R300 Econometrics

Metrics Enjoyers

Michaelmas Term, 2023-2024

Contents

1 Basic Probability. Conditional expectation function.

1.1 Random Variables

Definition 1.1.1: Cumulative distribution function

The cumulative distribution function of X is defined as $F_X(x) \equiv P(X \leq x)$. A function F is a cdf iff:

1. $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$;
2. $F(\cdot)$ non-decreasing;
3. $F(\cdot)$ right-continuous; i.e., $\forall x_0$, $\lim_{x\downarrow x_0} F(x) = F(x_0)$.

Definition 1.1.2: Probability density function

For a continuous r.v., $f_X(x)$ defined as the function which satisfies $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$ for all x. A function f_X is a pdf iff:

1. $\forall x$, $f_X(x) \geq 0$;
2. $\int_{\mathbb{R}} f_X(x)\,dx = 1$.

f_X gives the probability of any event: $P(X \in B) = \int_{\mathbb{R}} 1_{(x\in B)} f_X(x)\,dx$.
A continuous (in all dimensions) random vector X has joint pdf $f_X(x_1, \ldots, x_n)$ iff $\forall A \subseteq \mathbb{R}^n$, $P(X \in A) = \int \cdots \int_A f_X(x_1, \ldots, x_n)\,dx_1 \cdots dx_n$.

Exercise 1.1.1. Show that the standard normal density integrates to unity by showing (when $u > 0$):

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}uy^2} dy = \frac{1}{\sqrt{u}}.$$

Solution:-

$$\left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y^2} dy\right] \left[\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx\right] = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(x^2+y^2)} dxdy.$$

By changing to polar coordinates, $x^2+y^2 = r^2$ and $dxdy = rdrd\theta$. Thus, the desired integral becomes:

$$\frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}ur^2} rdrd\theta = \frac{1}{u}$$

Setting $u = 1$ yields the desired result.

Definition 1.1.3: τ-th quantile

Let X be a random variable with distribution function F_X. The τ-th quantile of X is defined as the value x_τ such that

$$F_X^{-1}(\tau) = \inf\{x : F_X(x) \geq \tau\}$$

where $0 \leq \tau \leq 1$.

Why inf and not min?

Because F is right-continuous and non-decreasing, the superlevel sets of F are of the form $[a, \infty]$ where $a > -\infty$ or else the entire real line. When the superlevel set is the whole line, there is no min (among the reals), while the inf is -∞. For a $= +\infty$ the superlevel set is empty and so the inf $= +\infty$. These cases can potentially arise when $\tau = 0$ or $\tau = 1$ respectively. *If $\tau \in (0, 1)$ then we can replace inf with min.*

If X is discrete, then using minimum and infimum are equivalent, since the support is finite and attains a minimum at some point. However, a continuous X with infinite support will not achieve a minimum, hence the infimum is needed.

Example. The CDF of an Exponential distribution with parameter λ is given by

$$F(x; \lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0 & x < 0. \end{cases}$$

The quantile function for Exponential(λ) is derived by finding the value of Q for which $1 - e^{-\lambda Q} = p$:

$$Q(p; \lambda) = \frac{-\ln(1-p)}{\lambda},$$

for $0 \leq p < 1$. The quartiles are therefore:

- First quartile ($p = 1/4$): $-\ln(3/4)/\lambda$
- Median ($p = 1/2$): $-\ln(1/2)/\lambda$
- Third quartile ($p = 3/4$): $-\ln(1/4)/\lambda$.

Definition 1.1.4: Expectation

For a function g, the expectation of $g(X)$ is defined as $\mathbb{E}g(X) = \int_{-\infty}^{\infty} g(x)f(x)\,dx$.

Exercise 1.1.2. Suppose that Y is a continuous random variable with density $f(y)$ that is positive only if $y \geq 0$. If $F(y)$ is the distribution function, show that

$$\mathbb{E}(Y) = \int_0^\infty [1 - F(y)]dy$$

Solution:-

$$E(Y) = \int_0^\infty yf(y)dy = \int_0^\infty \left(\int_0^y dt\right) f(y)dy = \int_0^\infty \left(\int_t^\infty f(y)dy\right) dt$$
$$= \int_0^\infty P(Y > y)dy = \int_0^\infty [1 - F(y)]dy$$

Definition 1.1.5: Moment

For $n \in \mathbb{Z}$, the nth moment of X is $\mu'_n \equiv \mathbb{E}X^n$. Also denote $\mu'_1 = \mathbb{E}X$ as μ. The nth central moment is $\mu_n \equiv \mathbb{E}(X-\mu)^n$.

Two different distributions *can* have all the same moments, but only if the variables have unbounded support sets. Note that $\mathbb{E}X^n$ may not exist (the integral might be infinite), then we say the nth moment does not exist.

Notable moments and properties:

- The first raw moment is the mean, $\mu = \mathbb{E}[X]$
  - $\mathbb{E}[ag_1(X)+bg_2(X)+c] = a\mathbb{E}(g_1(X))+b\mathbb{E}(g_2(X))+c$ (i.e., expectation is a linear operator)
  - The mean is the MSE minimizing predictor for X; i.e., $\min_b \mathbb{E}(X-b)^2 = \mathbb{E}(X-\mathbb{E}X)^2$
  - If $X_1, \ldots, X_n$ mutually independent, then $\mathbb{E}[g_1(X_1)\cdot \cdots \cdot g_n(X_n)] = \mathbb{E}[g_1(X_1)] \cdot \cdots \cdot \mathbb{E}[g_n(X_n)]$.
- The second central moment is the variance, $\mathbb{E}[(x-\mu)^2]$
  - $Var(aX+bY) = a^2 VarX + b^2 VarY + 2abCov(X,Y)$
  - $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$ (i.e.: residual variance + regression variance)
  - $Var\mathbf{X} \equiv \mathbb{E}[\mathbf{X}\mathbf{X}'] - \mathbb{E}[\mathbf{X}]\mathbb{E}[\mathbf{X}]'$
  - $Var(\mathbf{X}+\mathbf{Y}) = Var(\mathbf{X}) + Cov(\mathbf{X},\mathbf{Y}) + Cov(\mathbf{X},\mathbf{Y})' + Var(\mathbf{Y})$;
  - $Var(\mathbf{AX}) = \mathbf{A}Var(\mathbf{X})\mathbf{A}'$.
  - $Cov(\mathbf{AX},\mathbf{BY}) = \mathbf{A}Cov(\mathbf{X},\mathbf{Y})\mathbf{B}'$;
  - $Cov(\mathbf{X},\mathbf{Y}) = Cov(\mathbf{Y},\mathbf{X})'$.
- The third central moment is the measure of lopsidedness of the distribution. When standardised by the standard deviation it is known as the skewness. Any symmetric distribution will have skewness of 0.
- The fourth central moment is a measure of the heaviness of the tail. When standardised by the standard deviation, it is known as the kurtosis:

$$\text{Kurt}[X] = \mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mu_4}{\mu_2^2}.$$

Example. Find μ'_n for the uniform random variable with $\theta_1 = 0$ and $\theta_2 = \theta$.
By definition,

$$\mu'_n = E(Y^n) = \int_{-\infty}^{\infty} y^n f(y)\,dy = \int_0^{\theta} y^n \left(\frac{1}{\theta}\right) dy = \left.\frac{y^{n+1}}{\theta(n+1)}\right|_0^{\theta} = \frac{\theta^n}{n+1}.$$

Thus,

$$\mu'_1 = \mu = \frac{\theta}{2}, \quad \mu'_2 = \frac{\theta^2}{3}, \quad \mu'_3 = \frac{\theta^3}{4},$$

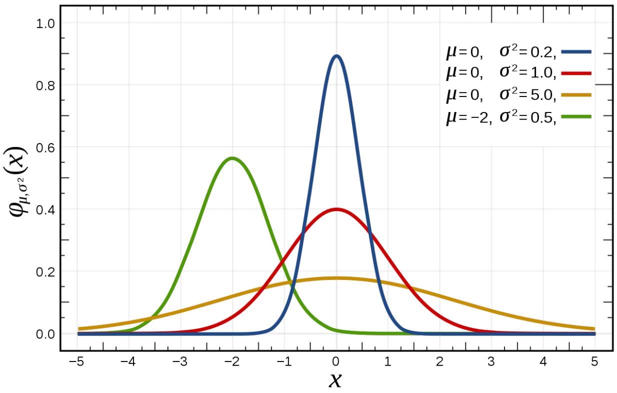
and so on.

