(x_t - \bar{x})^2\right] = \sigma_o^2$)

    $\rightarrow$ Let $g(\hat{v}_{MLE}) = \frac{T}{T-1}\hat{v}_{MLE}$.

    $\rightarrow$ Clearly, $E[g(\hat{v}_{MLE})] = E\left[\frac{1}{T-1}\Sigma_t(x_t - \bar{x})^2\right] = \sigma_o^2$.

    $\rightarrow$ Thus, $g(\hat{v}_{MLE})$ is MVUE of σ^2.

(3) Extension to Conditional density

Definition:

- Conditional density of y_t: $f(y_t, \theta_o \mid x_{t\bullet})$, $\theta = [\theta_1, \theta_2, \ldots, \theta_p]'$.
- $L_T(\theta) = \Pi_{t=1}^{T} f(y_t, \theta \mid x_{t\bullet})$.
- $l_T(\theta) = L_T(\theta) = \Sigma_{t=1}^{T} \ln(f(y_t \mid \theta, x_t))$.

Example:

- Assume that $(y_t, x_{t\bullet}')$ iid and $f(y_t, \beta_o, v_o \mid x_{t\bullet}) \sim N(x_{t\bullet}'\beta_o, v_o)$.
- $f(y_t, \beta, v \mid x_{t\bullet}) = \dfrac{1}{\sqrt{2\pi v}} \exp\left(-\dfrac{1}{2v}(y_t - x_{t\bullet}'\beta)^2 \right)$.
- $$\begin{aligned} l_T(\beta, v) &= \Sigma_t \ln f(y_t, \beta, v \mid x_{t\bullet}) \\ &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln v - \frac{1}{2v}\Sigma_t (y_t - x_t{}'\beta)^2 \\ &= -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln v - \frac{1}{2v}(y - X\beta)'(y - X\beta) \end{aligned}.$$

End of Digression

Return to Efficiency of OLS estimator

Proof:

We already know that $E(\hat{\beta}) = \beta_o$ and $E(s^2) = \sigma_o^2$. Thus, it is sufficient to show that $\hat{\beta}$ and s^2 are MLE or some functions of MLE. Under (SIC.1) – (SIC.6) and (SIC.8),

$$\varepsilon \sim N(0_{T\times 1}, v_o I_T) \rightarrow y \sim N(X\beta_o, v_o I_T), \text{ where } v_o = \sigma_o^2.$$

Therefore, we have the following likelihood function of y,

$$L_T(\beta,v) = \frac{1}{(2\pi)^{T/2}\sqrt{|vI_T|}} \exp\left[-\frac{1}{2}(y - X\beta)'(vI_T)^{-1}(y - X\beta)\right]$$

$$= \frac{1}{(2\pi)^{T/2}v^{T/2}} \exp\left[-\frac{1}{2}(y - X\beta)'(vI_T)^{-1}(y - X\beta)\right]$$

Then,

$$l_T(\beta,v) = -(T/2)\ln(2\pi) - (T/2)\ln(v) - (y-X\beta)'(y-X\beta)/(2v)$$

$$= -(T/2)\ln(2\pi) - (T/2)\ln(v) - (1/2v)[y'y - 2\beta'X'y + \beta'X'X\beta].$$

$\rightarrow$ FOC: $\partial l_T(\beta,v)/\partial\beta = -(1/2v)[-2X'y + 2X'X\beta] = 0_{k\times 1}$ (i)

$\partial l_T(\beta,v)/\partial v = -(T/2v) + (1/2v^2)(y-X\beta)'(y-X\beta) = 0$ (ii)

$\rightarrow$ From (i), $X'y - X'X\beta = 0_{k\times 1} \rightarrow \hat{\beta}_{MLE} = (X'X)^{-1}X'y = \hat{\beta}$.

$\rightarrow$ From (ii), $\hat{v}_{MLE}$ = SSE/T $\rightarrow$ s^2 is a function of $\hat{v}_{MLE}$.

[$s^2 = [T/(T-k)]\ \hat{v}_{MLE}$]

[7] Testing Linear Hypotheses

(1) Testing a single restriction on β:

- H_o: $R\beta_o - r = 0$, where R is $1\times k$ and r is a scalar.

Example: $y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + x_{t3}\beta_3 + \varepsilon_t$.

- We would like to test H_o: $\beta_{3,o} = 0$.
    - Define $R = [0\ 0\ 1]$ and $r = 0$.
    - Then, $R\beta_o - r = 0 \rightarrow \beta_{3,o} = 0$.
- H_o: $\beta_{2,o} - \beta_{3,o} = 0$ (or $\beta_{2,o} = \beta_{3,o}$).
    - Define $R = [0\ 1\ -1]$ and $r = 0$.
    - $R\beta_o - r = 0 \rightarrow \beta_{2,o} - \beta_{3,o} = 0$
- H_o: $2\beta_{2,o} + 3\beta_{3,o} = 3$.
    - $R = [0\ 2\ 3]$ and $r = 3$.
    - $R\beta - r = 0 \rightarrow H_o$.

Theorem: (T-Statistics Theorem)

Assume that (SIC.1)-(SIC.6) and (SIC.8) hold. Under H_o: $R\beta_o - r = 0$,

$$t = \frac{R\hat{\beta} - r}{s_R} \sim t(T-k),$$

where $s_R = \sqrt{R[s^2(X'X)^{-1}]R'}$.

Corollary:

Let se($\hat{\beta}_j$) = square root of the j'th diagonal of $s^2(X'X)^{-1}$. Then, under H_o: $\beta_j = \beta_j^*$,

$$t = \frac{\hat{\beta}_j - \beta_j^*}{se(\hat{\beta}_j)} \sim t(T-k) \cdot$$

Proof:

Let R = [0 0 ... 1 ... 0]; that is, only the j′th entry of R equals 1. Let $r = \beta_j^*$. Then,

$$t = \frac{R\hat{\beta} - r}{s_R} = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{Rs^2(X'X)^{-1}R'}} = \frac{\hat{\beta}_j - \beta_j^*}{\sqrt{\mathrm{var}(\hat{\beta}_j)}} = \frac{\hat{\beta}_j - \beta_j^*}{se(\hat{\beta}_j)}.$$

Comment:

- T-Statistics Theorem implies the following:
    - Imagine that you collect billions and billions (b) of different samples.
    - For each sample, compute the t statistic for the same hypothesis H_o. Denote the population of these t statistics by $\{t^{[1]}, t^{[2]}, ..., t^{[b]}\}$.
    - The above theorem indicates that the population of t-statistics is distributed as t(T-k).

How to reject or accept H_o

<Case 1> H_o: $R\beta_o = r$ and H_a: $R\beta_o \neq r$.

- For simplicity, consider a case with T-k = 25.
- H_o: $\beta_{j,o} = 0$ and H_a: $\beta_{j,o} \neq 0$.

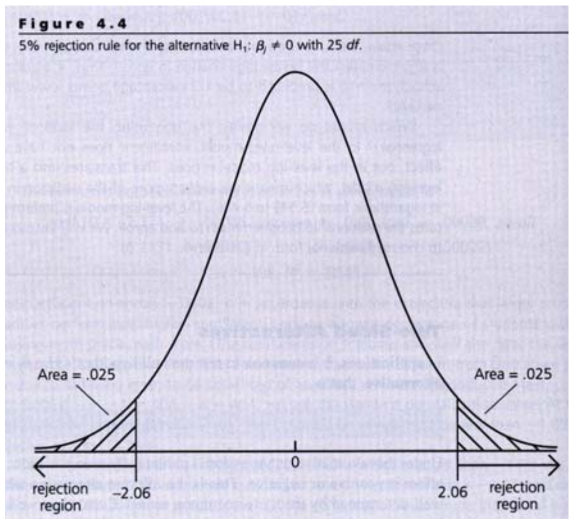
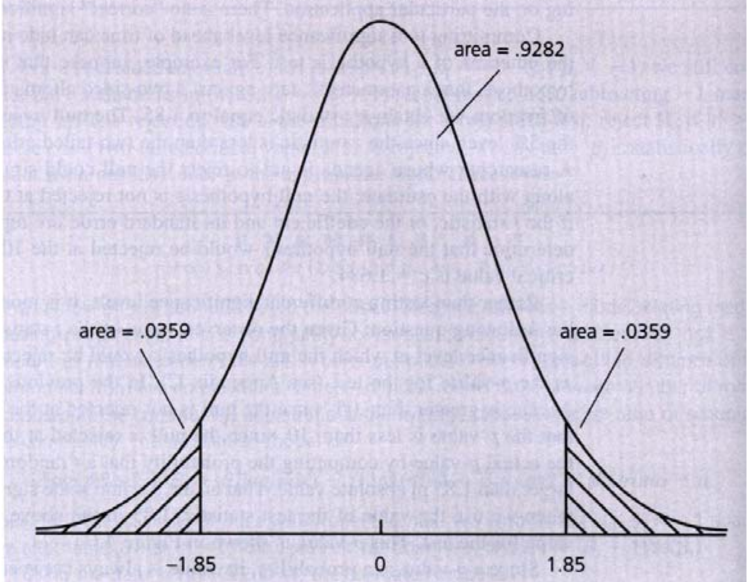


Figure 4.4
5% rejection rule for the alternative H_1: $\beta_j \neq 0$ with 25 *df*.

- If you choose α = 5% (significance level), the probability that your t-statistic computed with a sample lies between –2.06 and 2.06 is 95% (confidence level). Call 2.06 "critical value" (c).
- So, if the value of your t-statistic is outside of (-2.06, 2.06) [(-c, c)], you could say, "My t-value is quite an unlikely number I can obtain, if H_o is indeed correct". In this sense, you reject H_o.
- If the value of your t-statistic is inside of (-2.06,2.06), you can say, "My t-value is a possible number I can get if H_o is correct." In this sense, you accept (do not reject) H_o.

- Another way to determine acceptance/rejection (P-value):
    - Suppose you have t = 1.85 and T-k = 40
    - Find the probability that a t-random variable is outside of (-1.85, 1.85).

Figure 4.6
Obtaining the *p*-value against a two-sided alternative, when t = 1.85 and df = 40.



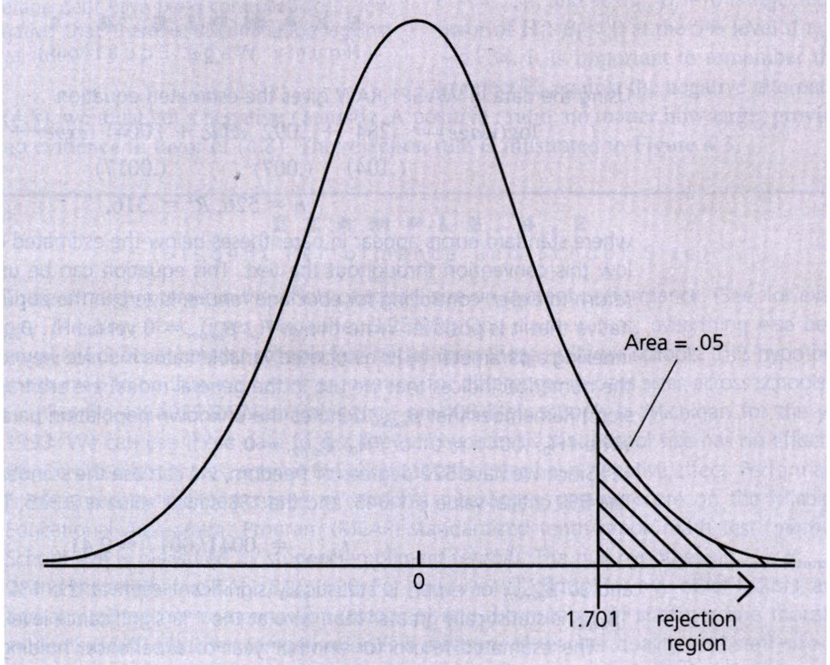
- This probability is called p-value. This value is the minimum α value with which you can reject H_o. Thus, your choice of α > p-value, reject H_o. If your choice of α < p-value, do not reject H_o.

<Case 2> H_o: $R\beta_o = r$ and H_a: $R\beta_o > r$.

- $T\text{-}k = 28$, H_o: $\beta_{j,o} = 0$ and H_a: $\beta_{j,o} > 0$.

Figure 4.2

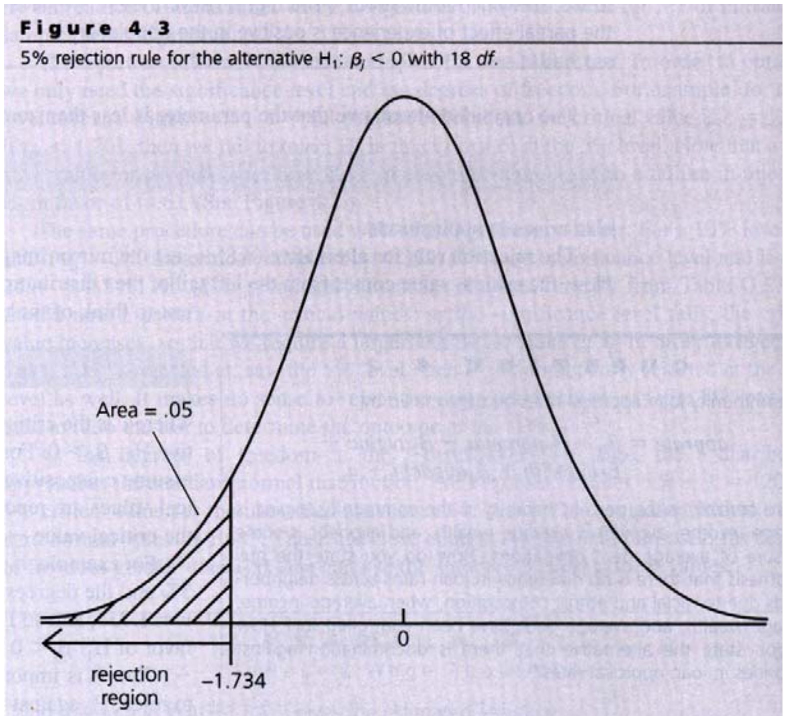
5% rejection rule for the alternative H_1: $\beta_j > 0$ with 28 *df*.