1.2 Common Distributions

Normal (Gaussian)

PDF:

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- $\mathbb{E}[X] = \mu$
  $\mathbb{E}[(X-\mu)] = 0$
- $\mathbb{E}[X^2] = \mu^2 + \sigma^2$
  $\mathbb{E}[(X-\mu)^2] = \sigma^2$
- $\mathbb{E}[X^3] = \mu^3 + 3\mu\sigma^2$
  $\mathbb{E}[(X-\mu)^3] = 0$
- $\mathbb{E}[X^4] = \mu^4 + 6\mu^2\sigma^2 + 3\sigma^4$
  $\mathbb{E}[(X-\mu)^4] = 3\sigma^4$

Properties

- The distribution is entirely characterised by the first two moments
- Square of standard normal is χ_1^2.
- If $X \sim N(\mu,\sigma^2)$, $Y \sim N(\gamma,\tau^2)$, and $X \perp\!\!\!\perp Y$, then $X+Y \sim N(\mu+\gamma,\sigma^2+\tau^2)$ (i.e., independent normals are additive in mean and variance).
- For a standard normal: $\mathbb{E}[Z^k] = 0$ if k odd, $\mathbb{E}[Z^k] = 1 \cdot 3 \cdot 5 \cdots (n-1)$ if k even.
- Ratio of independent standard normals is Cauchy ($\sigma = 1$, $\theta = 0$)

Lemma 1.2.1 (Stein's Lemma). If $g(\cdot)$ is differentiable with $\mathbb{E}|g'(X)| < \infty$, then $\mathbb{E}[g(X)(X-\mu)] = \sigma^2\mathbb{E}g'(X)$.

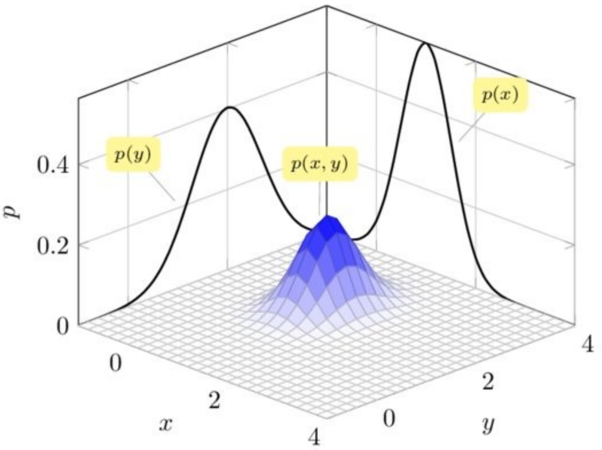
Proof. We shall prove in the case of a standard normal: $\phi(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}$
Since $\int x\exp(-x^2/2)\,dx = -\exp(-x^2/2)$ we get from integration by parts:
$E[g(X)X] = \frac{1}{\sqrt{2\pi}}\int g(x)x\exp(-x^2/2)\,dx = \frac{1}{\sqrt{2\pi}}\int g'(x)\exp(-x^2/2)\,dx = E[g'(X)]$. □

Multivariate Normal

PDF:

$$\frac{1}{\sqrt{(2\pi)^k|\mathbf{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

where $\mu = \mathbb{E}[\mathbf{X}]$ and $\mathbf{\Sigma}_{ij} = Cov(X_i, X_j)$



Bivariate Case

- $\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}$
- $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$

Properties

- A linear transformation of a normal is normal: if $\mathbf{X} \sim N_p(\mu, \boldsymbol{\Sigma})$, then for any $\mathbf{A} \in \mathbb{R}^{q\times p}$ with full row rank ($\Rightarrow q \leq p$), and any $\mathbf{b} \in \mathbb{R}^q$, we have $\mathbf{AX} + \mathbf{b} \sim N_q(\mathbf{A}\mu + \mathbf{b}, \mathbf{A\Sigma A'})$. In particular, $\boldsymbol{\Sigma}^{-1/2}(\mathbf{X} - \mu) \sim N(\mathbf{0}, \mathbf{I})$.
- The following transformations of $\mathbf{X} \sim N_p(\mu, \boldsymbol{\Sigma})$ are independent iff $\mathbf{A\Sigma B'} = Cov(\mathbf{AX}, \mathbf{BX}) = \mathbf{0}$:
  - $\mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{A\Sigma A'})$ and $\mathbf{BX} \sim N(\mathbf{B}\mu, \mathbf{B\Sigma B'})$,
  - $\mathbf{AX} \sim N(\mathbf{A}\mu, \mathbf{A\Sigma A'})$ and $\mathbf{X'BX} \sim \chi^2_{rk(\mathbf{B}\Sigma)}$ (where $\mathbf{B\Sigma}$ is an idempotent matrix),
  - $\mathbf{X'AX} \sim \chi^2_{rk(\mathbf{A}\Sigma)}$ and $\mathbf{X'BX} \sim \chi^2_{rk(\mathbf{B}\Sigma)}$ (where $\mathbf{A\Sigma}$ and $\mathbf{B\Sigma}$ are idempotent matrices).
- If X and Y are both normal and independent, this implies they are jointly normally distributed (i.e. (X, Y) is multivariate normal). However, a pair of jointly normal distributed variables need not be independent (would only be of if uncorrelated, $\rho = 0$).
- Independence and zero-covariance are equivalent for linear functions of normally distributed r.v.s.

Example (Individual normality $\not\Rightarrow$ joint normality). Consider $X \sim N(0, 1)$, and:

$$Y = \begin{cases} X, & \text{if } |X| \leq c \\ -X, & \text{if } |X| > c \end{cases} \quad \text{where } c > 0$$

When c is very small, $\operatorname{corr}(X, Y) \approx -1$ and when c is very large, $\operatorname{corr}(X, Y) \approx 1$. If the correlation is a continuous function of c, then there exists some c such that the correlation is 0. X and Y are uncorrelated, but clearly not independent since X completely determines Y. To show Y is normal:

$$\begin{aligned} P(Y \leq x) &= P(|X| < c \text{ and } X \leq x) + P(|X| > c \text{ and } -X \leq x) \\ &= P(|X| < c \text{ and } X \leq x) + P(|X| > c \text{ and } X \geq -x) \\ &= P(X \leq x) \end{aligned}$$

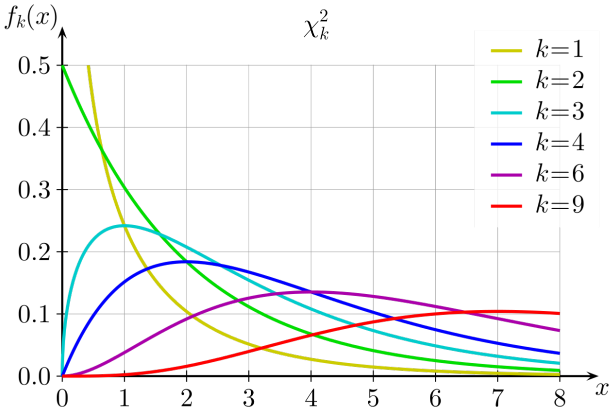
using the symmetry of $|X|$ and $|X| \leq c$. Note that $X - Y$ is not normally distributed due to the non-zero probability of $X - Y = 0$. However, a normal has no discrete part, i.e.: the probability of any point is 0. Thus, X and Y are not jointly normally distributed, even though they are individually normally distributed.

Chi-Squared (χ^2)

PDF:

$$\chi_k^2 = \sum_{i=1}^{k} Z_i^2$$

where $Z_i \overset{\text{iid}}{\sim} N(0,1)$



- $\mathbb{E}[X] = k$
  $\mathbb{E}[(X-k)] = 0$
- $\mathbb{E}[X^2] = k(k+2)$
  $\mathbb{E}[(X-k)^2] = 2k$
- $\mathbb{E}[X^3] = k(k+2)(k+4)$
  $\mathbb{E}[(X-k)^3] = 8k$
- $\mathbb{E}[X^4] = k(k+2)(k+4)(k+6)$
  $\mathbb{E}[(X-k)^4] = 12k^2 + 48k$

Properties

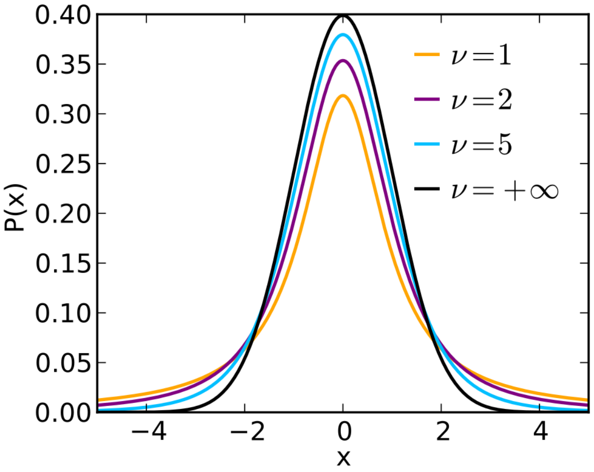
- If $X_1, \ldots, X_n$ are independent with $X_i \sim \chi^2_{p_i}$, then $\sum X_i \sim \chi^2_{\sum p_i}$ (i.e., independent chi squared variables add to a chi squared, and the degrees of freedom add).
- If $\mathbf{X} \sim N_n(\mu, \mathbf{\Sigma})$, then $(\mathbf{X}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu) \sim \chi^2_n$.
- If $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I})$ and $\mathbf{P}_{n\times n}$ is an idempotent matrix, then $\mathbf{X}'\mathbf{P}\mathbf{X} \sim \chi^2_{rk(\mathbf{P})} = \chi^2_{\mathrm{tr}(\mathbf{P})}$.
- If $\mathbf{X} \sim N_n(\mathbf{0}, \mathbf{I})$ then the sum of the squared deviations from the sample mean $\mathbf{X}'\mathbf{M}_\iota\mathbf{X} \sim \chi^2_{n-1}$.