- Here, you strongly believe that $\beta_{j,o}$ cannot be negative. If so, you would regard negative t-statistics as evidence for H_o. So, your acceptance/rejection decision depends on how positively large the value of your t-statistic is.
- Choose a critical value ($c = 1.701$) as in the above graph at 5% significance level. Then, reject H_o in favor of H_a, if $t > c$ (=1.701). Do not reject H_o, if $t < c$.

<Case 3> H_o: $R\beta_o = r$ and H_a: $R\beta_o < r$.

- $T\text{-}k = 18$, H_o: $\beta_{j,o} = 0$ and H_a: $\beta_{j,o} < 0$.



- Here, you strongly believe that $\beta_{j,o}$ cannot be positive. If so, you would regard a positive value of a t-statistic as evidence favoring H_o. So, your acceptance/rejection decision depends on how negatively large the value of your t-statistic is.
- Choose a critical value ($-c = -1.734$) as in the above graph at a given significance level. Then, reject H_o in favor of H_a, if $t < -c$ ($= -1.734$). Do not reject H_o, if $t > -c$.

Numerical Example:

- Use 95% of confidence level.
- $y = \beta_1 + \beta_2 x_{2t} + \beta_3 x_{3t} + \varepsilon_t$.
- $s^2(X'X)^{-1} = \begin{bmatrix} 1.45 & 0 & 0 \\ 0 & 72.57 & -101.60 \\ 0 & -101.60 & 145.14 \end{bmatrix}$; $\hat{\beta} = \begin{bmatrix} 1.2 \\ -1 \\ 2 \end{bmatrix}$; T = 10.

- H_o: $\beta_{2,o} = \beta_{3,o}$ against H_a: $\beta_{2,o} \neq \beta_{3,o}$

  → H_o: $\beta_{2,o} - \beta_{3,o} = 0$.

  → H_o: $1 \bullet \beta_{2,o} + (-1) \bullet \beta_{3,o} = 0$.

  → R = (0,1,-1) and r = 0.

  → t = -0.14

  → df = 10 – 3 = 7 → c = 2.365

  → Since –2.365 (-c) < t < 2.365 (c), do not reject H_o.

- H_o: $\beta_{2,o} + \beta_{3,o} = 1$; H_a: $\beta_{2,o} + \beta_{3,o} \neq 1$

  → t = 0, c = 2.365.

[Proof of T-Statistics Theorem]

Digression to Probability Theory

1) Standard Normal Distribution: ($z \sim N(0,1)$)

- Pdf: $\phi(z) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{z^2}{2}\right)$, $-\infty < z < \infty$.

2) χ^2 (Chi-Square) Distribution

- Let $z_1, \ldots, z_k$ be random variables iid with $N(0,1)$.
- Then, $y = \Sigma_{i=1}^{k} z_i^2 \sim \chi^2(k)$.
- Here, $y > 0$, k = degrees of freedom.
- $E(y) = k$ and $var(y) = 2k$.

3) Student t Distribution

- Let $z \sim N(0,1)$ and $y \sim \chi^2(k)$. Assume that z and y are stochastically independent.
- Then, $t = \frac{z}{\sqrt{y/k}} \sim t(k)$.
- $E(t) = 0$, $k > 1$; $var(t) = k/(k-2)$, $k > 2$.
- As $k \to \infty$, $var(t) \to 1$. In fact, $t \to z$.
- The pdf of t is similar to that of z, but t has ticker tails.
- $f(t)$ is symmetric around $t = 0$.

4) F Distribution

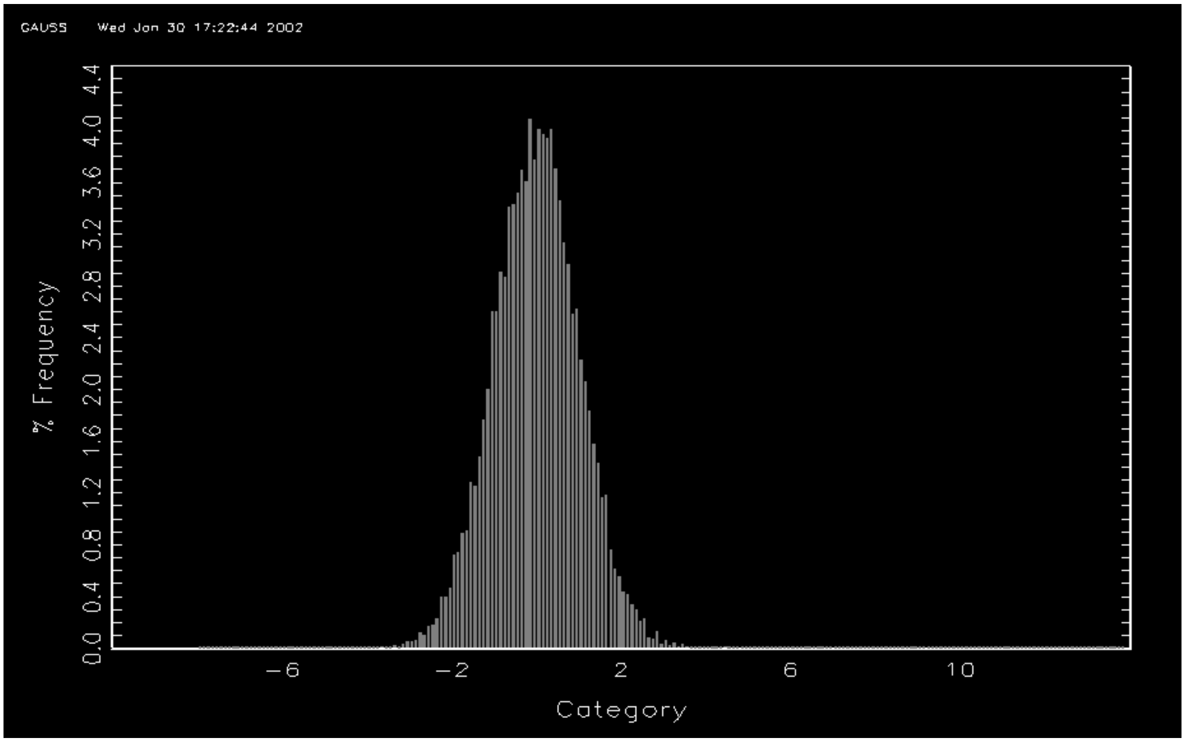
- Let $y_1 \sim \chi^2(k_1)$ and $y_2 \sim \chi^2(k_2)$ be stochastically independent.
- Then, $f = \dfrac{y_1/k_1}{y_2/k_2} \sim f(k_1,k_2)$.
- $f(1,k_2) = [t(k_2)]^2$.
- If $f \sim f(k_1,k_2)$, $k_1 f \rightarrow \chi^2(k_1)$ as $k_2 \rightarrow \infty$.

Gauss Exercise:

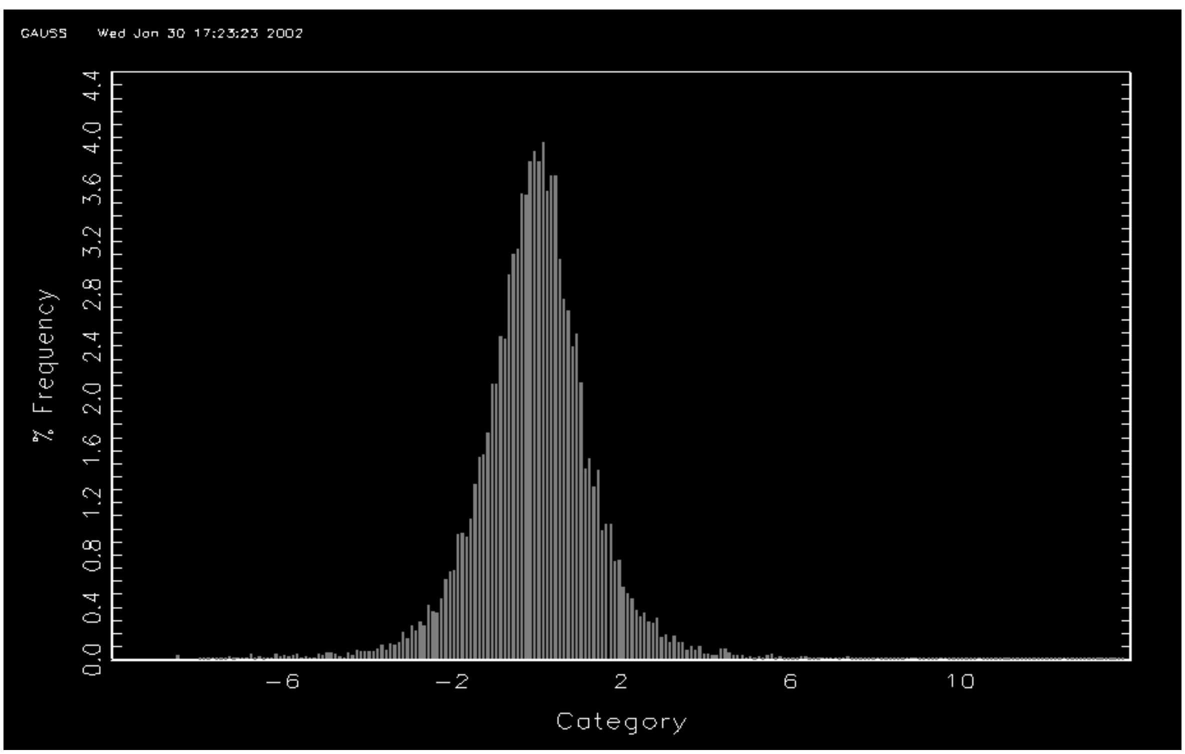
- $z \sim N(0,1)$; $t \sim t(4)$; $y \sim \chi^2(2)$; $f \sim f(2,10)$.
- Gauss program name: dismonte.prg

```
/*
** Monte Carlo Program for z, x-square, t and f distribution
*/

@ Data generation under Classical Linear Regression Assumptions @
new;
seed   = 1;
iter   = 10000; @ # of sets of different data points @

z = zeros(iter,1);
t = zeros(iter,1);
x = zeros(iter,1);
f = zeros(iter,1);

i = 1; do while i <= iter;
z[i,1] = rndns(1,1,seed);
t[i,1] = rndns(1,1,seed)./sqrt( sumc(rndns(4,1,seed)^2)/4 );
x[i,1] = sumc(rndns(2,1,seed)^2);
f[i,1] = ( sumc( rndns(2,1,seed)^2 )/2 )./ (sumc( rndns(10,1,seed)^2 )/10) ;
i = i + 1; endo ;

@ Histograms @

library pgraph;
graphset;
ytics(0,6,0.1,0) ;
v = seqa(-8,0.1,220);
@ {a1,a2,a3}=histp(z,v); @
@ {b1,b2,b3}=histp(t,v); @

library pgraph;
graphset;
ytics(0,10,0.1,0);
w = seqa(0, 0.1, 330);
```

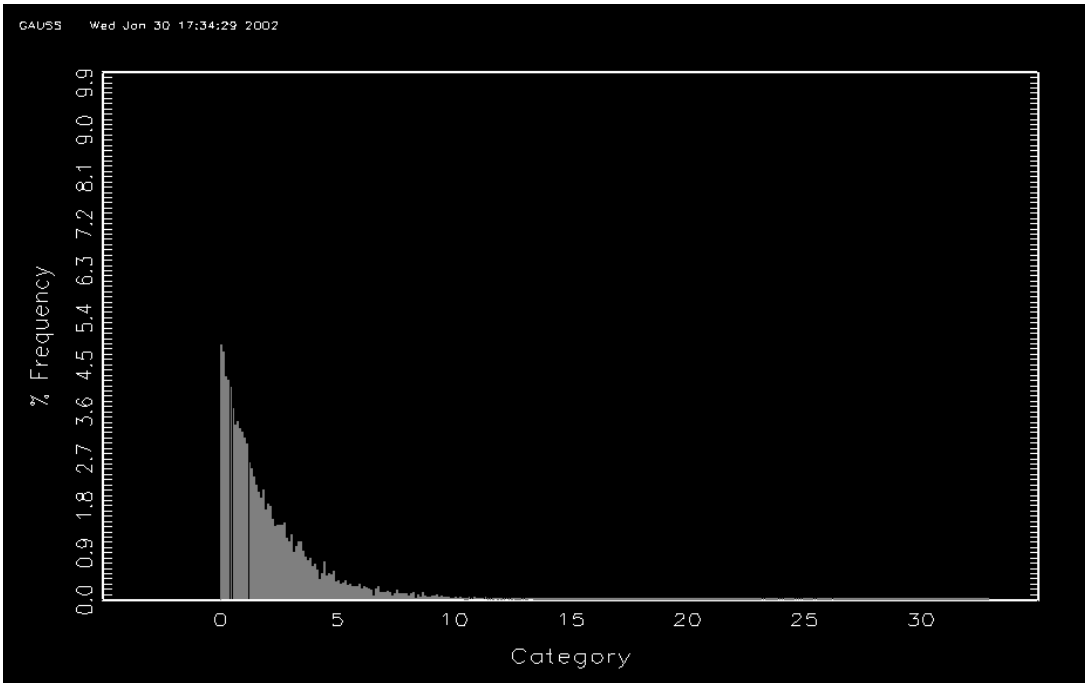
```
@ {c1,c2,c3} = histp(x,w); @
{d1,d2,d3} = histp(f,w);
```



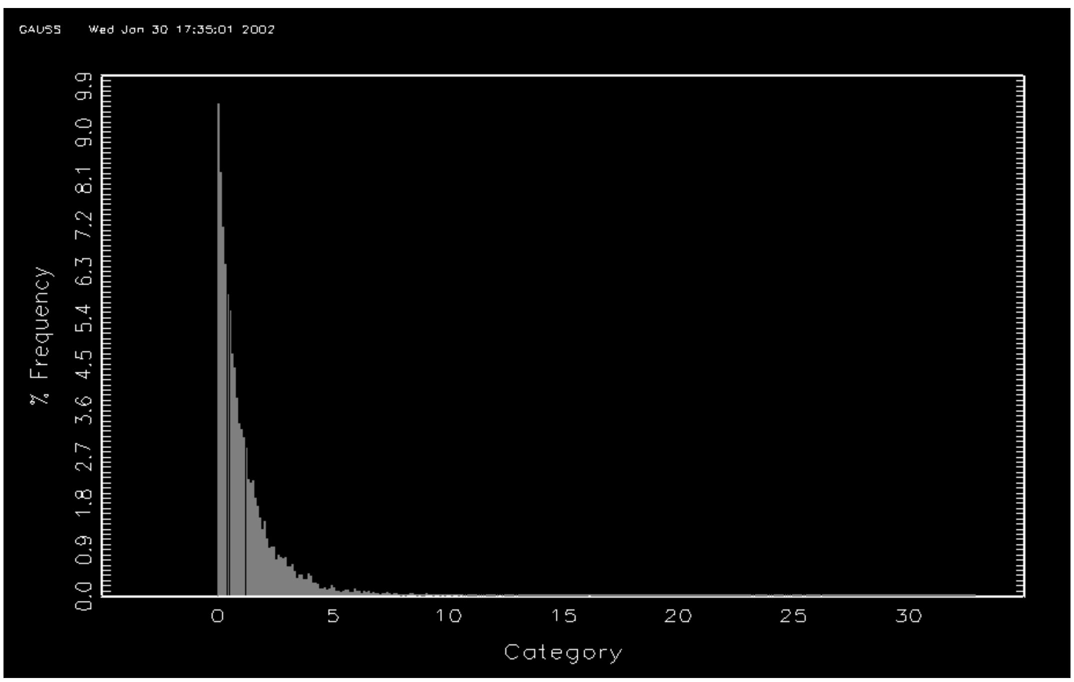
z ~ N(0,1)



t ~ t(4)



$y \sim \chi^2(2)$



$f \sim f(2,10)$

End of Digression

Lemma T.1:

Under (SIC.1)-(SIC.6) and (SIC.8), $\hat{\beta}$ and s^2 are stochastically independent. (See Schmidt.)

Lemma T.2:

Under (SIC.1)-(SIC.6) and (SIC.8),

$$\frac{R(\hat{\beta}-\beta)}{s_R} \sim t(T-k).$$

Proof:

Define $\sigma_R = \sqrt{\sigma_o^2 R(X'X)^{-1}R'}$. Note that:

$$E\left[\frac{R(\hat{\beta}-\beta)}{\sigma_R}\right] = 0; \ \text{var}\left[\frac{R(\hat{\beta}-\beta)}{\sigma_R}\right] = 1.$$

[Why?] Furthermore, since $\hat{\beta}$ is normal, so is $R(\hat{\beta}-\beta)/\sigma_R$. That is,

$$q_1 \equiv \frac{R(\hat{\beta}-\beta)}{\sigma_R} \sim N(0,1).$$

Note that

$$q_2 \equiv \frac{s_R}{\sigma_R} = \frac{\sqrt{Rs^2(X'X)^{-1}R'}}{\sqrt{R\sigma_o^2(X'X)^{-1}R'}} = \sqrt{\frac{s^2}{\sigma_o^2}} = \sqrt{\frac{(T-k)s^2}{(T-k)\sigma_o^2}} == \sqrt{\frac{\chi^2(T-k)}{T-k}}.$$

Note that q_1 and q_2 are stochastically independent because $\hat{\beta}$ and s^2 are stochastically independent by Lemma T.1. Therefore, we have:

$$\frac{R(\hat{\beta}-\beta)}{s_R} = \frac{q_1}{q_2} = \frac{N(0,1)}{\sqrt{\chi^2(T-K)/(T-k)}} \sim t(T-k).$$

Proof of T-Statistics Theorem:

Under H_o,

$$t = \frac{R\hat{\beta} - r}{s_R} = \frac{R\hat{\beta} - R\beta_o}{s_R} = \frac{R(\hat{\beta} - \beta_o)}{s_R} \sim t(T-k).$$

Then, the result immediately follows from Lemma T.2.

(2) Testing several restrictions

Assume that R is m×k and r is m×1 vector, and H_o: $R\beta_o = r$.

Example:

- A model is given: $y_t = x_{t1}\beta_{1,o} + x_{t2}\beta_{2,o} + x_{t3}\beta_{3,o} + \varepsilon_t$.
- Wish to test for H_o: $\beta_{1,o} = 0$ and $\beta_{2,o} + \beta_{3,o} = 1$.
- Define:

$$R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix};\ r = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Then, $H_o \to R\beta_o = r$.

Theorem: (F-Statistics Theorem)

Assume that all of SIC holds. Under H_o: $R\beta_o = r$,

$$F \equiv \frac{(R\hat{\beta} - r)'[Rs^2(X'X)^{-1}R']^{-1}(R\hat{\beta} - r)}{m} \sim f(m, T-k).$$

Comment:

$$\frac{(R\hat{\beta}-r)'[Rs^2(X'X)^{-1}R']^{-1}(R\hat{\beta}-r)}{m}$$
$$= \frac{R(\hat{\beta}-r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-r)/m}{SSE/(T-k)}.$$

Comment:

F-Statistics Theorem implies the following:

- Imagine that you collect billions and billions (b) of different samples.
- For each sample, compute the F statistic for the same hypothesis H_o. Denote the population of these F statistics as $\{F^{[1]}, F^{[2]}, \ldots, F^{[b]}\}$.
- The above theorem indicates that the population of the F-statistics is distributed as f(m,T-k).

How to reject or accept H_o

- When you use the F-test, it is important to note that the hypothesis you actually test is not H_o: $R\beta_o = r$. It is rather (with some exaggerations) the hypothesis that:

  $$H_o': (R\beta_o-r)'[R(X'X)^{-1}R']^{-1}(R\beta_o-r) = 0.$$

  If so, your alternative hypothesis should be that

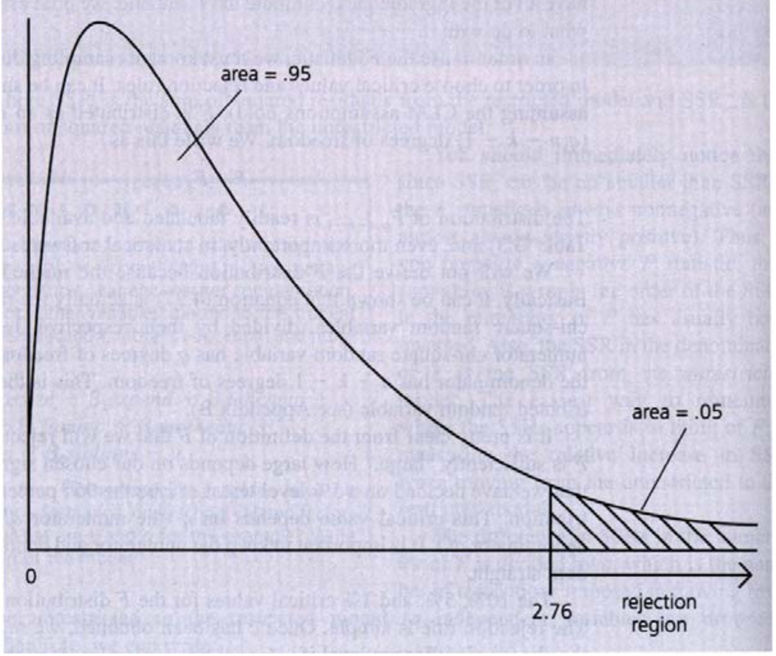
  $$H_a': (R\beta_o-r)'[R(X'X)^{-1}R']^{-1}(R\beta_o-r) > 0,$$

  because $R(X'X)^{-1}R'$ is pd. So, the F-test is a one-tail by nature.

- Suppose m = 3 and T-k = 60.

Figure 4.7

The 5% critical value and rejection region in an $F_{3,60}$ distribution.



- If you choose $\alpha = 5\%$ (significance level), the probability that your F-statistic computed with a sample is greater than 2.76 (confidence level). Call 2.76 "critical value" (c).
- So, if the value of your F-statistic is greater (smaller) than c, reject (do not reject) H_o.

An Alternative Representation of F-Statistic

Definition: (Restricted OLS)

Restricted OLS estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$ are defined as follows: $\tilde{\beta}$ minimizes $S_T(\beta) = (y-X\beta)'(y-X\beta)$ subject to the restriction $R\beta = r$. Given $\tilde{\beta}$, $\tilde{\sigma}^2$ is computed by $(y - X\tilde{\beta})'(y - X\tilde{\beta})/(T - k + m)$.

Theorem:

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'[R(X'X)^{-1}R'](R\hat{\beta} - r).$$

Proof: See Greene.

Theorem:

Under H_o: $R\beta_o - r = 0$,

$$E(\tilde{\beta}) = \beta_o.$$

$$Cov(\tilde{\beta}) = Cov(\hat{\beta}) - \sigma_o^2(X'X)^{-1}R'[R(X'X)^{-1}R']R(X'X)^{-1}.$$

Proof:

Show it by yourself. Use the fact that for any pd matrix A, BAB′ is a psd matrix whatever nonzero conformable matrix B.

Theorem:

Assume that (SIC.1)-(SIC.6) and (SIC.8) hold (whether (SIC.7) holds or not). If H_o is correct, then, $\tilde{\beta}$ is more efficient than $\hat{\beta}$.

Proof: Show it by yourself.

Theorem

Let $SSE = (y - X\hat{\beta})'(y - X\hat{\beta})$; $SSE_r = (y - X\tilde{\beta})'(y - X\tilde{\beta})$. Then,

$$F = \frac{(SSE_r - SSE)/m}{s^2} = \frac{(SSE_r - SSE)/m}{SSE/(T-k)}.$$

Proof: See Greene.

Remark:

- Consider a model: $y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + x_{t3}\beta_3 + x_{t4}\beta_4 + \varepsilon_t$.
- Wish to test for H_o: $\beta_{3,o} = \beta_{4,o} = 0$.
  - To find $\tilde{\beta}$, do OLS on:

    (*) $y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \varepsilon_t$.
  - Denote the OLS estimates by $\tilde{\beta}_1$ and $\tilde{\beta}_2$. Then, the restricted OLS estimate of β is given by $= [\tilde{\beta}_1, \tilde{\beta}_2, 0, 0]'$.
  - Also, set SSE from (*) as SSE_r.
- Test H_o: $\beta_{2,o} + \beta_{3,o} = 1$ and $\beta_{4,o} = 0$.
  - $y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + x_{t3}\beta_3 + x_{t4}\beta_4 + \varepsilon_t$.

    $\rightarrow$ $y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + x_{t3}(1-\beta_2) + \varepsilon_t$.

    $\rightarrow$ $y_t - x_{t3} = x_{t1}\beta_1 + (x_{t2}-x_{t3})\beta_2 + \varepsilon_t$. (**)
  - Do OLS on (**) and get $\tilde{\beta}_1$ and $\tilde{\beta}_2$. Set $\tilde{\beta}_3 = 1 - \tilde{\beta}_2$ and $\tilde{\beta}_4 = 0$. Set SSE_r = SSE from OLS on (**).

Theorem

Let $\tilde{\beta}_1$ be the OLS estimator β_1 for a model $y_t = \beta_1 + \varepsilon_t$. Then, $\tilde{\beta}_1 = \bar{y}$.

Proof: Do this by yourself.

Theorem: (Overall Significance F Test)

The model is given:

$$y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + ... + x_{tk}\beta_k + \varepsilon_t. \quad (*)$$

Assume that this model satisfies all of SIC (including SIC.7). Consider H_o: $\beta_{2,o} = ... = \beta_{k,o} = 0$. The F-statistic for this hypothesis is given by

$$F = \frac{T-k}{k-1}\frac{R^2}{1-R^2} \sim f(k-1,T-k),$$

where R^2 is from the original model (*).

Example:

- Consider WAGE2.WF1
- Data: (WAGE2.WF1 or WAGE2.TXT – from Wooldridge's website)

# of observations (T): 935	
1. wage	monthly earnings
2. hours	average weekly hours
3. IQ	IQ score
4. KWW	knowledge of world work score
5. educ	years of education
6. exper	years of work experience
7. tenure	years with current employer
8. age	age in years
9. married	=1 if married
10. black	=1 if black
11. south	=1 if live in south
12. urban	=1 if live in SMSA
13. sibs	number of siblings
14. brthord	birth order
15. meduc	mother's education
16. feduc	father's education
17. lwage	natural log of wage

- Estimate the Mincerian wage equation:

$$\log(\text{wage}) = \beta_1 + \beta_2\text{Educ} + \beta_3\text{Exper} + \beta_4\text{Exper}^2 + \varepsilon$$

Estimation Results by Eviews:

Dependent Variable: LWAGE
Method: Least Squares
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.517432	0.124819	44.20360	0.0000
EDUC	0.077987	0.006624	11.77291	0.0000
EXPER	0.016256	0.013540	1.200595	0.2302
EXPER^2	0.000152	0.000567	0.268133	0.7887

R-squared	0.130926	Mean dependent var	6.779004
Adjusted R-squared	0.128126	S.D. dependent var	0.421144
S.E. of regression	0.393240	Akaike info criterion	0.975474
Sum squared resid	143.9675	Schwarz criterion	0.996183
Log likelihood	-452.0343	F-statistic	46.75188
Durbin-Watson stat	1.788764	Prob(F-statistic)	0.000000

- H_o: Education does not improve individuals' productivity.
- Ha: Education matters, but its effect could be either positive or negative.

→ H_o: $\beta_{2,o} = 0$ Vs. H_a: $\beta_{2,o} \neq 0$.

→ $t = \dfrac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = 11.77291$; c = 1.96 at 5% significance level.

→ Since t $\notin$ (-1.96, 1.96), reject Ho!

→ P-value for this t statistic = 0.0000; $\alpha = 0.05$.

- H_o: Education does not improve individuals' productivity.

  H_a: Education improves individuals' productivity.

  → H_o: $\beta_{2,o} = 0$ Vs. H_a: $\beta_{2,o} > 0$.