Student's t

PDF:

$$t_\nu = \frac{Z}{\sqrt{X/\nu}} = c\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

where $Z \sim N(0,1)$, $X \sim \chi^2_\nu$



- Mean: 0 for $\nu > 1$
- Variance: $\frac{\nu}{\nu-2}$ for $\nu > 2$, ∞ for $1 < \nu \leq 2$
- Skewness: 0 for $\nu > 3$
- Ex. kurtosis: $\frac{6}{\nu-4}$ for $\nu > 4$, ∞ for $2 < \nu \leq 4$

Why does the ν-th moment of t_ν not exist?

Consider the ν-th raw moment: $\int x^\nu c\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx \approx \int c\nu^{\frac{\nu+1}{2}} x^{-1} dx$ when x is large. This integral diverges, meaning the ν-th raw moment does not exist. A more rigorous proof requires the use of the Beta and Gamma functions.

Properties

- If $X_1, \dots, X_n$ are iid $N(\mu, \sigma^2)$, then $\sqrt{n}(\bar{X} - \mu)/\sigma \sim N(0,1)$. However, we will generally not know σ. Using the sample variance rather than the true variance gives $\sqrt{n}(\bar{X} - \mu)/s \sim t_{n-1}$.
- If a t distribution has ν degrees of freedom, there are only $\nu - 1$ defined moments. ν has thicker tails than normal.
- t_1 is Cauchy distribution (the ratio of two independent standard normals). t_∞ is standard normal.

Example (Derive variance of Student's t). Consider $X \sim t_\nu$. When $\nu > 1$:

$$E(X) = 0$$

$$(t_\nu)^2 \sim F_{1,\nu} \Rightarrow E(X^2) = E(Y)$$

with $Y \sim F_{1,\nu}$, where $F_{1,\nu}$ is the F-distribution with $(1, \nu)$ degrees of freedom. $E(Y)$ exists if and only if $\nu > 2$:

$$E(Y) = E(X^2) = \frac{\nu}{\nu - 2}$$

We therefore have:

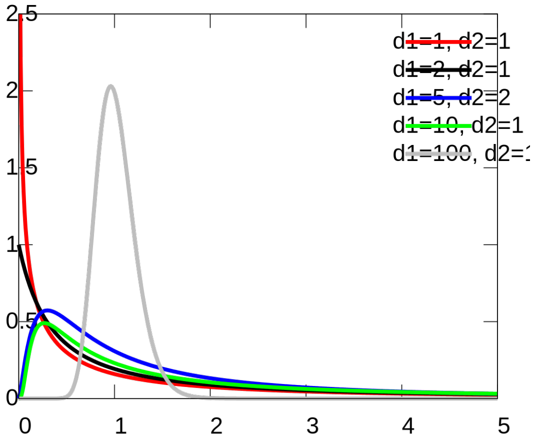
$$\text{var}(X) = E(X^2) - (E(X))^2 = \frac{\nu}{\nu - 2}$$

Snedecor's F

PDF:

$$F_{d_1,d_2} = \frac{X_1/d_1}{X_2/d_2}$$

where $X_1 \sim \chi^2_{d_1}$, $X_2 \sim \chi^2_{d_2}$



- Mean: $\frac{d_2}{d_2 - 1}$ for $d_2 > 2$
- Variance: $\frac{2d_2^2(d_1+d_2-2)}{d_1(d_2-2)^2(d_2-4)}$, for $d_2 > 4$

Properties

- $1/F_{p,q} \sim F_{q,p}$ (i.e., the reciprocal of an F r.v. is another F with the degrees of freedom switched);
- $(t_q)^2 \sim F_{1,q}$;
- If $X \sim F_{p,q}$ then $Y = \lim_{q \to \infty} pX \sim \chi^2_p$

1.3 Conditional expectation function

Definition 1.3.1: Conditional distribution

Conditional distribution of Y given X is defined as

$$f_{Y|X}(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} \quad \text{if } f_X(x) \neq 0$$

Conditional expectation $E(Y|X=x)$ is defined as

$$E(Y|X=x) = \int_y y f_{Y|X}(y|x)dy$$

Often, we will skip $X = x$ having in mind that $E(Y|X)$ is a function of random variable X. Hence, it is itself a random variable.
We can also condition for/on multiple coordinates: e.g., for (X_1, X_2, X_3, X_4) a continuous random vector, $f(x_3, x_4|x_1, x_2) \equiv f(x_1, x_2, x_3, x_4)/f_{X_1X_2}(x_1, x_2)$, where f is a joint pdf, and $f_{X_1X_2}$ is the marginal pdf in X_1 and X_2.

Note:-

Borel Paradox: Be careful when we condition on events of probability zero: two events of probability zero may be equivalent, but the probabilities conditional on the two events is different!

Theorem 1.3.1 (Law of Iterated Expectations). $\mathbb{E}X = \mathbb{E}[\mathbb{E}(X|Y)]$, provided the expectations exist. More generally, when $\mathcal{L} \subseteq \mathcal{M}$ (i.e., $\mathcal{L}$ contains less information, $\mathcal{M}$ contains more),

$$\mathbb{E}[X|\mathcal{L}] = \mathbb{E}[\mathbb{E}(X|\mathcal{M})|\mathcal{L}] = \mathbb{E}[\mathbb{E}(X|\mathcal{L})|\mathcal{M}].$$

Proof.

$$E(Y) = \int_y y f_Y(y)dy = \int_x \int_y y f_{XY}(x,y)dxdy = \int_x \int_y y f_{YX}(x,y)dydx$$
$$= \int_x \int_y y f_{Y|X}(y|x) f_X(x) dydx = \int_x E(Y|X=x) f_X(x)dx = E(E(Y|X)).$$

□

Theorem 1.3.2. $\mathbb{E}(Y|X)$ is the MSE $= E(Y - g(X))^2$ minimising predictor of Y based on knowledge of X.

Proof.

$$\begin{aligned} E(Y-g(X))^2 &= E[Y - E(Y|X) + E(Y|X) - g(X)]^2 \\ &= E[Y - E(Y|X)]^2 + 2E[(Y - E(Y|X))(E(Y|X) - g(X))] + E[E(Y|X) - g(X)]^2 \end{aligned}$$

Using the law of iterated expectations: $E(Z) = E(E(Z|X))$

$$E[(Y - E(Y|X))(E(Y|X) - g(X))] = E(E[(Y - E(Y|X))(E(Y|X) - g(X))]|X)$$

Bring terms explained fully by X outside expectation

$$= E([E(Y|X) - g(X)]E\{[Y - E(Y|X)]|X\})$$

Expand conditional expectation

$$
\begin{aligned}
&= E([E(Y|X) - g(X)]\{E(Y|X) - E(Y|X)\}) \\
&= 0
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
2E[(Y - E(Y|X))(E(Y|X) - g(X))] &= 0 \Rightarrow \\
E(Y - g(X))^2 &= E[Y - E(Y|X)]^2 + E[E(Y|X) - g(X)]^2 \\
&\geq E[Y - E(Y|X)]^2.
\end{aligned}
$$

and CEF is the best conditional predictor of Y □

Lemma 1.3.1 (Leibniz Rule). Let $f(x,t)$ be a continuously differentiable function then, for the function

$$F(t) = \int_{a(t)}^{b(t)} f(x,t)\,dx$$

the derivative of $F(t)$ with respect to t is given by

$$\frac{dF}{dt} = \int_{a(t)}^{b(t)} \frac{\partial f}{\partial t}\,dx + f(b(t),t)\cdot\frac{db}{dt} - f(a(t),t)\cdot\frac{da}{dt}$$

Theorem 1.3.3. The conditional median $med(Y|X)$ is the expected absolute error $= E(|Y - g(X)||X = x)$ minimizing predictor of Y based on knowledge of X.

The following proof is a complete version of the outline Alexei presents in the notes. A brief (similar) proof is given at the end.

Proof.

$$
\begin{aligned}
E(|Y - g(X)||X = x) &= \int_{-\infty}^{\infty} |y - g(x)| f_{Y|X}(y|x)dy \\
&= \int_{g(x)}^{\infty} (y - g(x)) f_{Y|X}(y|x)dy + \int_{-\infty}^{g(x)} (g(x) - y) f_{Y|X}(y|x)dy.
\end{aligned}
$$

Assume that $f_{Y|X}$ is zero to the left of some constant A, and is unity to the right of some constant B. The problem is:

$$\min_{g(x)}\left\{\phi = \int_{g(x)}^{A} (y - g(x)) f_{Y|X}(y|x)dy + \int_{-B}^{g(x)} (g(x) - y) f_{Y|X}(y|x)dy\right\}$$

Applying Leibniz rule, we have:

$$
\begin{aligned}
\frac{d\phi}{dg(x)} &= \int_{A}^{g(x)} (1) f_{Y|X}(y|x)dy + (g(x) - g(x))(1) - (g(x) - A)(0) \\
&+ \int_{g(x)}^{B} (-1) f_{Y|X}(y|x)dy + (B - g(x))(0) - (g(x) - g(x))(1)
\end{aligned}
$$

FOC:

$$0 = \int_A^{g(x)} f_{Y|X}(y|x)dy - \int_{g(x)}^{B} f_{Y|X}(y|x)dy \Rightarrow \int_A^{g(x)} f_{Y|X}(y|x)dy = \int_{g(x)}^{B} f_{Y|X}(y|x)dy$$

Hence, $g(x)$ must be the value of Y such that $P(Y \leq g(x)|X = x) = P(Y > g(x)|X = x)$. That is, $g(x)$ must be the median of the conditional distribution $F_{Y|X}$.
To verify that we have minimized $E(|Y - g(x)||X = x)$:

$$\frac{d^2\phi}{dg(x)^2} = \frac{\partial}{\partial g(x)} \left(\int_A^{g(x)} f_{Y|X}(y|x)dy - \int_{g(x)}^{B} f_{Y|X}(y|x)dy \right)$$

$$= \int_A^{g(x)} 0 f_{Y|X}(y|x)dy + 1\left(\frac{dg(x)}{dg(x)}\right) - 1\left(\frac{dA}{dg(x)}\right) - \int_{g(x)}^{B} 0 f_{Y|X}(y|x)dy + 1\left(\frac{dB}{dg(x)}\right) - 1\left(\frac{dg(x)}{dg(x)}\right)$$

$= [0 + 1 - 0] - [0 + 0 - 1] = 2(> 0)$ so we are characterising a minimum. □

Also, note that if we let $A \to -\infty$ and $B \to \infty$, the support of F can be taken to be the whole real line, so there is no loss of generality in establishing the above result with a support of $[A, B]$.

Alternative Proof

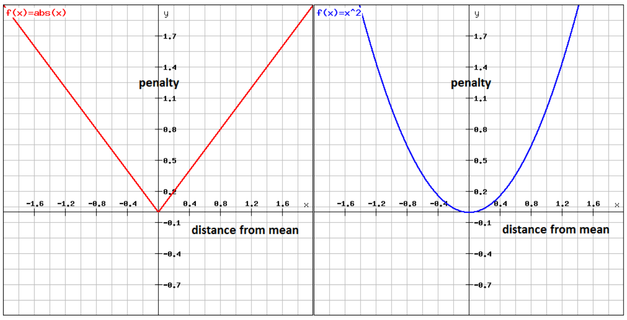
$$\frac{d}{dc}E(|X - c|) = E\left(\frac{d}{dc}|X - c|\right) = E\left(\frac{-(X - c)}{|X - c|}\right)$$

$$= E\left[1_{\{X<c\}} - 1_{\{X>c\}}\right] = P(X < c) - P(X \geq c)$$

$$\frac{d}{dc}E(|X - c|) = 0 \Rightarrow P(X < c) = P(X > c) = \frac{1}{2}$$

By definition of the median, $c = med(X)$ □

MAE vs MSE



- MAE $= E|Y - g(X)|$
- MSE $= E(Y - g(X))^2$

- MAE imposes a linear penalty on errors, i.e.: each deviation from the mean is given a proportional corresponding error.
- MSE is a squared proportional relationship between deviation and penalty. This will make sure that the further you are away from the mean, the proportionally more you will be penalized. Using this penalty function, outliers are deemed proportionally more informative than observations near the mean.

Because the MAE is a more robust estimator of scale than the sample variance or standard deviation, it works better with distributions without a mean or variance, such as the Cauchy distribution.

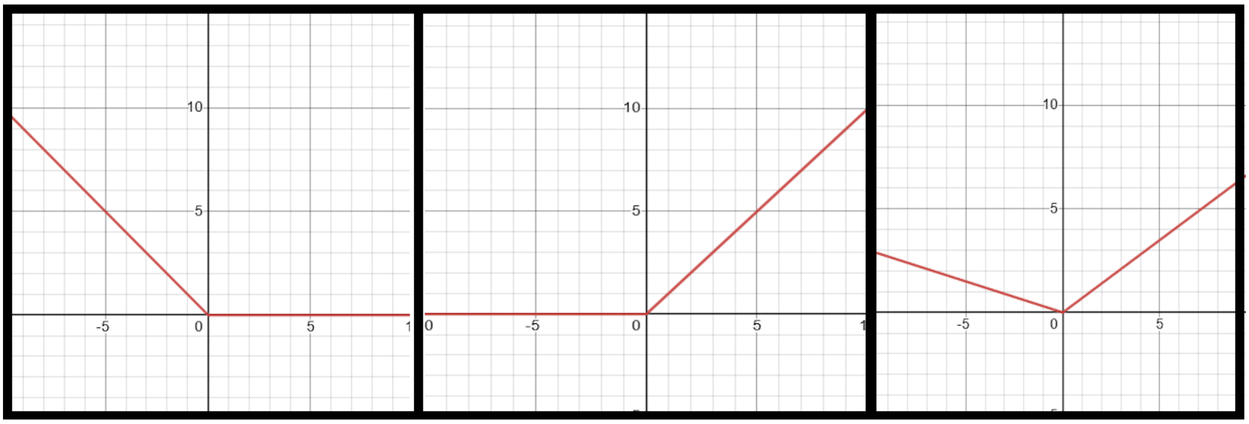
Weighted MAE

If underprediction is marginally less or more costly as overprediction, it makes sense to minimize the expectation of

$$\tau 1(Y > g(X))(Y - g(X)) + (1 - \tau)1(Y \leq g(X))(g(X) - Y)$$

with $\tau \in (0,1)$. For example, parameter $\tau < 1/2$ would correspond to situations where the underprediction is less costly than overprediction. Following the same logic as above, we can show that *the corresponding best predictor would be τ-th quantile $\tau(X)$ of the conditional distribution of Y given X.*

Below we have (from left to right): $\tau = 1$ (no cost to overprediction), $\tau = 0$ (no cost to underprediction) and $\tau = 0.3$ (cost to both, but relatively more to overprediction.)



2 Causal interpretation of regression. Least Squares.

2.1 Regression and Causality

A variable x_1 can be said to have a causal effect on the response variable y if the latter changes when all other inputs are held constant. We can write a full model for the response variable y as:

$$y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$$

where x_1 and $\mathbf{x_2}$ are the observed variables, $\boldsymbol{\varepsilon}$ is an $\ell \times 1$ unobserved random factor and h is a functional relationship.

Definition 2.1.1: Causal effect

In the model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ the **causal effect** of x_1 on y is

$$C(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon}) = \nabla_1 h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon}),$$

the change in y due to a change in x_1, holding $\mathbf{x_2}$ and $\boldsymbol{\varepsilon}$ constant.

Note:-

This is just a definition, and does not necessarily describe causality in a fundamental or experimental sense. It might be more appropriate to label this a structural effect (the effect within the structural model).

Example. Suppose firms have Cobb-Douglas production functions:

$$y = AK^{\alpha}L^{\beta}$$

where K, L are observed capital and labour, A is an unobserved production technology and y is output. Here $x_1 = K, x_2 = L, \varepsilon = A$. Then the causal effect of capital on output is

$$C(K, L, A) = y'(K, L, A) = \alpha AK^{\alpha-1}L^{\beta}.$$

Even for firms with identical inputs, this effect differs due to unobserved A.

Sometimes it is useful to write this relationship as a potential outcomes function

$$y(x_1) = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$$

where the notation implies that $y(x_1)$ is holding $\mathbf{x_2}$ and $\boldsymbol{\varepsilon}$ constant. A popular example arises in the analysis of treatment effects with a binary regressor x_1. Let $x_1 = 1$ indicate treatment (e.g., a medical procedure) and $x_1 = 0$ indicate non-treatment. In this case $y(x_1)$ can be written

$$y(0) = h(0, x_2, \boldsymbol{\varepsilon}), \;\; y(1) = h(1, x_2, \boldsymbol{\varepsilon})$$

where $y(0)$ and $y(1)$ are known as the latent outcomes associated with non-treatment and treatment, respectively. The causal effect of treatment for the individual is the change in their health outcome due to treatment; the change in y as we hold both x_2 and $\boldsymbol{\varepsilon}$ constant:

$$C(x_2, \boldsymbol{\varepsilon}) = y(1) - y(0).$$

This is random as both potential outcomes $y(0)$ and $y(1)$ are different across individuals.

Example. Suppose there are two individuals Yinfeng and Charles, and both have the possibility of being a PhD graduate or dropping out. Suppose Yinfeng would earn £8/hour without a PhD and £12/hour as a PhD grad, while Charles would earn £20/hour without and £30/hour with a PhD. The causal effect of a PhD on wages is £4/hour for Yinfeng and £10/hour for Charles.