  → $t = \dfrac{\hat{\beta}_2 - 0}{se(\hat{\beta}_2)} = 11.77291$; c = 1.645 at 5% significance level.

  Since c < t, reject H_o in favor of H_a.

- H_o: Work experience does not improve individuals' productivity.

  → H_o: $\beta_{3,o} = \beta_{4,o} = 0$.

  H_a: Work experience matters.

  → H_a: $\beta_{3,o} \neq 0$ and/or $\beta_{4,o} \neq 0$.

Wald Test:
Equation: Untitled

Null Hypothesis:	C(3)=0		
	C(4)=0		
F-statistic	17.94867	Probability	0.000000
Chi-square	35.89734	Probability	0.000000

→ F = 17.94867; c from f(2,931) = 2.6 (at α = 5%).

→ Reject H_o.

→ Or, p-val of F = 0.0000 < 0.05 = α. So, reject H_o.

Example: (Cobb-Douglas production function)

- Setup: L = labor; K = capital; Q = output.
- The Cobb-Douglas production function is given:

$$Q_t = AL_t^{\beta_2} K_t^{\beta_3} e^{\varepsilon_t},$$

  where A is constant. Taking log for both sides, we have:

$$(*)\ \log(Q_t) = \beta_1 + \beta_2 \log(L_t) + \beta_3 \log(K_t) + \varepsilon_t,$$

  where $\beta_1 = \ln(A)$.
- Estimation: Do OLS on (*), and estimate β's.
- Interpretation of β's:

  $\beta_2 = \partial \log(Q_t) / \partial \log(L_t)$ = Elasticity of output with respect to labor.

  $\beta_3 = \partial \log(Q_t) / \partial \log(K_t)$ = Elasticity of output with respect to capital.

  $\beta_2 + \beta_3$ = scale of economy (r)

  [increasing returns to scale if r > 1]
- Using F- or t-test methods, we can test H_o: $\beta_{2,o} + \beta_{3,o} = 1$.
- A drawback of Cobb-Douglas
  - When you use the Cobb-Douglas production function, you are assuming that the elasticities are constant over different levels of L and L. In reality, elasticities might change over different L and K.

Example: (Translog Production Function)

- Setup:

$$\log(Q_t) = \begin{Bmatrix} \beta_1 + \beta_2 \log(L_t) + \beta_3 \log(K_t) \\ +\beta_4 \dfrac{(\log(L_t))^2}{2} + \beta_5 \dfrac{(\log(K_t))^2}{2} + \beta_6 (\log(L_t))(\log(K_t)) + \varepsilon_t \end{Bmatrix}.$$

- Testing Cobb-Douglas:
  - Do a F-test for H_o: $\beta_{4,o} = \beta_{5,o} = \beta_{6,o} = 0$.

- Estimating elasticities:
  - Let $\overline{log(L)}$ and $\overline{\log(K)}$ be chosen values of $\log(L_t)$ and $\log(K_t)$.

    [You may choose sample means.]
  - Observe that $\eta_{QL} = \partial \log(Q) / \partial \log(L) = \beta_2 + \beta_4 \log(L) + \beta_6 \log(K)$.
  - Thus, a natural estimate of η_{QL} is given:

$$\hat{\eta}_{QL} = \hat{\beta}_2 + \hat{\beta}_4 \overline{\log(L)} + \hat{\beta}_6 \overline{\log(K)} = R\hat{\beta},$$

    where $R = (0,1,0,\overline{\log(L)},0,\overline{\log(K)})$.
  - $\text{var}(\hat{\eta}_{QL}) = \text{var}(R\hat{\beta}) = RCov(\hat{\beta})R'$.

    Thus, $se(\hat{\eta}_{QL}) = \sqrt{RCov(\hat{\beta})R'}$.

[Proofs of the theorems related with F-statistic]

Theorem:

Under H_o: $R\beta_o = r$,

$$F = \frac{(R\hat{\beta}-r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-r)/m}{SSE/(T-k)} \sim f(m,T-k).$$

Proof:

Let g = $(R\hat{\beta}-r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-r)/\sigma_o^2$; and let h = SSE/σ_o^2 = $(T-k)s^2/\sigma_o^2$. Note that F = (g/m)/[h/(T-k)]. We already know that h ~ χ^2(T-k). Therefore, we can complete the proof by showing that (i) g is χ^2(m), and that (ii) g and h are stochastically independent.

(i) Note that under H_o,

$$R\hat{\beta}-r = R\hat{\beta}-R\beta = R(\hat{\beta}-\beta) = R(X'X)^{-1}X'\varepsilon.$$

Therefore, we have:

$$g = \frac{\varepsilon'X(X'X)^{-1}R'[R(X'X)^{-1}R']^{-1}R(X'X)^{-1}X'\varepsilon}{\sigma_o^2} \equiv \frac{\varepsilon'Q\varepsilon}{\sigma_o^2}.$$

We can see that Q is symmetric and idempotent with Rank(Q) = m. Since $\varepsilon \sim N(0_{T\times 1},\sigma_o^2 I_T)$, g ~ χ^2(m). [See Schmidt.]

(ii) $h = SSE/\sigma^2 = \varepsilon'M(X)\varepsilon/\sigma^2 \sim \chi^2(T-k)$. Note that M(X)Q = 0. Therefore, g and h are stochastically independent. [See Schmidt.]

Theorem:

Under H_o: $R\beta_o - r = 0$,

$$E(\tilde{\beta}) = \beta_o) ;$$

$$Cov(\tilde{\beta}) = Cov(\hat{\beta}) - \sigma_o^2 (X'X)^{-1} R'[R(X'X)^{-1}R']^{-1} R(X'X)^{-1}.$$

Proof:

(i) $\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{\beta} - r)$

$$= \hat{\beta} - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}(R\beta_o + R(X'X)^{-1}X'\varepsilon - r)$$

$$= \beta_o + (X'X)^{-1}X'\varepsilon - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X'\varepsilon$$

$$= \beta_o + [(X'X)^{-1}X' - (X'X)^{-1}R'\left[R(X'X)^{-1}R'\right]^{-1}R(X'X)^{-1}X']\varepsilon$$

$$\rightarrow E(\tilde{\beta}) = \beta_o.$$

(ii) Derive $Cov(\tilde{\beta})$ by yourself.

Theorem (Overall Significance Test)

The model is given:

$$(*) \quad y_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \ldots + x_{tk}\beta_k + \varepsilon_t.$$

The null hypothesis is given by H_o: $\beta_{2,o} = \ldots = \beta_{k,o} = 0$. Assume that (SIC.1)-(SIC.8) (including (SIC.7)) hold. Then, the F-statistic for H_o is given by:

$$F = \frac{T-k}{k-1}\frac{R^2}{1-R^2} \sim f(k-1, T-k),$$

where R^2 is from the above-unrestricted model (*).

Proof:

The restricted model is given by: $y_t = \beta_1 + \varepsilon_t$. Since $\tilde{\beta}_1 = \bar{y}$,

$$
\begin{aligned}
\text{SSE}_r &= (y - X\tilde{\beta})'(y - X\tilde{\beta}) = \Sigma_t (y_t - x_{t\bullet}'\tilde{\beta})^2 \\
&= \Sigma_t (y_t - \tilde{\beta}_1 - x_{t2}\tilde{\beta}_2 - \ldots - x_{tk}\tilde{\beta}_k)^2 \\
&= \Sigma_t (y_t - \tilde{\beta}_1)^2 = \Sigma_t (y_t - \bar{y})^2 = \text{SST}.
\end{aligned}
$$

Observe that:

$$
\begin{aligned}
F &= \frac{(SSE_r - SSE_u)/(k-1)}{SSE_u/(T-k)} = \frac{T-k}{k-1}\frac{SST - SSE}{SSE} \\
&= \frac{T-k}{k-1}\frac{1 - SSE/SST}{SSE/SST} = \frac{T-k}{k-1}\frac{R^2}{1-R^2}.
\end{aligned}
$$

[8] Tests of Structural Changes

(1) Motivation:

Relationships among economic variables may change over time or across different genders (Ch. 7.4 in Greene)

Example 1:

Oil shocks during 70's may have changed firms' production functions permanently.

Example 2:

Effects of schooling on wages may be different over different regions. [Why? Perhaps because of different industries across different regions.]

- Data: (WAGE2.WF1 or WAGE2.TXT – from Wooldridge's website)

# of observations (T): 935

1. wage	monthly earnings
2. hours	average weekly hours
3. IQ	IQ score
4. KWW	knowledge of world work score
5. educ	years of education
6. exper	years of work experience
7. tenure	years with current employer
8. age	age in years
9. married	=1 if married
10. black	=1 if black
11. south	=1 if live in south
12. urban	=1 if live in SMSA
13. sibs	number of siblings
14. brthord	birth order
15. meduc	mother's education
16. feduc	father's education
17. lwage	natural log of wage

- Mincerian wage equation for people living in South (A):

Dependent Variable: LWAGE
Sample(adjusted): 28 935 IF SOUTH = 1
Included observations: 319 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.860469	0.233695	20.79831	0.0000
EDUC	0.101053	0.012594	8.024086	0.0000
EXPER	0.053960	0.024386	2.212751	0.0276
EXPER^2	-0.001007	0.001009	-0.997829	0.3191

R-squared	0.179628	Mean dependent var	6.665056
Adjusted R-squared	0.171815	S.D. dependent var	0.450349
S.E. of regression	0.409838	Akaike info criterion	1.066352
Sum squared resid	**52.90976**	Schwarz criterion	1.113565
Log likelihood	-166.0832	F-statistic	22.99070
Durbin-Watson stat	1.755004	Prob(F-statistic)	0.000000

- Mincerian wage equation for people living in Non-South (B):

Dependent Variable: LWAGE
Sample(adjusted): 1 910 IF SOUTH = 0
Included observations: 616 after adjusting endpoints

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.893468	0.143314	41.12270	0.0000
EDUC	0.063453	0.007563	8.389865	0.0000
EXPER	-0.002798	0.015758	-0.177542	0.8591
EXPER^2	0.000744	0.000664	1.120953	0.2627

R-squared	0.103200	Mean dependent var	6.838013
Adjusted R-squared	0.098804	S.D. dependent var	0.392769
S.E. of regression	0.372861	Akaike info criterion	0.871250
Sum squared resid	**85.08351**	Schwarz criterion	0.899973
Log likelihood	-264.3451	F-statistic	23.47553
Durbin-Watson stat	1.852473	Prob(F-statistic)	0.000000

- Question:
  - $\beta_{A1,o} = \beta_{B1,o}$, $\beta_{A2,o} = \beta_{B2,o}$, $\beta_{A3,o} = \beta_{B3,o}$ and $\beta_{A4,o} = \beta_{B4,o}$?
  - If so, we can pool all observations to estimate:

$$(C) \quad lwage_t = \beta_1 + \beta_2 educ_t + \beta_3 exper_t + \beta_4 exper_t^2 + \varepsilon_t,\ t = 1, \ldots, T.$$

Dependent Variable: LWAGE
Method: Least Squares
Date: 02/05/02 Time: 13:57
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.517432	0.124819	44.20360	0.0000
EDUC	0.077987	0.006624	11.77291	0.0000
EXPER	0.016256	0.013540	1.200595	0.2302
EXPER^2	0.000152	0.000567	0.268133	0.7887

R-squared	0.130926	Mean dependent var	6.779004
Adjusted R-squared	0.128126	S.D. dependent var	0.421144
S.E. of regression	0.393240	Akaike info criterion	0.975474
Sum squared resid	**143.9675**	Schwarz criterion	0.996183
Log likelihood	-452.0343	F-statistic	46.75188
Durbin-Watson stat	1.788764	Prob(F-statistic)	0.000000

- Question:

How can we test H_o: $\beta_{A1,o} = \beta_{B1,o}$, $\beta_{A2,o} = \beta_{B2,o}$, $\beta_{A3,o} = \beta_{B3,o}$ and $\beta_{A4,o} = \beta_{B4,o}$?

(2) General Framework

Model For Group A:

(A) $y_{At} = \beta_{A1} + \beta_{A2}x_{At2} + ... + \beta_{Ak}x_{Atk} + \varepsilon_{At}$, $t = 1, ... , T_A$.

Model For Group B:

(B) $y_{Bt} = \beta_{B1} + \beta_{B2}x_{Bt2} + ... + \beta_{Bk}x_{Btk} + \varepsilon_{Bt}$, $t = 1, ... , T_B$.

Under H_o: $\beta_{Aj,o} = \beta_{Bj,o}$ for any $j = 1, ... , k$ (k restrictions),

we can pool the data to estimate

(C) $y_t = \beta_1 + \beta_2x_{t2} + ... + \beta_kx_{tk} + \varepsilon_t$, $t = 1, ... , T$ ($= T_A+T_B$).

Assume that $\mathrm{var}(\varepsilon_{At}) = \mathrm{var}(\varepsilon_{Bt}) = \sigma_o^2$.

(3) Chow-Test Procedure.

STEP 1: Do OLS on (C) and get SSE_C.

STEP 2: Do OLS on (A) and (B); then get SSE_A and SSE_B.

STEP 3: Compute the Chow-Test statistic.

Under H_o,

$$F_{CHOW} = \frac{(SSE_C - SSE_A - SSE_B)/k}{(SSE_A + SSE_B)/(T_A + T_B - 2k)} \sim f(k, T_A + T_B - 2k).$$

Example: Back to the Mincerian wage equation.

STEP 1: OLS results from all (SSE_C = 143.9675; T_A+T_B = 935).

STEP 2: OLS results from South (SSE_A = 52.90976; T_A = 319).

OLS results from Non-South (SSE_B = 85.08351; T_B = 616).

STEP 3: Compute the Chow statistic:

$$F_{CHOW} = \frac{(SSE_C - SSE_A - SSE_B)/k}{(SSE_A + SSE_B)/(T_A + T_B - 2k)}$$

$$= \frac{(143.9675 - 85.08351 - 52.90976)/4}{(85.08351 + 52.90976)/(935 - 8)}$$

$$= 10.033299$$

c from f(4,927) = 2.37 at 5% significance level. Since F > c, we reject H_o. There is a structural difference between South and Non-South.