In a sample, we cannot observe both outcomes from the same individual, we only observe the realised value. As the causal effect varies across individuals and is not observable, it cannot be measured on the individual level. We therefore focus on aggregate causal effects, in particular what is known as the average causal effect.

Definition 2.1.2: Average causal effect

In the model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ the **average causal effect** of x_1 on y conditional on $\mathbf{x_2}$ is

$$\begin{aligned} ACE(x_1, \mathbf{x_2}) &= \mathbb{E}(C(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \,|\, x_1, \boldsymbol{x_2}) \\ &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \end{aligned}$$

where $f(\boldsymbol{\varepsilon} \,|\, x_!, \boldsymbol{x_2})$ is the conditional density of $\boldsymbol{\varepsilon}$ given $x_1, \boldsymbol{x_2}$.

Example. In the Cobb-Douglas example, the ACE of capital on output will be:

$$ACE(K, L) = \mathbb{E}(\alpha A K^{\alpha-1} L^\beta | K, L) = \alpha \mathbb{E}(A|K, L) K^{\alpha-1} L^\beta$$

Example. Considering again Yinfeng and Charles, suppose half our population are Yinfeng's and the other half Charles's, then the average causal effect of a PhD is $(10+4)/2 =$£7/hour. This is not the individual causal effect, it is the average of the causal effect across all individuals in the population.

We can think of $ACE(x_1, \mathbf{x_2})$ as the average effect in the general population. When we conduct regression analysis we might hope that regression reveals the ACE, i.e.: what is the relationship between $ACE(x_1, \mathbf{x_2})$ and the regression derivative $\nabla_1 m(x_1, \mathbf{x_2})$? The model $h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$ implies that the CEF is

$$\begin{aligned} m(x_1, \boldsymbol{x_2}) &= \mathbb{E}(h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \,|\, x_1, \boldsymbol{x_2}) \\ &= \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon}, \end{aligned}$$

the average causal equation, averaged over the conditional distribution of the unobserved component $\boldsymbol{\varepsilon}$.

Applying the marginal effect operator [1], the regression derivative is:

$$\begin{aligned} \nabla_1 m(x_1, \boldsymbol{x_2}) &= \int_{\mathbb{R}^\ell} \nabla_1 h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} + \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \\ &= ACE(x_1, \boldsymbol{x_2}) + \int_{\mathbb{R}^\ell} h(x_1, \boldsymbol{x_2}, \boldsymbol{\varepsilon}) \nabla_1 f(\boldsymbol{\varepsilon} \,|\, x_1, \boldsymbol{x_2}) d\boldsymbol{\varepsilon} \end{aligned}$$

In general we see that the regression derivative does not equal the average causal effect. They are only equal in the special case when the second term equals zero, which occurs when the conditional

[1] Alexei uses $\frac{\partial}{\partial x_1}$ throughout, this is equivalent to the marginal effect operator used here with continuous x_1.

density of $\boldsymbol{\varepsilon}$ given $(x_1, \boldsymbol{x_2})$ does not depend on x_1 ($\nabla_1 f(\boldsymbol{\varepsilon}\,|\,x_1, \boldsymbol{x_2}) = 0$). When this condition holds then the regression derivative equals the ACE, which means that regression analysis can be interpreted causally, in the sense that it uncovers average causal effects.

Definition 2.1.3: Condiional Independence Assumption (CIA)

Conditional on $\mathbf{x_2}$, the random variables x_1 and $\boldsymbol{\varepsilon}$ are statistically independent.

The CIA implies $f(\boldsymbol{\varepsilon}\,|\,x_1, \boldsymbol{x_2}) = f(\boldsymbol{\varepsilon}\,|\,\boldsymbol{x_2})$ does not depend on x_1, and thus $\nabla_1 f(\boldsymbol{\varepsilon}\,|\,x_1, \boldsymbol{x_2}) = 0$. Thus the CIA implies that the regression derivative equals the ACE.

Theorem 2.1.1. In the structural model $y = h(x_1, \mathbf{x_2}, \boldsymbol{\varepsilon})$, the CIA implies

$$\nabla_1 m(x_1, \boldsymbol{x_2}) = ACE(x_1, \boldsymbol{x_2})$$

the regression derivative equals the average causal effect for x_1 on y conditional on $\mathbf{x_2}$.

Example (Nerlove: Returns to scale in electriciy supply)**.** Nerlove investigated returns to scale in a regulated industry (U.S. electricity) using Cobb-Douglas production. The market had the following features:

1. Privately owned local monopolies supply electricity on demand
2. These local monopolies face competitive factor prices
3. Electricity prices are set by the government

Notably Y is exogenously given (by consumer demand). Nerlove assumes firms pick K, L to minimise the cost of producing $Y = AK^\alpha L^\beta$, i.e. K, L both depend on A, Y, in particular $f(A|K, L)$ depends on K. Thus a regression of Y on K, L will not identify the ACE.

$$\min_{K,L} p_K K + p_L L \;\; s.t. \;\; Y = AK^\alpha L^\beta$$

The Lagrangian and FOCs for this problem are:

$$\mathcal{L} = p_K K + p_L L + \lambda(Y - AK^\alpha L^\beta)$$

$$\frac{\partial \mathcal{L}}{\partial K} = p_K - \lambda \alpha A K^{\alpha-1} L^\beta = 0, \quad \frac{\partial \mathcal{L}}{\partial L} = p_L - \lambda \beta A K^\alpha L^{\beta-1} = 0$$

$$\Rightarrow K = \frac{\alpha p_L}{\beta p_K} L$$

We can substitute this into the production function to solve for L and K, giving:

$$TC = p_K \left(\frac{\alpha p_L}{\beta p_K} \left(\frac{Y}{A \left(\frac{\alpha p_L}{\beta p_K} \right)^\alpha} \right)^{\frac{1}{\alpha+\beta}} \right) + p_L \left(\left(\frac{Y}{A \left(\frac{\alpha p_L}{\beta p_K} \right)^\alpha} \right)^{\frac{1}{\alpha+\beta}} \right)$$

$$TC = p_L \left(\frac{Y \left(\frac{p_L \alpha}{p_K \beta} \right)^{-\alpha}}{A} \right)^{\frac{1}{r}} \left(\frac{r}{\beta} \right) = r\alpha^{-\alpha/r} \beta^{-\beta/r} A^{-1/r} Y^{1/r} p_K^{\alpha/r} p_L^{\beta/r}$$

Taking logs we obtain the following log-linear relationship for each firm:

$$\log(TC_i) = \mu_i + \frac{1}{r}\log(Y_i) + \frac{\alpha}{r}\log(p_K) + \frac{\beta}{r}\log(p_L)$$

where $\mu_i = \log[r(A_i\alpha^\alpha\beta^\beta)^{-\frac{1}{r}}]$. Coefficients in this equation are elasticities, for example $\frac{\beta}{r}$ is the elasticity of total cost with respect to the wage rate, i.e.: the percentage change in in total cost when the wage rate changes by 1%. The degree of returns to scale (the reciprocal of the output elasticity of total costs), is independent of the level of output.
To estimate this define $\mu \equiv \mathbb{E}[\mu_i]$, $\varepsilon_i \equiv \mu - \mu_i$ so $\mathbb{E}[\varepsilon_i] = 0$, firms with positive ε_i are high-cost firms.

$$\log(TC_i) = \beta_0 + \beta_1 log(Y_i) + \beta_2\log(p_K) + \beta_3\log(p_L),$$

where

$$\beta_0 = \mu, \beta_1 = \frac{1}{r}, \beta_2 = \frac{\alpha}{r}, \beta_3 = \frac{\beta}{r}$$

This equation is overidentified, the 4 coefficients are not free parameters, they are a function of three technology parameters (α, β, μ). Clearly $\beta_2 + \beta_3 = 1$ (as expected, cost function is linearly homogenous in factor prices). To fix this we can subtract p_L from each side and consider relative prices:

$$\log\left(\frac{TC_i}{p_L}\right) = \beta_0 + \beta_1 log(Y_i) + \beta_2\log\left(\frac{p_K}{p_L}\right)$$

To test constant returns to scale $(r = 1)$, just t-test $\beta_1 = 1$ in this restricted model.

2.2 Estimating population regression by least squares

If CIA holds, regression captures the causal effect of x's on y. However even if it doesn't, it still provides the best predictor of y given x's. We assume the regression function $\mathbb{E}[y|x] = m(x)$ is parametrised by a finite dimensional vector $\beta = [\beta_1, ..., \beta_k]^T$, so that estimating the population regression $m(x; \beta)$ is equivalent to estimating β. One approach to estimation is using the analogy principle.

Definition 2.2.1: Analogy principle

Consider finding an estimator that satisfies the same properties in the sample that the parameter satisfies in the population; i.e., seek to estimate $\beta(P)$ with $\beta(P_n)$ where P_n is the empirical distribution which puts mass $\frac{1}{n}$ at each sample point. Note this distribution converges uniformly to P.