[Proof for Chow test]

- Assume ε_{At} and ε_{Bt} are iid $N(0,\sigma_o^2)$.
- Unrestricted Model: Merge Models (A) and (B):

  Model A: $y_A = X_A\beta_A + \varepsilon_A$

  Model B: $y_B = X_B\beta_B + \varepsilon_B$

$\rightarrow$ (*) $\begin{pmatrix} y_A \\ y_B \end{pmatrix} = \begin{pmatrix} X_A & 0_{T_A\times k} \\ 0_{T_B\times k} & X_B \end{pmatrix}\begin{pmatrix} \beta_A \\ \beta_B \end{pmatrix} + \begin{pmatrix} \varepsilon_A \\ \varepsilon_B \end{pmatrix} \rightarrow y = X_*\beta_* + \varepsilon_*$.

(# of obs (T) = T_A+T_B; # of regressors = 2k)

$\rightarrow$ OLS on (*): $\hat{\beta}_* = (X_*'X_*)^{-1}X_*'y = \begin{pmatrix} \hat{\beta}_A \\ \hat{\beta}_B \end{pmatrix}$.

$\rightarrow$ SSE from this regression = $SSE_* = SSE_A + SSE_B$ [Why?].

- Restricted model:

  $\beta_{A,o} = \beta_{B,o}$ (let us denote them by β): k restrictions.

$\rightarrow$ Merge model (A) and (B) with the restriction (Model C):

(**) $\begin{pmatrix} y_A \\ y_B \end{pmatrix} = \begin{pmatrix} X_A \\ X_B \end{pmatrix}\beta + \begin{pmatrix} \varepsilon_A \\ \varepsilon_B \end{pmatrix} \rightarrow y = X\beta + \varepsilon$

$\rightarrow$ OLS on this model (restricted OLS): $\hat{\beta} = (X'X)^{-1}X'y$.

$\rightarrow SSE_r = SSE_C$.

- F-test for $\beta_{A,o} = \beta_{B,o}$:

$$F = [(SSE_r\text{-}SSE_u)/k]/[SSE_u/(T\text{-}2k)]$$
$$= [(SSE_C\text{-}SSE_A\text{-}SSE_B)/k]/[(SSE_A+SSE_B)/(T\text{-}2k)].$$

- Chow test when $\text{var}(\varepsilon_{At}) \neq \text{var}(\varepsilon_{Bt})$.

  Under H_o: $\beta_{A,o} = \beta_{B,o}$,

  W_T (wald test) = $(\hat{\beta}_A - \hat{\beta}_B)'[s_A{}^2(X_A{}'X_A)^{-1} + s_B{}^2(X_B{}'X_B)^{-1}]^{-1}(\hat{\beta}_A - \hat{\beta}_B)$

  $\rightarrow \chi^2(k)$.

- Alternative form of Chow test [Assuming $\text{var}(\varepsilon_{At}) = \text{var}(\varepsilon_{Bt})$.]
  - Define a dummy variable:

    $d_t = 1$ if $t \in A$; $d_t = 0$ if t $t \in B$.

  - Using all T observations, build up a model:

    (*) $y_t = x_{t1}\beta_1 + ... + x_{tk}\beta_k + (d_t x_{t1})\beta_{k+1} + ... + (d_t x_{tk})\beta_{2k} + \varepsilon_t$.

  - Note that

    $y_t = x_{t1}(\beta_1+\beta_{k+1}) + ... + x_{tk}(\beta_k+\beta_{2k}) + \varepsilon_t$, for $t \in A$,

    $y_t = x_{t1}\beta_1 + ... + x_{tk}\beta_k + \varepsilon_t$, for $t \in B$.

  - If no difference between A and B, $\beta_{k+1} = ... = \beta_{2k} = 0$.

    F test for H_o: $\beta_{k+1,o} = ... = \beta_{2k,o} = 0$ using OLS on (*) = Chow test!!!

Example: Return to South V.S. Non-South

Dependent Variable: LWAGE
Method: Least Squares
Date: 02/05/02 Time: 16:01
Sample: 1 935
Included observations: 935

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	5.893468	0.148297	39.74107	0.0000
EDUC	0.063453	0.007826	8.107984	0.0000
EXPER	-0.002798	0.016306	-0.171577	0.8638
EXPER^2	0.000744	0.000687	1.083291	0.2790
SOUTH	-1.032999	0.265316	-3.893462	0.0001
SOUTH*EDUC	0.037600	0.014206	2.646802	0.0083
SOUTH*EXPER	0.056757	0.028159	2.015637	0.0441
SOUTH*EXPER^2	-0.001751	0.001172	-1.493727	0.1356

R-squared	0.166990	Mean dependent var	6.779004
Adjusted R-squared	0.160700	S.D. dependent var	0.421144
S.E. of regression	0.385824	Akaike info criterion	0.941648
Sum squared resid	137.9933	Schwarz criterion	0.983064
Log likelihood	-432.2203	F-statistic	26.54749
Durbin-Watson stat	1.825679	Prob(F-statistic)	0.000000

Wald Test:
Equation: Untitled

Null Hypo.:	C(5)=0		
	C(6)=0		
	C(7)=0		
	C(8)=0		
F-statistic	10.03332	Probability	0.000000
Chi-square	40.13328	Probability	0.000000

(4) What if $T_B < k$?

- Can't estimate β for Group B.
- Alternative test procedure (Chow predictive test):

  STEP 1: Do OLS on (C) and get SSE_C.

  STEP 2: Do OLS on (A); then get SSE_A.

  STEP 3: Compute an alternative Chow-test statistic. Under H_o,

$$F_{ACHOW} = \frac{(SSE_C - SSE_A)/T_B}{(SSE_A)/(T_A - k)} \sim f(T_B, T_A - k).$$

- What is this?

  - $$y_A = X_A\beta + \varepsilon_A \quad \textit{for Group A;}$$
    $$y_B = X_B\beta + I_{T_B}\gamma + \varepsilon_B \quad \textit{for Group B,}$$

    where $\gamma = \left(\gamma_1, ..., \gamma_{T_B}\right)'$.

  - $$\begin{pmatrix} y_{B,1} \\ y_{B,2} \\ \vdots \\ y_{B,T_B} \end{pmatrix} = \begin{pmatrix} x_{B,1\bullet}' & 1 & 0 & \dots & 0 \\ x_{B,2\bullet}' & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ x_{B,T_B\bullet}' & 0 & 0 & \dots & 1 \end{pmatrix} \begin{pmatrix} \beta \\ \gamma_1 \\ \vdots \\ \gamma_{T_B} \end{pmatrix} + \begin{pmatrix} \varepsilon_{B,1} \\ \varepsilon_{B,2} \\ \vdots \\ \varepsilon_{B,T_B} \end{pmatrix}.$$

  - $$\begin{pmatrix} y_A \\ y_B \end{pmatrix} = \begin{pmatrix} X_{T_A \times k} & 0_{T_A \times T_B} \\ X_{T_B \times k} & I_{T_B \times T_B} \end{pmatrix} \begin{pmatrix} \beta \\ \gamma_1 \\ \vdots \\ \gamma_{T_B} \end{pmatrix} + \begin{pmatrix} \varepsilon_A \\ \varepsilon_B \end{pmatrix}$$

  - SSE_A = SSE from regression of the above model.
  - F_{ACHOW} = F for H_o: $\gamma_1 = ... = \gamma_{T_B} = 0$.

[9] Forecasting

- Model: $y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + ... + \beta_k x_{tk} + \varepsilon_t$.
- Wish to predict y_0 given $x_{01}, x_{02}, ... , x_{0k}$.
  - $y_0 = x_0'\beta + \varepsilon_0$, $x_0' = (x_{01}, ..., x_{0k})$.
  - $\hat{y}_0 = x_0'\hat{\beta}$ (point forecast of y_0).

Theorem:

Under (SIC.1)-(SIC.6) and (SIC.8), $(y_0 - \hat{y}_0) \sim N(0, \sigma_o^2[1 + x_0'(X'X)^{-1}x_0])$.

Proof:

$\hat{y}_0 = x_0'\hat{\beta} = x_0'[\beta_o + (X'X)^{-1}X'\varepsilon] = x_0'\beta_o + x_0'(X'X)^{-1}X'\varepsilon$.

$y_0 = x_0'\beta_o + \varepsilon_0$.

→ $y_0 - \hat{y}_0 = \varepsilon_0 - x_0'(X'X)^{-1}X'\varepsilon$.

→ Since ε_0 and ε are normal, so is $(y_0 - \hat{y}_0)$.

→ $E(y_0 - \hat{y}_0) = 0$.

→ $var(y_0 - \hat{y}_0) = \text{var}(\varepsilon_0 - x_0'(X'X)^{-1}X'\varepsilon)$

$$= \text{var}(\varepsilon_0) + \text{var}[x_0'(X'X)^{-1}X'\varepsilon]$$

$$= \sigma_o^2 + x_0'(X'X)^{-1}X'Cov(\varepsilon)[x_0'(X'X)^{-1}X']'$$

$$= \sigma_o^2 + \sigma_o^2 x_0'(X'X)^{-1}x_0.$$

Theorem:

Under (SIC.1)-(SIC.6) and (SIC.8), $\dfrac{y_0 - \hat{y}_0}{\sqrt{s^2(1 + x_0'(X'X)^{-1}x_0)}} \sim t(T-k)$.

Implication:

Let c be a critical value for two-tail t-test given a significance level (say, 5%):

$$\Pr\left(-c < \frac{y_0 - \hat{y}_0}{se(y_0 - \hat{y}_0)} < c\right) = 0.95,$$

where $se(y_0 - \hat{y}_0) = \sqrt{s^2(1 + x_0'(X'X)^{-1}x_0)}$. This implies that:

$$\Pr(\hat{y}_0 - c \times se < y_0 < \hat{y}_0 + c \times se) = 0.95.$$

Forecasting Procedure:

STEP 1: Let $x_0' = (x_{01}, x_{02}, \ldots, x_{0k})$.

STEP 2: Compute $\hat{y}_0 = x_0'\hat{\beta}$.

STEP 3: Compute $se(y_0 - \hat{y}_0) = \sqrt{s^2(1 + x_0'(X'X)^{-1}x_0)}$.

STEP 4: From given df = T-k and confidence level, find c.

STEP 5: Compute $\Pr(\hat{y}_0 - c \times se < y_0 < \hat{y}_0 + c \times se) = 0.95$.

Numerical Example:

$$(X'X)^{-1} = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 5 & -7 \\ 0 & -7 & 10 \end{pmatrix}; \hat{\beta} = \begin{pmatrix} 1.2 \\ -1 \\ 2 \end{pmatrix}; T = 10; s^2 = 14.514.$$

And $x_{02} = 1$ and $x_{03} = 1$.

STEP 1: Let $x_0' = (1, x_{02}, x_{03}) = (1,1,1)$.

STEP 2: Compute $\hat{y}_0 = x_0'\hat{\beta} = \begin{pmatrix} 1 & 1 & 1 \end{pmatrix}\begin{pmatrix} 1.2 \\ -1 \\ 2 \end{pmatrix} = 2.2$.

STEP 3: Compute se = $\sqrt{s^2(1 + x_0'(X'X)^{-1}x_0)}$ = $\sqrt{14.514 \times (1 + 1.1)}$ = 5.52.

STEP 4: From given df = 10-3 = 7 and α = 5%, c = 2.365.

STEP 5: $\hat{y}_0$- c×se = 2.2 - 2.365×5.52 = -10.855.

$\hat{y}_0$+c×se = 2.2 + 2.365×5.52 = 15.255.

Pr(-10.855 < y_o < 15.255) = 0.95.

“Dynamic” and “Static” Forecasts in Eviews

- For the analysis of cross-section data, they are the same.
- For the analysis of time-series data, they could be different.
- When a regression model uses lagged dependent variables as regressors, it is called a dynamic model.
  - Consider a simple dynamic model $y_t = \beta_1 + \beta_2 y_{t-1} + \varepsilon_t$.
  - “Dynamic” Forecast [Multiple Period Forecast]: Suppose you estimate β’s using observations up to t = 100. Using the estimates, you would like to forecast y_{101} and y_{102}. For this case, if you use “dynamic forecast”, Eviews will compute point forecasts of y_{101} and y_{102} by

    $$\hat{y}_{101} = \hat{\beta}_1 + \hat{\beta}_2 y_{100}\,;\ \hat{y}_{102} = \hat{\beta}_1 + \hat{\beta}_2 \hat{y}_{101}\,.$$

  - “Static” Forecast [One Period Forecast]: If you choose “static forecast”, Eviews will compute point forecasts of y_{101} and y_{102} by

    $$\hat{y}_{101} = \hat{\beta}_1 + \hat{\beta}_2 y_{100}\,;\ \hat{y}_{102} = \hat{\beta}_1 + \hat{\beta}_2 y_{101}\,.$$

    Observe that “static forecast” use y_{101} instead of $\hat{y}_{101}$ to forecast y_{102}.

- If you have data points up to t = 100, and if you would like to forecast y at t = 101 and t = 102, you’d better to use “dynamic forecast.”
- The formula of forecasting standard errors taught in the class can be used for static forecasts. But the standard errors for dynamic forecasts are much more complicated.

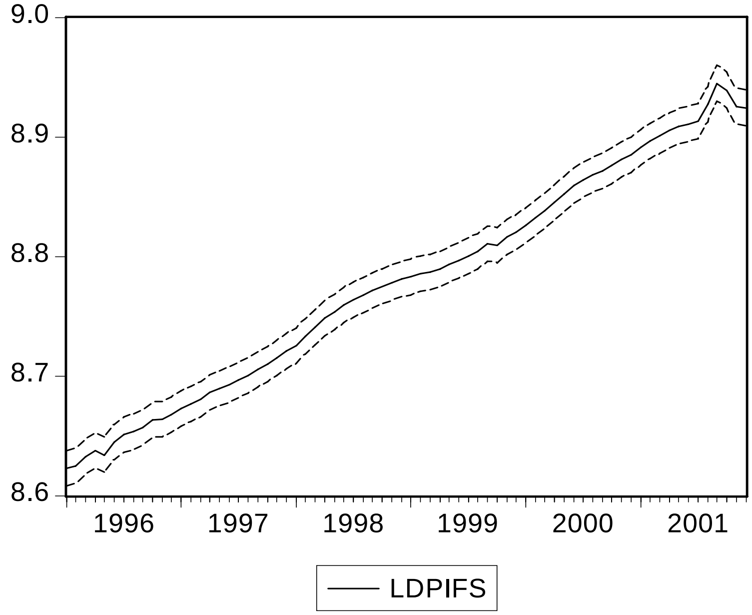
[Exercise for Static Forecast]

- Use ECN2002.wf1 (data from 1959:1 to 1995:12).
- For the definitions of the variables, see ECN2002.XLS.
  - Forecasting ldpi = log(DPI) using regression results from 1959:1 to 1995:12.