In the regression context:

$$\beta = \arg\min_b \mathbb{E}(y - m(x; b))^2$$

The sample analogue of expectation is the average:

$$\hat{\beta} = \arg\min_b \frac{1}{n}\sum_{i=1}^{n}(y - m(x; b))^2$$

When $m(x; b)$ is linear in b, the method is called OLS. We assume that the observations of the data (y_i, x_i) are independent and come from the same joint distribution. Let

$$\underbrace{X_i}_{K\times 1} = \begin{bmatrix} x_{i1} \\ \vdots \\ x_{iK} \end{bmatrix}, \underbrace{\beta}_{K\times 1} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix},$$

$$\underbrace{Y}_{n\times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \underbrace{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \underbrace{X}_{n\times K} = \begin{bmatrix} X_1' \\ \vdots \\ X_n' \end{bmatrix}.$$

When our model contains a constant, one of the columns of X will contain only ones. Our linear model can thus be represented as:

$$Y = X\beta + \varepsilon$$

When estimating we select the $\hat{\beta}$ such that the sum of squared residuals ($e'e$) is minimised[1].

$$\begin{aligned} e'e &= (y - X\hat{\beta})'(y - X\hat{\beta}) \\ &= (y' - \hat{\beta}'X')(y - X\hat{\beta}) \\ &= y'y - \hat{\beta}'X'y - y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta} \\ &= y'y - 2\hat{\beta}'X'y + \hat{\beta}'X'X\hat{\beta} \end{aligned}$$

Where $y'X\hat{\beta} = (y'X\hat{\beta})' = \hat{\beta}'X'y$ since the transpose of a scalar is itself.

Note:-

Matrix differentiation

$$\frac{\partial \mathbf{a}'\mathbf{b}}{\partial \mathbf{b}} = \frac{\partial \mathbf{b}'\mathbf{a}}{\partial \mathbf{b}} = \mathbf{a} \tag{2.1}$$

when $\mathbf{a}$ and $\mathbf{b}$ are $K \times 1$ vectors.

$$\frac{\partial \mathbf{b}'\mathbf{A}\mathbf{b}}{\partial \mathbf{b}} = 2\mathbf{A}\mathbf{b} = 2\mathbf{A}'\mathbf{b} \tag{2.2}$$

when $\mathbf{A}$ is any symmetric matrix.

$$\frac{\partial 2\mathbf{b}'\mathbf{X}'\mathbf{y}}{\partial \mathbf{b}} = \frac{\partial 2\mathbf{b}'(\mathbf{X}'\mathbf{y})}{\partial \mathbf{b}} = 2\mathbf{X}'\mathbf{y} \tag{2.3}$$

and

$$\frac{\partial \mathbf{b}'\mathbf{X}'\mathbf{X}\beta}{\partial \mathbf{b}} = \frac{\partial 2\mathbf{A}\beta}{\partial \mathbf{b}} = 2\mathbf{A}\beta = 2\mathbf{X}'\mathbf{X}\beta \tag{2.4}$$

when $\mathbf{X}'\mathbf{X}$ is a $K \times K$ matrix.

Solving for the minimum:

$$\begin{aligned} \frac{\partial e'e}{\partial \hat{\beta}} &= -2X'y + 2X'X\hat{\beta} = 0 \\ &\Rightarrow X'X\hat{\beta} = X'y \\ &\Rightarrow (X'X)^{-1}(X'X)\hat{\beta} = (X'X)^{-1}X'y \\ &\Rightarrow I_K\hat{\beta} = (X'X)^{-1}X'y \\ &\Rightarrow \hat{\beta} = (X'X)^{-1}X'y \end{aligned}$$

Here we have assumed that the inverse of $X'X$ exists, i.e. X is full rank[2]. To check this is a minimum, take second derivative which gives us $2X'X$ which is clearly positive semi-definite (when X is full rank). Note that $X'X$ is always square ($k \times k$) and always symmetric.

[1] Note that $e \neq \varepsilon$, residuals e are observed, whilst disturbances ε are unobserved.

[2] The inverse of $X'X$ may not exist, it does not exist in the following two cases: 1) When $n < k$; we have more independent variables than observations 2) One or more of the independent variables are a linear combination of the other variables i.e. perfect multicollinearity.

We can further show that $X'e = 0$, consider the normal form equations $X'X\hat{\beta} = X'y$:

$$\begin{aligned}(\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= \mathbf{X}'(\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{e}) \\ (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})\hat{\boldsymbol{\beta}} + \mathbf{X}'\mathbf{e} \\ \mathbf{X}'\mathbf{e} &= \mathbf{0}\end{aligned}$$

Proposition 2.2.1 (Properties of OLS). From $X'e$=0 we can derive a number of properties.

1. The observed values of X are uncorrelated with the residuals.
2. The sum of the residuals is zero.
3. The sample mean of the residuals is zero.
4. The regression hyperplane passes through the sample means of observables.
5. The predicted values of y are uncorrelated with the residuals.

Where 2-5 hold when the regression includes a constant term.

Proof. Using $X'e = 0$

1. $\mathbf{X}'\mathbf{e} = 0$ implies that for every column $\mathbf{x}_k$ of $\mathbf{X}$, $\mathbf{x}_k'\mathbf{e} = 0$. In other words, each regressor has zero sample correlation with the residuals. Note that this does not mean that $\mathbf{X}$ is uncorrelated with the disturbances; we'll have to assume this.
2. If there is a constant, then the first column in $\mathbf{X}$ (i.e. $\mathbf{X}_1$) will be a column of ones. This means that for the first element in the $\mathbf{X}'\mathbf{e}$ vector (i.e. $\mathbf{X}_{11}e_1 + \mathbf{X}_{12}e_2 + \ldots + \mathbf{X}_{1n}e_n$), to be zero, it must be the case that $\sum_i e_i = 0$.
3. This follows straightforwardly from the previous property i.e. $\bar{e} = \frac{\sum e_i}{n} = 0$.
4. This follows from the fact that $\bar{e} = 0$. Recall that $e = y - \mathbf{X}\hat{\boldsymbol{\beta}}$. Dividing by the number of observations, we get $\bar{e} = \bar{y} - \bar{\mathbf{X}}\hat{\boldsymbol{\beta}} = 0$. This implies that $\bar{y} = \bar{\mathbf{X}}\hat{\boldsymbol{\beta}}$.
5. $\hat{y}'e = (\mathbf{X}\hat{\boldsymbol{\beta}})'e = \hat{\boldsymbol{\beta}}'\mathbf{X}'e = 0$

□

5 Finite sample tests of linear hypotheses.

5.1 Linear hypotheses

The t-test is appropriate when the null hypothesis is a real valued restriction. However, more generally there may be multiple restrictions on the coefficient vector $\boldsymbol{\beta}$. Suppose we have $p > 1$ restrictions, we can express a linear hypothesis about $\boldsymbol{\beta}$ in the form $\boldsymbol{R}_{p\times k}\boldsymbol{\beta}_{k\times 1} = \boldsymbol{q}_{p\times 1}$.

Example (Nerlove's returns to scale). Nerlove studied the regression of the total cost of electricity production on demand (Q_i) and factor prices (capital, labour and fuel):

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log p_{C_i} + \beta_4 \log p_{L_i} + \beta_5 \log p_{F_i} + \varepsilon_i$$

Economic theory suggests that $\beta_2 = \frac{1}{r}$ where r is the degree of returns to scale. To test constant returns we can use $H_0 : \beta_2 = 1$, which is trivially linear in components of $\boldsymbol{\beta}$. Alternatively we can write

$$R\beta = q$$

with $R = (0, 1, 0, 0, 0)$ and $q = 1$.

Further the total cost must be homogenous of degree 1 with respect to factor prices (doubling cost of all inputs doubles total cost). To test this we can consider $H_0 : \beta_3 + \beta_4 + \beta_5 = 1$. If we were to reject this it would suggest model misspecification.

To test these hypotheses simultaneously consider:

$$R\beta = q \quad \text{with} \quad R = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \quad \text{and} \quad q = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

To test H_0: $\boldsymbol{R\beta} = \boldsymbol{q}$ vs. H_1: $\boldsymbol{R\beta} \neq \boldsymbol{q}$ we compute the vector $\boldsymbol{R\hat{\beta}} = \boldsymbol{q}$ and reject the null if this vector is "too large" depending on the distribution of $\hat{\boldsymbol{\beta}}$ under H_0.

Definition 5.1.1: Wald statistic

When restrictions are a linear function of coefficients $\boldsymbol{\beta}$, we can write the Wald statistic as

$$W = (R\hat{\beta} - q)'(R\hat{V}_{\hat{\beta}}R')^{-1}(R\hat{\beta} - q)$$

i.e. a weighted Euclidean measure of the length of the vector $R\hat{\beta} - q$.

Note:-

As the Wald statistic is symmetric in the argument $R\hat{\beta} - q$ it treats positive and negative alternatives symmetrically. Thus the inherent alternative is always two-sided.