Dependent Variable: LDPI
Method: Least Squares
Date: 02/07/02 Time: 11:31
Sample(adjusted): 1959:07 1995:12
Included observations: 438 after adjusting endpoints

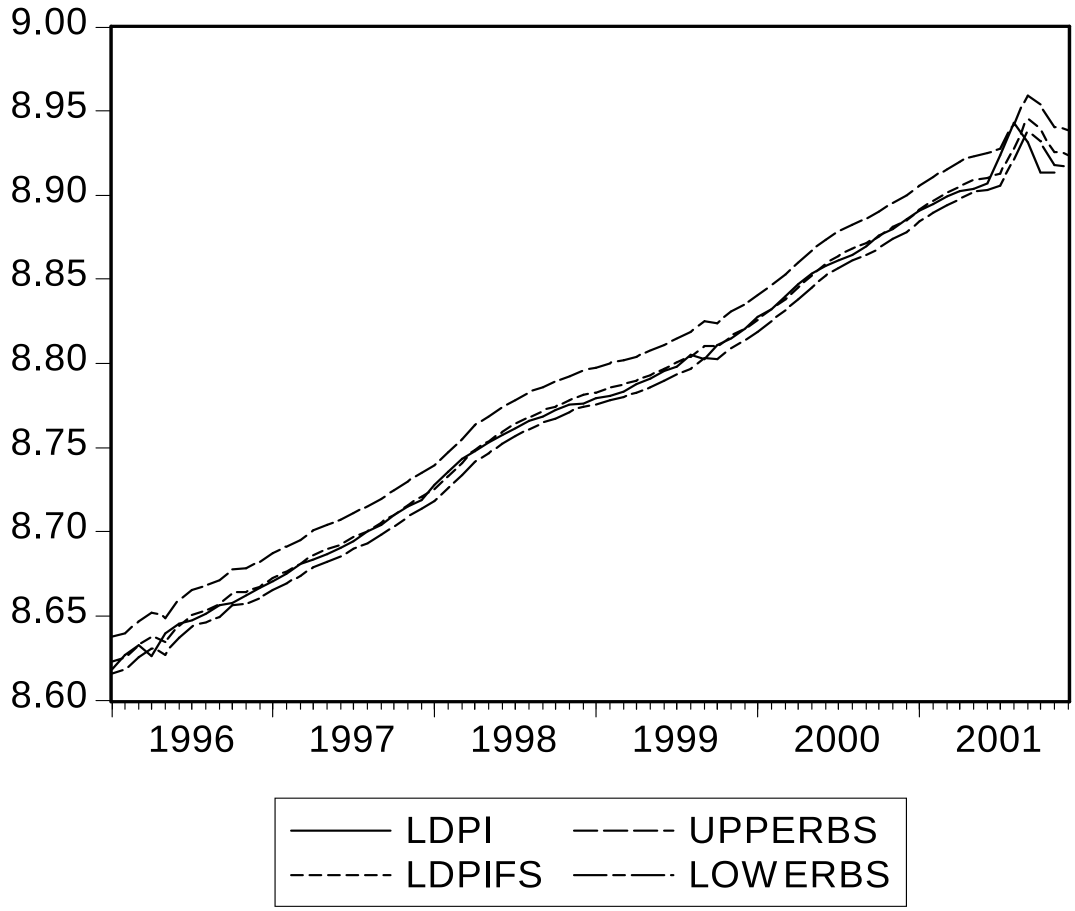
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.008851	0.003062	2.890236	0.0040
LDPI(-1)	0.802184	0.047680	16.82446	0.0000
LDPI(-2)	0.130495	0.061254	2.130386	0.0337
LDPI(-3)	0.086545	0.061535	1.406419	0.1603
LDPI(-4)	0.045344	0.061534	0.736894	0.4616
LDPI(-5)	0.078119	0.061248	1.275461	0.2028
LDPI(-6)	-0.143010	0.047695	-2.998423	0.0029

R-squared	0.999933	Mean dependent var	7.280527
Adjusted R-squared	0.999932	S.D. dependent var	0.889422
S.E. of regression	0.007340	Akaike info criterion	-6.97510
Sum squared resid	0.023220	Schwarz criterion	-6.90986
Log likelihood	1534.547	F-statistic	1069361.
Durbin-Watson stat	2.014603	Prob(F-statistic)	0.000000

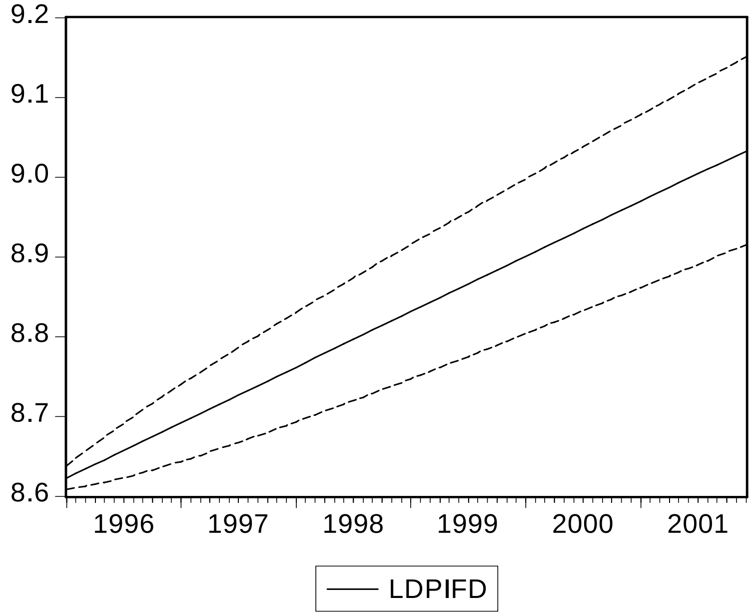


Forecast: LDPIFS
Actual: LDPI
Forecast sample: 1996:01 2001:12
Included observations: 71

Root Mean Squared Error	0.005262
Mean Absolute Error	0.003230
Mean Abs. Percent Error	0.036666
Theil Inequality Coefficient	0.000300
Bias Proportion	0.106970
Variance Proportion	0.005447
Covariance Proportion	0.887582

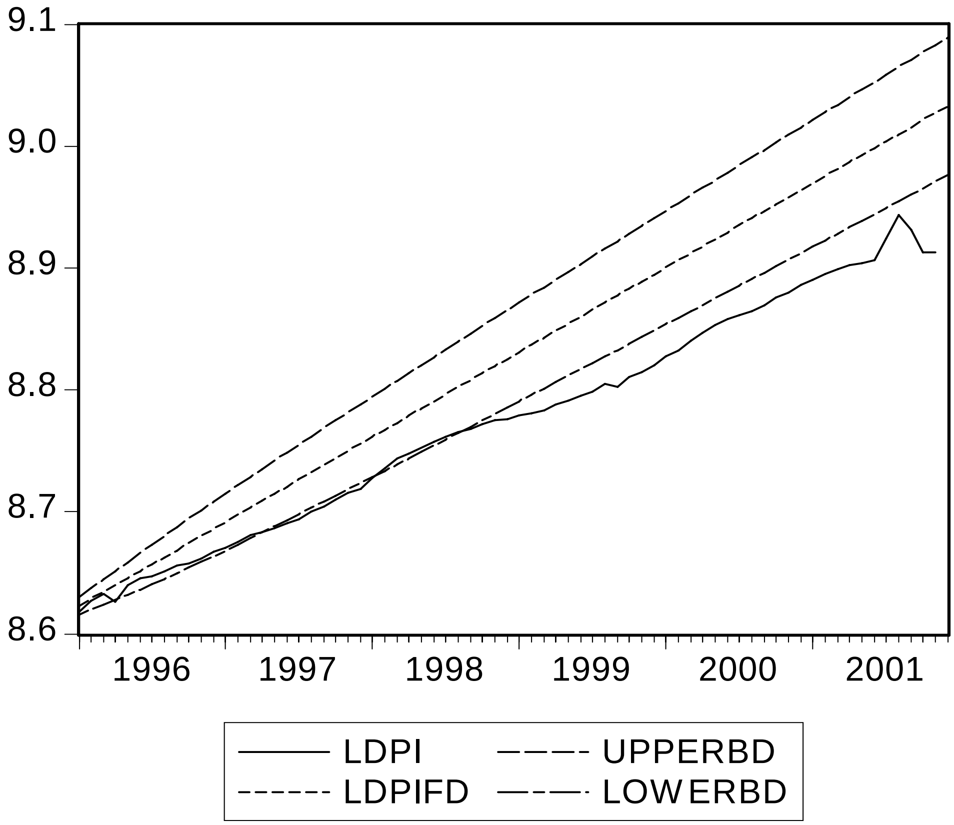


[Exercise for Dynamic Forecast]



Forecast: LDPIFD
Actual: LDPI
Forecast sample: 1996:01 2001:12
Included observations: 71

Root Mean Squared Error	0.057155
Mean Absolute Error	0.049899
Mean Abs. Percent Error	0.565546
Theil Inequality Coefficient	0.003247
Bias Proportion	0.762216
Variance Proportion	0.221227
Covariance Proportion	0.016557



[10] Nonnormal ε and Stochastic Regressors

(1) Motivation

- If the regressors $x_{t\bullet}$ are stochastic, all t and F tests are wrong (bad news).
  - The t and F tests require the OLS estimator $\hat{\beta}$ to be unbiased.
  - Recall how we have shown the unbiasedness of $\hat{\beta}$ under (SIC.8):

$$\hat{\beta} = \beta_o + (X'X)^{-1}X'\varepsilon$$

$$\rightarrow E(\hat{\beta}) = \beta + E[(X'X)^{-1}X'\varepsilon] \overset{?}{=} \beta + (X'X)^{-1}X'E(\varepsilon).$$

  - Unbiasedness of $\widehat{\beta}$ does not require nonstochastic regressors. It only requires:

$$E\left(\varepsilon_t \mid x_{1\bullet},...,x_{T\bullet}\right) = 0, \text{ for all t.} \qquad (*)$$

Or $E(\varepsilon \mid X) = 0_{T\times 1}$.

Under this assumption,

$$E(\widehat{\beta}) = E_X\left(E(\widehat{\beta} \mid X)\right) = E_X\left(E\left(\beta + (X'X)^{-1}X'\varepsilon \mid X\right)\right)$$
$$= E_X\left(\beta + (X'X)^{-1}X'E(\varepsilon \mid X)\right) = E_X(\beta) = \beta.$$

  - But, for some cases, condition (*) does not hold. For example, $x_{t\bullet} = y_{t-1}$. In this case, $E(\varepsilon_{t-1}|y_{t-1}) \neq 0$. For this case, we can no longer say that $\widehat{\beta}$ is an unbiased estimator.
  - An example for models with lagged dependent variables as regressors: $y_t = \beta_1 x_{t1} + \beta_2 x_{t2} + \beta_3 y_{t-1} + \varepsilon_t$. $\rightarrow \beta_2/(1-\beta_3)$ = long-run effect of x_{t2}.

- If the ε_t are not normally distributed, all t and F tests are wrong (bad news).
  - Can we use them if T is large?
  - Recall that the t and F statistics follow t and f distributions, respectively, only if $\hat{\beta}$ is normally distributed. But if the ε_t are not normally distributed, $\hat{\beta}$ is no longer normal.

Digression to Mathematical Statistics

Large-Sample Theories

1. Motivation:
   - $\hat{\theta}_T$: An estimator from a sample of size T, $\{x_1, \ldots , x_T\}$.

     I use subscript "T" to emphasize the fact that an estimator is a function of sample size T.
   - What would be the statistic properties of $\hat{\theta}_T$ when T is infinitely large?
   - What do we wish?

     [We wish the distribution of $\hat{\theta}_T$ would become more condensed around θ_o as T increases.]

2. Main Points:

Rough Definition of Consistency

- Suppose that the distribution of $\hat{\theta}_T$ becomes more and more condensed around θ_o as T increases. Then, we say that $\hat{\theta}_T$ is a consistent estimator. And we use the following notation:

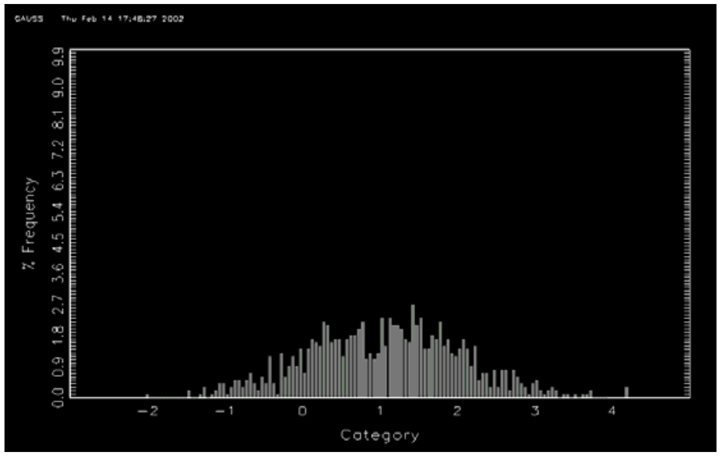
$$\text{plim}_{T\to\infty} \hat{\theta}_T = \theta_o \text{ (or } \hat{\theta}_T \to_p \theta_o).$$

- The law of large numbers (LLN) says that a sample mean $\bar{x}_T$ ($\bar{x}$ from a sample size equal to T) is a consistent estimator of μ_o. What does it mean?