The Wald statistic is not-invariant to a non-linear transformation/reparametrisation of the hypothesis. For example, asking whether $\beta_1 = 1$ is the same as asking whether $\log \beta_1 = 0$; but the Wald statistic for $\beta_1 = 1$ is not the same as the Wald statistic for $\log \beta_1 = 0$. This is because there is in general no neat relationship between the standard errors of β_1 and $\log \beta_1$, so it needs to be approximated.

Assuming normal regression:

$$\begin{aligned}
\hat{\beta}|X &\sim N(\beta, \sigma^2(X'X)^{-1}) \\
R\hat{\beta}|X &\sim N(R\beta, \sigma^2 R(X'X)^{-1}R') \\
R\hat{\beta} - q|X &\sim N(R\beta - q, \sigma^2 R(X'X)^{-1}R') \\
&\overset{H_0}{\sim} N(0, \sigma^2 R(X'X)^{-1}R')
\end{aligned}$$

We can thus standardise:

$$(\sigma^2 R(X'X)^{-1}R')^{-\frac{1}{2}}(R\hat{\beta} - q)|X \overset{H_0}{\sim} N(0, I_P)$$

$$(R\hat{\beta} - q)'(\sigma^2 R(X'X)^{-1}R')^{-1}(R\hat{\beta} - q)|X \overset{H_0}{\sim} \chi^2(p) \tag{5.1}$$

However, the true variance σ^2 is unknown, we thus replace it with the estimated $\hat{\sigma}^2$ to obtain the Wald statistic:

$$\begin{aligned}
W &= (R\hat{\beta} - q)'(\hat{\sigma}^2 R(X'X)^{-1}R')^{-1}(R\hat{\beta} - q) \\
&= \frac{(R\hat{\beta} - q)'(\sigma^2 R(X'X)^{-1}R')^{-1}(R\hat{\beta} - q)}{\hat{\sigma}^2/\sigma^2}
\end{aligned}$$

Note that this distribution is not $\chi^2(p)$ since $\hat{\sigma}^2$ is itself a random variable. We must consider the joint distribution of $\boldsymbol{\hat{\sigma}^2}$ and $\boldsymbol{\hat{\beta}}$ to make progress.

5.2 The joint distribution of $\hat{\sigma}^2$ and $\hat{\beta}$

Recall the definition of the variance estimator:

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}$$

To express this in terms of the population $\boldsymbol{\varepsilon}$'s examine the following, where we denote the residual maker matrix by $\mathbf{M_X} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$:

$$\begin{aligned}
(n-k)\hat{\sigma}^2 &= \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \\
&= (\mathbf{M_X y})'\mathbf{M_X y} \\
&= (\mathbf{M_X}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}))'\mathbf{M_X}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) \\
&= \boldsymbol{\varepsilon}'\mathbf{M_X'}\mathbf{M_X}\boldsymbol{\varepsilon} \qquad (\text{since } \mathbf{M_X X} = \mathbf{0}) \\
&= \boldsymbol{\varepsilon}'\mathbf{M_X}\boldsymbol{\varepsilon} \qquad (\text{since } \mathbf{M_X'M_X} = \mathbf{M_X M_X} = \mathbf{M_X})
\end{aligned}$$

Since $\mathbf{M_X}$ is symmetric, it is positive definite when all eigenvalues are positive. Since it is also idempotent, $\mathbf{M_X^2} = \mathbf{M_X}$, all eigenvalues are either zero or one, meaning $\mathbf{M_X}$ is positive semi-definite.[1]

Lemma 5.2.1 (Spectral decomposition). For every $n \times n$ real symmetric matrix, the eigenvalues are real and the eigenvectors can be chosen real and orthonormal. Thus a real symmetric matrix $\mathbf{A}$ can be decomposed as

$$\mathbf{A} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}'$$

where $\mathbf{Q}$ is an orthogonal matrix whose columns are the real, orthonormal eigenvectors of $\mathbf{A}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$.

[1] Alternatively since $\mathbf{M_X^2} = \mathbf{M_X}$ and $\mathbf{M_X}' = \mathbf{M_X}$, note that $\mathbf{v}'\mathbf{M_X v} = \mathbf{v}'\mathbf{M_X^2 v} = \mathbf{v}'\mathbf{M_X}'\mathbf{M_X v} = (\mathbf{v}'\mathbf{M_X})'(\mathbf{M_X v}) = \|\mathbf{M_X v}\|^2$ for all $\mathbf{v} \in \mathbb{R}^n$.

The spectral decomposition of $\mathbf{M_X}$ is $\mathbf{M_X} = \mathbf{H\Lambda H'}$ where $\mathbf{HH'} = \mathbf{I_n}$ and $\mathbf{\Lambda}$ is diagonal with the eigenvalues of $\mathbf{M_X}$ along the diagonal. Since $\mathbf{M_X}$ is idempotent with rank $n-k$, it has $n-k$ eigenvalues equalling 1 and k eigenvalues equalling 0, so:

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_k \end{bmatrix}$$

In the normal regression $\boldsymbol{\varepsilon} \sim N(0, \mathbf{I_n}\sigma^2)$, we want to find the distribution of $\mathbf{H}'\boldsymbol{\varepsilon}$. A linear combination of normals is also normal, meaning $\mathbf{H}'\boldsymbol{\varepsilon}$ is normal with mean $\mathbb{E}[\mathbf{H}'\boldsymbol{\varepsilon}] = \mathbf{H}'\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ and variance $\mathrm{Var}(\mathbf{H}'\boldsymbol{e}) = \mathbf{H'I_n}\sigma^2\mathbf{H} = \sigma^2\mathbf{H'H} = \mathbf{I_n}\sigma^2$. Thus $\mathbf{H}'\boldsymbol{\varepsilon} \sim N(0, \mathbf{I_n}\sigma^2)$.

Let $\mathbf{u} = \mathbf{H}'\boldsymbol{\varepsilon}$, and partition $\underset{n\times 1}{\mathbf{u}} = \begin{bmatrix} \underset{(n-k)\times 1}{\mathbf{u_1}} \\ \underset{k\times 1}{\mathbf{u_2}} \end{bmatrix}$ where $\mathbf{u_1} \sim N(0, \mathbf{I_n}\sigma^2)$, then we have

$$\begin{aligned}
(n-k)\hat{\sigma}^2 &= \boldsymbol{\varepsilon}'\mathbf{M_X}\boldsymbol{\varepsilon} \\
&= \boldsymbol{\varepsilon}'\mathbf{H\Lambda H'}\boldsymbol{\varepsilon} \\
&= \mathbf{u}' \begin{bmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_k \end{bmatrix} \mathbf{u} \\
&= [\mathbf{u_1'} \ \ \mathbf{u_2'}] \begin{bmatrix} \mathbf{I}_{n-k} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_k \end{bmatrix} \begin{bmatrix} \mathbf{u_1} \\ \mathbf{u_2} \end{bmatrix} \\
&= \mathbf{u_1'u_1}
\end{aligned}$$

where $\mathbf{u_1'u_1}$ is the sum of $n-k$ squared standard normals , thus it is distributed χ^2_{n-k}. Since $\boldsymbol{\varepsilon}$ is independent of $\hat{\boldsymbol{\beta}}$ it follows that $\hat{\sigma}^2$ is independent of $\hat{\boldsymbol{\beta}}$ as well.

Theorem 5.2.1. In normal regression,

$$\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-k}$$

and is independent of $\hat{\boldsymbol{\beta}}$.

Corollary 5.2.1. In normal regression satisfying GM1-3, the normalised Wald statistic $\frac{W}{p}$, is distributed as $F(p, n-k)$ under the null.

Proof.

$$\frac{W}{p} = \frac{(R\hat{\beta}-q)'(\sigma^2 R(X'X)^{-1}R')^{-1}(R\hat{\beta}-q)/p}{\hat{\sigma}^2/\sigma^2} \sim \frac{\chi^2(p)/p}{\chi^2(n-k)/(n-k)} \sim F(p, n-k).$$

Where we have used 5.1 in the numerator, and Theorem 5.2.1 in the denominator. □

Consider a special case of testing a single restriction, that the j-th coefficient is zero. Then $R\hat{\beta}_j - q = \beta_j$:

$$\hat{\beta}_j | X \overset{H_0}{\sim} N(0, \sigma^2(X'X)^{-1}_{ij})$$

$$\frac{\hat{\beta}_j}{\sqrt{\sigma^2(X'X)^{-1}_{jj}}} | X \overset{H_0}{\sim} N(0,1)$$

As before σ^2 is unknown, we can substitute in $\hat{\sigma}^2$, but the distribution will change:

$$\begin{aligned}
t &= \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{jj}}} \\
&= \frac{\hat{\beta}_j / \sqrt{\sigma^2 (X'X)^{-1}_{jj}}}{\sqrt{\frac{(n-k)\hat{\sigma}^2}{\sigma^2}/(n-k)}} \\
t|X &\overset{H_0}{\sim} \frac{N(0,1)}{\sqrt{\chi^2(n-k)/(n-k)}} \\
&\overset{H_0}{\sim} t(n-k)
\end{aligned}$$

Where we are using the fact that the numerator and denominator are independent conditional on X. Note that the square of the t-statistic equals the F-statistic for testing the single restriction.

$$\begin{aligned}
t^2(n-k) &= \left(\frac{N(0,1)}{\sqrt{\chi^2(n-k)/(n-k)}} \right)^2 \\
&= \frac{\chi^2(1)/1}{\chi^2(n-k)/(n-k)} \\
&= F(1, n-k)
\end{aligned}$$

It is preferable to use the t-statistic since we can test one-sided alternatives, by squaring it we kill the sign of $\hat{\beta}_j$, making it impossible to differentiate between left and right sided alternatives.