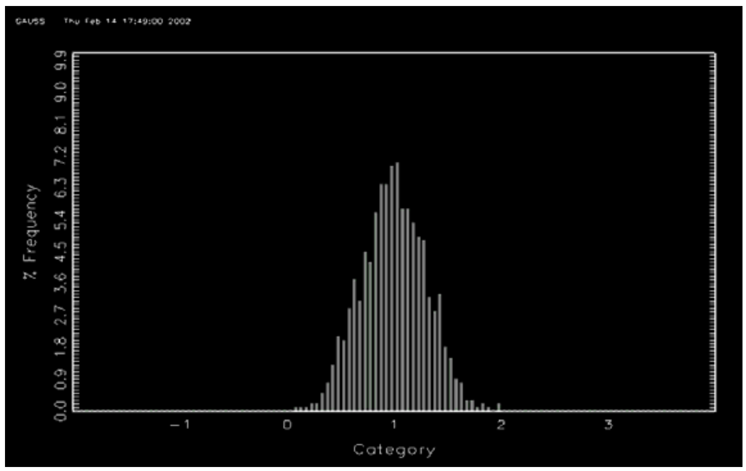
- Gauss Exercise:
  - A population with N(1,9).
  - 1000 different random samples of T = 10 to compute $\bar{x}_{10}$.
  - 1000 different random samples of T = 100 to compute $\bar{x}_{100}$.
  - 1000 different random samples of T =5000 to compute $\bar{x}_{5000}$.

- conmonte.prg

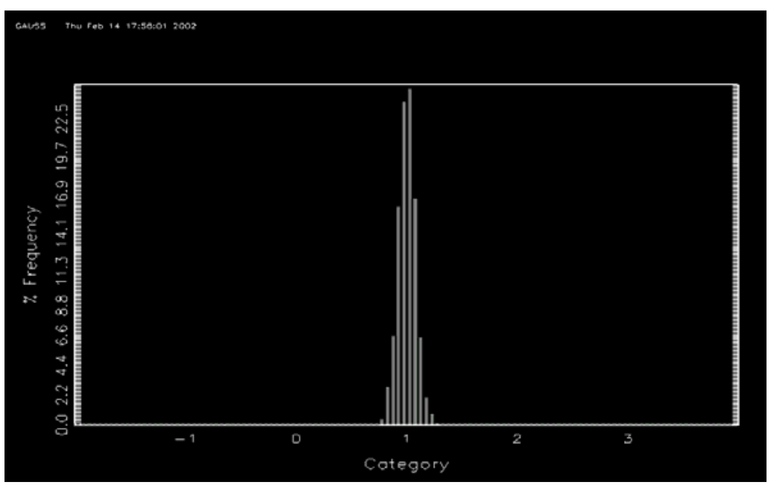
```
/*
** Monte Carlo Program to Demonstrate Efficiency of Sample Mean
*/

@ Data generation from N(1,9) @

seed   = 1;
tt1   = 10; @ # of observations @
tt2   = 100; @ # of observations @
tt3  = 1500; @ # of observations @
iter   = 1000; @ # of sets of different data @

storx10     = zeros(iter,1) ;
storx100   = zeros(iter,1) ;
storx5000 = zeros(iter,1);

i = 1; do while i <= iter;

@ compute sample mean for each sample @

x10     =  1 + 3*rndns(tt1,1,seed);
x100   =  1 + 3*rndns(tt2,1,seed);
x5000 =  1 + 3*rndns(tt3,1,seed);
storx10[i,1]    = meanc(x10);
storx100[i,1]  = meanc(x100);
storx5000[i,1]  = meanc(x5000);

i = i + 1; endo;

@ Reporting Monte Carlo results @

library pgraph;
graphset;

v = seqa(-2, .05, 120);
ytics(0,25,0.1,0);
@ {a1,a2,a3}=histp(storx10,v); @
@ {b1,b2,b3}=histp(storx100,v); @
            {b1,b2,b3}=histp(storx5000,v);
```



$\overline{x}_{10}$



$\overline{x}_{100}$



$\overline{x}_{5000}$

- Relation between unbiasedness and consistency:
- Biased estimators could be consistent.

  Example: Suppose that $\tilde{\theta}_T$ is unbiased and consistent.

  Define $\hat{\theta}_T = \tilde{\theta}_T + 1/T$.

  Clearly, $E(\hat{\theta}_T) = \theta_o + 1/T \neq \theta_o$ (biased).

  But, $\text{plim}_{T\to\infty} \hat{\theta}_T = \text{plim}_{T\to\infty} \tilde{\theta}_T = \theta_o$ (consistent).

- A unbiased estimator $\hat{\theta}_T$ is consistent if $\text{var}(\hat{\theta}_T) \to 0$ as $T \to 4$.

  Example: Suppose that $\{x_1, ..., x_T\}$ is a random sample from $N(\mu_o, \sigma_o^2)$.

  $E(\bar{x}_T) = \mu_o$.

  $\text{var}(\bar{x}_T) = \sigma_o^2/T \to 0$ as $T \to \infty$.

  Thus, $\bar{x}_T$ is a consistent estimator of μ_o.

Law of Large Numbers (LLN)

Case of scalar random variables

- Komogorov's Strong LLN:

  Suppose that $\{x_1, \ldots, x_T\}$ is a random sample from a population with finite μ and σ^2. Then, $\mathrm{plim}_{T\to\infty}\ \bar{x}_T = \mu_o$.

- Generalized Weak LLN (GWLLN):
  - $\{x_1, \ldots, x_T\}$ is a sample (not necessarily a random sample)
  - Define $E(x_1) = \mu_{1,o}, \ldots, E(x_T) = \mu_{T,o}$.
  - The variances of the x_t ($t = 1, \ldots, T$) are finite and may be different over different t.
  - Then, under suitable assumptions, $\mathrm{plim}_{T\to\infty}\ \bar{x}_T = \lim_{T\to\infty} \frac{1}{T}\Sigma_t \mu_{o,t}$.

Case of Vector Random Variables

- GWLLN
  - x_t: $p\times 1$ random vector.
  - $\{x_1, \ldots, x_T\}$ is a sample.
  - Let $E(x_1) = \mu_{1,o}$ ($p\times 1$), $\ldots$, $E(x_T) = \mu_{T,o}$.
  - Assume that $\mathrm{Cov}(x_j)$ are well-defined and finite.
  - Then, under suitable assumptions. $\mathrm{plim}_{T\to\infty}\ \bar{x}_T = \lim_{T\to\infty} \frac{1}{T}\Sigma_t \mu_{o,t}$.

Central Limit Theorems (CLT) –Asymptotic Normality

Case of scalar random variables

- Motivation:
  - Suppose that $\{x_1, \ldots, x_T\}$ is a random sample from a population with finite μ and σ^2.
  - We know $\bar{x}_T \to \mu_o$ as $T \to \infty$. But we can never have an infinitely large sample!!!
  - For finite T, $\bar{x}_T$ is still a random variable. What statistical distribution could approximate the true distribution of $\bar{x}_T$?
  - Lindberg-Levy CLT:
    - Suppose that $\{x_1, \ldots, x_T\}$ is a random sample from a population with finite μ and σ^2.
    - Then, $\sqrt{T}(\bar{x}_T - \mu_o) \to_d N(0,\sigma_o^2)$ and $\sqrt{T}\,\frac{\bar{x}_T - \mu_o}{\sigma_o} \to_d N(0,1)$.
  - Implication of CLT:
    - $\sqrt{T}(\bar{x}_T - \mu_o) \approx N(0,\sigma_o^2)$, if T is large.
    - $E\left[\sqrt{T}(\bar{x}_T - \mu_o)\right] = \sqrt{T}[E(\bar{x}_T) - \mu_o] \approx 0 \to$ $E(E(\bar{x}_T) \approx \mu_o$.
    - $\mathrm{var}\left[\sqrt{T}(\bar{x}_T - \mu_o)\right] = T \bullet \mathrm{var}(\bar{x}_T - \mu_o) = T \bullet \mathrm{var}(\bar{x}_T) \approx \sigma_0^2$

      $\to$ $\mathrm{var}(\bar{x}_T) \approx \sigma_0^2 / T$.
    - $\bar{x}_T \approx N(\mu_o, \sigma_o^2 / T)$, if T is large.

Case of random vectors

- GCLT
  - $\{y_1, \ldots, y_T\}$: a sequence of $p\times 1$ random vectors.
  - For any t, $E(y_t) = 0_{p\times 1}$ and $Cov(y_t)$ is well defined and finite.
  - Under some suitable conditions (acceptable for Econometrics I, II),

$$\frac{1}{\sqrt{T}}\Sigma_t y_t \rightarrow_d N\left(0_{p\times 1}, \lim_{T\rightarrow\infty}\frac{1}{T}Cov(\Sigma_t y_t)\right)$$

  - Note:
    - $Cov(y_t)$ [$var(y_t)$ if y_t is a scalar] could differ across different t.
    - The y_t could be correlated as long as $\lim_{n\rightarrow\infty}cov(y_t,y_{t+n}) = 0$ (ergodic).
    - If $E(y_t \mid y_{t-1}, y_{t-2}, \ldots) = 0$ (Martingale Difference Sequence), the y_t's are linearly uncorrelated. Then,

$$\frac{1}{\sqrt{T}}\Sigma_t y_t \rightarrow_d N\left(0_{p\times 1}, \lim_{T\rightarrow\infty}\frac{1}{T}\Sigma_t Cov(y_t)\right).$$

End of Digression

(2) Weak Ideal Conditions (WIC)

Consider the following linear regression model:

$$y_t = x_{t\bullet}'\beta + \varepsilon_t = x_{t1}\beta_1 + x_{t2}\beta_2 + \ldots + x_{tk}\beta_k + \varepsilon_t .$$

(WIC.1) The conditional mean of y_t (dependent variable) given $x_{1\bullet}$, $x_{2\bullet}$, ... , $x_{t\bullet}$, ε_1, ... , and ε_{t-1} is linear in $x_{t\bullet}$:

$$y_t = E(y_t \mid x_{1\bullet},\ldots,x_{t\bullet},\varepsilon_1,\ldots,\varepsilon_{t-1}) + \varepsilon_t = x_{t\bullet}'\beta_o + \varepsilon_t .$$

Comment:

- Implies $E(\varepsilon_t \mid x_{1\bullet}, x_{2\bullet},\ldots,x_{t\bullet},\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_{t-1}) = 0$.
- No autocorrelation in the ε_t: $\mathrm{cov}(\varepsilon_t,\varepsilon_s) = 0$ for all $t \neq s$.
- Regressors are weakly exogenous and need not be strictly exogenous.
- $E(x_{s\bullet}\varepsilon_t) = 0_{k\times 1}$ for all $t \geq s$, but could be that $E(x_{s\bullet}\varepsilon_t) \neq 0$ for some $s > t$.

(WIC.2) β_o is unique.

(WIC.3) The series $\{x_{t\bullet}\}$ are covariance-stationary and ergodic.

Comment:

- (WIC.2)-(WIC.3) implies that

  $p\lim_{T\to\infty} T^{-1}X'X = p\lim_{T\to\infty} T^{-1}\Sigma_t x_{t\bullet}x_{t\bullet}' \equiv Q_o$ is finite and pd.

- $Q_o = \lim_{T\to\infty} T^{-1}\Sigma E(x_{t\bullet}x_{t\bullet}')$ [By GWLLN].
- Rules out perfect multicollinearity among regressors.

(WIC.4) The data need not be a random sample.

(WIC.5) $\mathrm{var}(\varepsilon_t \mid x_{1\bullet}, x_{2\bullet}, ..., x_{t\bullet}, \varepsilon_1, ..., \varepsilon_{t-1}) = \sigma_o^2$ for all t.

(No-Heteroskedasticity Assumption).

(WIC.6) The error terms ε_t are normally distributed conditionally on $x_{1\bullet}$, … , $x_{t\bullet}$, ε_1, … , ε_{t-1}.

(WIC.7) $x_{t1} = 1$, for all t = 1, ... , T.

Comment:

SIC → WIC.

(3) Statistical Properties of the OLS estimator under WIC:

Theorem (Consistency/Asymptotic Normality Theorem):

Under (WIC.1)-(WIC.5),

$p\lim_{T\to\infty} \hat{\beta} = \beta_o$ (consistent).

$p\lim_{T\to\infty} s^2 = \sigma_o^2$ (consistent).

$\sqrt{T}(\hat{\beta} - \beta_o) \to_d N\left(0_{k\times 1}, \sigma_o^2 Q_o^{-1}\right)$.

Implication:

$$\hat{\beta} \approx N\left(\beta_o, \sigma_o^2 (TQ_o)^{-1}\right) \to \hat{\beta} \approx N(\beta_o, s^2(X'X)^{-1}),$$

if T is reasonably large.

Implication:

1) t test for H_o: $R\beta_o - r = 0$ (R: $1\times k$, r: scalar) is valid if T is large.
   Use z-table to find critical value.
2) For H_o: $R\beta_o - r = 0$ (R: $m\times k$, r: $m\times 1$),
   use $W_T = mF$ which is asymptotically $\chi^2(m)$ distributed. [Why?]

- $W_T = (R\hat{\beta} - r)'\left[R Cov(\hat{\beta}) R'\right]^{-1}(R\hat{\beta} - r)$

$$= (R\hat{\beta} - r)'\left[R s^2 (X'X)^{-1} R'\right]^{-1}(R\hat{\beta} - r) = mF.$$

Theorem (Efficiency Theorem):

Under (WIC.1)-(WIC.6), the OLS estimators are efficient asymptotically.

(4) Testing Nonlinear restrictions:

General form of hypotheses:

- Let $w(\theta) = [w_1(\theta), w_2(\theta), \ldots, w_m(\theta)]'$, where $w_j(\theta) = w_j(\theta_1, \theta_2, \ldots, \theta_p)$ = a function of $\theta_1, \ldots, \theta_p$.
- H_o: The true θ (θ_o) satisfies the m restrictions, $w(\theta) = 0_{m\times 1}$ ($m \le p$).

Examples:

1) θ: a scalar

H_o: $\theta_o = 2 \rightarrow H_o$: $\theta_o - 2 = 0 \rightarrow H_o$: $w(\theta) = 0$, where $w(\theta) = \theta - 2$.

2) $\theta = (\theta_1, \theta_2, \theta_3)'$.

H_o: $\theta_{1,o}^2 = \theta_{2,o} + 2$ and $\theta_{3,o} = \theta_{1,o} + \theta_{2,o}$.

$\rightarrow H_o$: $\theta_{1,o}^2 - \theta_{2,o} - 2 = 0$ and $\theta_{3,o} - \theta_{1,o} - \theta_{2,o} = 0$.

$$\rightarrow H_o:\ w(\theta) = \begin{pmatrix} w_1(\theta) \\ w_2(\theta) \end{pmatrix} = \begin{pmatrix} \theta_1^2 - \theta_2 - 2 \\ \theta_3 - \theta_1 - \theta_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

3) linear restrictions

$\theta = [\theta_1, \theta_2, \theta_3]'$.

H_o: $\theta_{1,o} = \theta_{2,o} + 2$ and $\theta_{3,o} = \theta_{1,o} + \theta_{2,o}$

$$\rightarrow H_o:\ w(\theta_o) = \begin{pmatrix} w_1(\theta_o) \\ w_2(\theta_o) \end{pmatrix} = \begin{pmatrix} \theta_{1,o} - \theta_{2,o} - 2 \\ \theta_{3,o} - \theta_{1,o} - \theta_{2,o} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

$$\rightarrow H_o:\ w(\theta_o) = \begin{pmatrix} 1 & -1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} \theta_{1,o} \\ \theta_{2,o} \\ \theta_{3,o} \end{pmatrix} - \begin{pmatrix} 2 \\ 0 \end{pmatrix} = R\theta_o - r.$$

Remark:

If all restrictions are linear in θ, H_o takes the following form:

$$H_o: R\theta_o - r = 0_{m\times 1},$$

where R and r are known m×p and m×1 matrices, respectively.

Definition:

$$W(\theta) \equiv \frac{\partial w(\theta)}{\partial \theta'} = \begin{pmatrix} \frac{\partial w_1(\theta)}{\partial \theta_1} & \frac{\partial w_1(\theta)}{\partial \theta_2} & \cdots & \frac{\partial w_1(\theta)}{\partial \theta_p} \\ \frac{\partial w_2(\theta)}{\partial \theta_1} & \frac{\partial w_2(\theta)}{\partial \theta_2} & \cdots & \frac{\partial w_2(\theta)}{\partial \theta_p} \\ \vdots & \vdots & & \vdots \\ \frac{\partial w_m(\theta)}{\partial \theta_1} & \frac{\partial w_m(\theta)}{\partial \theta_2} & \cdots & \frac{\partial w_m(\theta)}{\partial \theta_p} \end{pmatrix}_{m\times p}.$$

Example: (Nonlinear restrictions)

Let $\theta = [\theta_1,\theta_2,\theta_3]'$.

H_o: $\theta_{1,o}^2 - \theta_{2,o} = 0$ and $\theta_{1,o} - \theta_{2,o} - \theta_{3,o}^2 = 0$.

$$\rightarrow w(\theta) = \begin{pmatrix} \theta_1^2 - \theta_2 \\ \theta_1 - \theta_2 - \theta_3^2 \end{pmatrix}; \; W(\theta) = \begin{pmatrix} 2\theta_1 & -1 & 0 \\ 1 & -1 & -2\theta_3 \end{pmatrix}.$$

Example: (Linear restrictions)

$\theta = [\theta_1,\theta_2,\theta_3]'$

H_o: $\theta_{1,o} = 0$ and $\theta_{2,o} + \theta_{3,o} = 1$.

$$\to w(\theta) = \begin{pmatrix} \theta_1 \\ \theta_2 + \theta_3 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \to w(\theta) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{pmatrix} - \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

which is of form $w(\theta) = R\theta - r$.

Theorem:

Under (WIC.1)-(WIC.5),

$$\sqrt{T}\left(w(\hat{\beta}) - w(\beta_o)\right) \to_d N\left(0_{m\times 1}, W(\beta_o)\sigma^2 Q_o^{-1} W(\beta_o)'\right).$$

Proof:

Taylor's expansion around β_o:

$$w(\hat{\beta}) = w(\beta_o) + W(\overline{\beta})(\hat{\beta} - \beta_o),$$

where $\overline{\beta}$ is between $\hat{\beta}$ and β_o. Since $\hat{\beta}$ is consistent, so is $\overline{\beta}$. Thus,

$$\sqrt{T}\left(w(\hat{\beta}) - w(\beta_o)\right) \approx W(\beta_o)\sqrt{T}(\hat{\beta} - \beta_o)$$

$$\to_d N\left(0_{m\times 1}, W(\beta_o)\sigma^2 Q_o^{-1} W(\beta_o)'\right).$$

Implication:

$$\left(w(\hat{\beta}) - w(\beta_o)\right) \approx N\left(0_{m\times 1}, W(\hat{\beta}) s^2 (X'X)^{-1} W(\hat{\beta})'\right).$$

Theorem:

Under (WIC.1)-(WIC.5) and H_o: $w(\beta_o) = 0$,

$$W_T = w(\hat{\beta})'\left[W(\hat{\beta})Cov(\hat{\beta})W(\hat{\beta})'\right]^{-1} w(\hat{\beta}) \Rightarrow \chi^2(m).$$

Proof:

Under H_o: $w(\beta_o) = 0$,

$$w(\hat{\beta}) \approx N\left(0_{m\times 1}, W(\hat{\beta})Cov(\hat{\beta})W(\hat{\beta})'\right).$$

For a normal random vector $h_{m\times 1} \sim N(0_{m\times 1},\Omega_{m\times m})$, $h'\Omega^{-1}h \sim \chi^2(m)$. Thus, we obtain the desired result.

Question: What does "Wald test" mean?

A test based on the unrestricted estimator only.

(5) When the WIC are violated:

CASE 1: Simple dynamic model, $y_t = \beta y_{t-1} + \varepsilon_t$.

- SIC is violated. But WIC hold, if the ε_t i.i.d. $N(0,\sigma_o^2)$ and $-1 < \beta_o < 1$.
- If $\beta_o = 1$, WIC is also violated. For this case, the OLS is consistent, but not normally distributed.
- For simplicity, set $y_0 = 0$.
- $y_t = \Sigma_{s=1}^{t}\varepsilon_s \rightarrow \text{var}(y_t) = E(y_t^2) = t\sigma_o^2$.
- $\text{plim } (1/T)\Sigma_t x_{t\bullet}x_{t\bullet}' = \text{plim } (1/T)\Sigma_t y_{t-1}^2 = \lim (1/T)\Sigma_t E(y_{t-1}^2)$ (by GWLLN)

$$= \lim (1/T)\Sigma_t(t-1)\sigma_o^2 = \lim (1/T)[T(T-1)/2]\sigma_o^2$$

$$= \lim [(T-1)/2]\sigma_o^2 \rightarrow \infty \text{ (WIC.3 violated.)}$$

CASE 2: Deterministic trend model, $y_t = \beta t + \varepsilon_t$.

- $\text{plim } (1/T)\Sigma_t x_{t\bullet}x_{t\bullet}' = \text{plim } (1/T)\Sigma_t t^2 = \frac{1}{T}\frac{T(T+1)(2T+1)}{6} \rightarrow \infty.$
- WIC.3 is violated. But OLS estimator is consistent and asymptotically normal.

CASE 3: Simultaneous Equations models.

- (a) $c_t = \beta_{1,o} + \beta_{2,o}y_t + \varepsilon_t$; (b) $c_t + i_t = y_t$
- (a) → (b): $y_t = \beta_{1,o} + \beta_{2,o}y_t + \varepsilon_t + i_t$.
- $y_t = [\beta_{1,o}/(1-\beta_{2,o})] + i_t[1/(1-\beta_{2,o})] + \varepsilon_t/(1-\beta_{2,o})$.
- y_t is correlated with ε_t in (a).
- OLS is inconsistent.

CASE 4: Measurement errors:

- $y_t = \beta_o x_t^* + \varepsilon_t$ (true model).
- But we can observe $x_t = x_t^* + v_t$ (v_t: measurement error).
- If we use x_t for x_t^*,

$$y_t = x_t\beta_o + [\varepsilon_t - \beta_o v_t] \text{ (model we estimate).}$$

- x_t and $(\varepsilon_t - \beta_o v_t)$ correlated.
- OLS is inconsistent.

- $y_t^* = \beta_o x_t + \varepsilon_t$ (true model).
- But we can observe $y_t = y_t^* + v_t$.
- If we use y_t for y_t^*,

$$y_t = x_t\beta_o + [\varepsilon_t + v_t] \text{ (model we estimate)}$$

x_t and $(\varepsilon_t + v_t)$ uncorrelated.

- OLS is consistent.

[Proofs of Consistency and Asymptotic Normality Theorems]

(1) Show $p\lim \hat{\beta} = \beta_o$.

$$\hat{\beta} = \beta_o + (X'X)^{-1}X'\varepsilon = \beta_o + \left(T^{-1}\Sigma_t x_{t\bullet}x_{t\bullet}'\right)T^{-1}\Sigma_t x_{t\bullet}\varepsilon_t .$$

$p\lim_{T\to\infty} T^{-1}\Sigma_t x_{t\bullet}x_{t\bullet}' = Q_o$ (by WIC.3)

$p\lim_{T\to\infty} T^{-1}\Sigma_t x_{t\bullet}\varepsilon_t = \lim_{T\to\infty} \Sigma_t E(x_{t\bullet}\varepsilon_t)$ [by GWLLN]

$= \lim T^{-1}\Sigma_t 0$ [by WIC.1] $= 0$.

$\rightarrow$ plim $p\lim_{T\to\infty} \hat{\beta} = \beta_o + (Q_o)^{-1}0 = \beta_o$.

(2) Show plim $s^2 = \sigma_o^2$.

plim s^2 = plim SSE/T.

SSE/T $= \varepsilon' M(X)\varepsilon/T = \varepsilon'\varepsilon/T - \varepsilon'X(X'X)^{-1}X'\varepsilon/T$

$= T^{-1}\Sigma_t\varepsilon_t^2 - (T^{-1}\varepsilon'X)(T^{-1}X'X)^{-1}(T^{-1}X'\varepsilon)$

$= T^{-1}\Sigma_t\varepsilon_t^2 - (T^{-1}\Sigma_t\varepsilon_t x_{t\bullet}')(T^{-1}\Sigma_t x_{t\bullet}x_{t\bullet}')^{-1}(T^{-1}\Sigma_t x_{t\bullet}\varepsilon_t)$.

$$p\lim_{T\to\infty} T^{-1}\Sigma_t\varepsilon_t^2 = \lim_{T\to\infty} T^{-1}\Sigma_t E(\varepsilon_t^2) = \lim_{T\to\infty} T^{-1}\Sigma_t\sigma_o^2 = \sigma_o^2 .$$

$p\lim_{T\to\infty} T^{-1}\Sigma_t x_{t\bullet}\varepsilon_t = 0$.

$\rightarrow p\lim_{T\to\infty} s^2 = \sigma_o^2 - 0'(Q_o)^{-1}0 = \sigma_o^2$.

(3) Show $\sqrt{T}(\hat{\beta}-\beta) \to_d N\left(0_{k\times 1}, \sigma_o^2 Q_o^{-1}\right)$.

$$\hat{\beta} = \beta_o + \left(T^{-1}\Sigma_t x_{t\bullet} x_{t\bullet}'\right) T^{-1}\Sigma_t x_{t\bullet}\varepsilon_t$$

$$\to (\hat{\beta}-\beta) = \left(T^{-1}\Sigma_t x_{t\bullet} x_{t\bullet}'\right) T^{-1}\Sigma_t x_{t\bullet}\varepsilon_t$$

$$\to \sqrt{T}(\hat{\beta}-\beta) = [T^{-1}\Sigma_t x_{t\bullet} x_{t\bullet}']^{-1} \left(\frac{1}{\sqrt{T}}\Sigma_t x_{t\bullet}\varepsilon_t\right).$$

$\to$ By GCLT with martingale difference,

$$\frac{1}{\sqrt{T}}\Sigma_t x_{t\bullet}\varepsilon_t \to_{\mathrm{d}} \mathrm{N}(0, \lim \mathrm{T}^{-1}\Sigma_t \mathrm{Cov}(\mathrm{x}_{t\bullet}\varepsilon_t))$$

$$Cov(x_{t\bullet}\varepsilon_t) = E(x_{t\bullet}\varepsilon_t\varepsilon_t x_{t\bullet}') = E\left(\varepsilon_t^2 x_{t\bullet} x_{t\bullet}'\right)$$

$$= E_{x_{t\bullet}}[E({\varepsilon_t}^2 x_{t\bullet} x_{t\bullet}' \mid x_{t\bullet})] \text{ (by LIE)}$$

$$= E_{x_{t\bullet}}[E({\varepsilon_t}^2 \mid x_{t\bullet}) x_{t\bullet} x_{t\bullet}'] = E_{x_t}(\sigma_o^2 x_{t\bullet} x_{t\bullet}') = \sigma_o^2 E(x_{t\bullet} x_{t\bullet}').$$

$$\lim_{T\to\infty} T^{-1} Cov(x_{t\bullet}\varepsilon_{t\bullet}) = \sigma_o^2 \lim_{T\to\infty} T^{-1}\Sigma_t E(x_{t\bullet} x_{t\bullet}') = \sigma_o^2 Q_o.$$

$$\to \frac{1}{\sqrt{T}}\Sigma_t x_{t\bullet}\varepsilon_t \to_d N(0_{k\times 1}, \sigma_o^2 Q_o).$$

$$\to \sqrt{T}(\hat{\beta}-\beta_o) \to_d N((Q_o)^{-1} 0_{k\times 1}, (Q_o)^{-1}\sigma_o^2 Q_o (Q_o)^{-1}) = N(0_{k\times 1}, \sigma_o^2 (Q_o)^{-1})$$