5.3 The familiar form of the F-statistic

Consider the following test:

$$H_0 : R\beta = q \text{ vs. } H_1 : R\beta \neq q.$$

Proposition 5.3.1. The normalised Wald statistic is equivalent to the following formula for the F-statistic when testing linear restrictions:

$$F = \frac{W}{p} = \frac{(RSS_r - RSS_u)/p}{RSS_u/(n-k)}$$

Proof. Let us impose the null hypothesis $R\beta = q$ when minimising the sum of squared residuals, denote the solution as the restricted least squares estimator $\tilde{\beta}$:

$$\min_{\beta} (Y - X\beta)'(Y - X\beta) \quad \text{s.t.} \quad R\beta = q$$

$$\mathcal{L}(\beta) = (Y - X\beta)'(Y - X\beta) + \lambda'(R\beta - q)$$

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \beta} &= -2X'(Y - X\tilde{\beta}) + R'\lambda = 0 \\
&\Rightarrow X'Y - X'X\tilde{\beta} = R' \left(\frac{\lambda}{2} \right) \\
&\Rightarrow (X'X)^{-1}X'Y - (X'X)^{-1}X'X\tilde{\beta} = (X'X)^{-1}R' \left(\frac{\lambda}{2} \right)
\end{aligned}$$

Define the usual (unrestricted) OLS estimate as $\hat{\beta} = \hat{\beta}_{OLS} = (X'X)X'Y$

$$\begin{aligned}
\Rightarrow & \hat{\beta} - \tilde{\beta} = (X'X)^{-1}R'\left(\frac{\lambda}{2}\right) \\
\Rightarrow & \tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'\left(\frac{\lambda}{2}\right) \\
\Rightarrow & R\tilde{\beta} = R\hat{\beta} - R(X'X)^{-1}R'\left(\frac{\lambda}{2}\right)
\end{aligned}$$

Since $R\tilde{\beta} = q$:

$$\begin{aligned}
q &= R\hat{\beta} - R(X'X)^{-1}R'\left(\frac{\lambda}{2}\right) \\
R\hat{\beta} - q &= R(X'X)^{-1}R'\left(\frac{\lambda}{2}\right) \\
\Rightarrow (R(X'X)^{-1}R')^{-1}(R\hat{\beta} - q) &= \frac{\lambda}{2}
\end{aligned}$$

Thus,

$$\tilde{\beta} = \hat{\beta} - (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta} - q)$$

Now from the corresponding restricted and unrestricted residuals,

$$\hat{\varepsilon} = Y - X\hat{\beta}$$

$$\tilde{\varepsilon} = Y - X\tilde{\beta} = X\hat{\beta} + \hat{\varepsilon} - X\tilde{\beta} = \hat{\varepsilon} + X(\hat{\beta} - \tilde{\beta})$$

Since $\hat{\varepsilon}'X = 0$ [a]

$$\begin{aligned}
\tilde{\varepsilon}'\tilde{\varepsilon} &= (\hat{\varepsilon} + X(\hat{\beta} - \tilde{\beta}))'(\hat{\varepsilon} + X(\hat{\beta} - \tilde{\beta})) \\
&= \hat{\varepsilon}'\hat{\varepsilon} + \hat{\varepsilon}'X(\hat{\beta} - \tilde{\beta}) + (\hat{\beta} - \tilde{\beta})'X'\hat{\varepsilon} + (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta}) \\
&= \hat{\varepsilon}'\hat{\varepsilon} + (\hat{\beta} - \tilde{\beta})'X'X(\hat{\beta} - \tilde{\beta})
\end{aligned}$$

and substituting $\hat{\beta} - \tilde{\beta} = (X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta} - q)$,

$$\begin{aligned}
\tilde{\varepsilon}'\tilde{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon} &= ((X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta} - q))'\cancel{X'X(X'X)^{-1}}R'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta} - q) \\
&= (R\hat{\beta} - q)'\left(R(X'X)^{-1}R'\right)^{-1}\cancel{R(X'X)^{-1}R'\left(R(X'X)^{-1}R'\right)^{-1}}(R\hat{\beta} - q) \\
&= (R\hat{\beta} - q)'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta} - q)
\end{aligned}$$

Finally,

$$\frac{W}{p} = \frac{(R\hat{\beta} - q)'\left(R(X'X)^{-1}R'\right)^{-1}(R\hat{\beta} - q)/p}{\hat{\sigma}^2} = \frac{(\tilde{\varepsilon}'\tilde{\varepsilon} - \hat{\varepsilon}'\hat{\varepsilon})/p}{\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k}} = \frac{(RSS_r - RSS_u)/p}{RSS_u/(n-k)}$$

□

[a] I.e.: Unrestricted OLS residuals uncorrelated with regressors, see lecture 2 for an explanation

6 Convergence concepts. Asymptotics of OLS.

6.1 Convergence concepts

Definition 6.1.1: Convergence in probability

A sequence of random scalars $\{z_i\}_{i=1}^{\infty}$ converges in probability to z iff $\forall \varepsilon > 0$, $\lim_{n\to\infty} P(|z_n - z| \geq \varepsilon) = 0$, or equivalently $\lim_{n\to\infty} P(|z_n - z| < \varepsilon) = 1$. Written as $z_n \xrightarrow{p} z$ or $z_n - z = o_p(1)$ or $\text{plim}_{n\to\infty} z_n = z$.

This definition is extended to a sequence of random vectors or random matrices by requiring element-by-element convergence in probability. That is, a sequence of K-dimensional vectors $\mathbf{z_n}$ converges in probability to a K-dimensional vector $\mathbf{z}$ if, for any $\varepsilon > 0$

$$\lim_{n\to\infty} P(|z_{nk} - z_k| > \varepsilon) = 0 \quad \text{for all} k = 1, 2, ..., K$$

where z_{nk} is the k-th element of $\mathbf{z_n}$ and z_k the k-th element of $\mathbf{z}$.

Exercise 6.1.1. Let X_n be an IID sequence of continuous random variables having a uniform distribution over support

$$R_{X_n} = \left[-\frac{1}{n}, \frac{1}{n}\right]$$

with pdf

$$f_{X_n}(x) = \begin{cases} \frac{n}{2} & \text{if } x \in \left[-\frac{1}{n}; \frac{1}{n}\right] \\ 0 & \text{if } x \notin \left[-\frac{1}{n}; \frac{1}{n}\right] \end{cases}$$

Find the probability limit (if it exists) of the sequence X_n.

Solution:-

Intuitively as $n \to \infty$ the probability density becomes concentrated around $x = 0$; it seems reasonable to conjecture $X_n \xrightarrow{p} X = 0$. To show this formally, for any $\varepsilon > 0$:

$$\begin{aligned}
\lim_{n\to\infty} P(|X_n - X| > \varepsilon) &= \lim_{n\to\infty} P(|X_n - 0| > \varepsilon) \\
&= \lim_{n\to\infty} [1 - P(-\varepsilon \leq X_n \leq \varepsilon)] \\
&= 1 - \lim_{n\to\infty} \int_{-\varepsilon}^{\varepsilon} f_{X_n}(x)dx \\
&= 1 - \lim_{n\to\infty} \int_{\max(-\varepsilon,-1/n)}^{\min(\varepsilon,1/n)} \frac{n}{2}dx \quad (f(x) \text{ has no density outside } [-\tfrac{1}{n}, \tfrac{1}{n}]) \\
&= 1 - \lim_{n\to\infty} \int_{-1/n}^{1/n} \frac{n}{2}dx \quad (\text{when } n \text{ becomes large, } \frac{1}{n} < \varepsilon) \\
&= 1 - \lim_{n\to\infty} 1 \\
&= 0
\end{aligned}$$

Definition 6.1.2: Convergence in distribution

A sequence of random scalars $\{z_i\}_{i=1}^{\infty}$ converges in distribution to z iff, $\lim_{n\to\infty} F_{z_n}(z) = F_z(z)$ at all points where F_z is continuous. Written as $z_n \xrightarrow{d} z$ or $z_n - z = O_p(1)$ or as "z is the limiting distribution of z_n".

Convergence in distribution is also known as weak convergence or the convergence in law.

Theorem 6.1.1. $\mathbf{z_n} \xrightarrow{d} \mathbf{z}$ iff $\mathbb{E}f(\mathbf{z_n}) \to \mathbb{E}f(\mathbf{z})$ for all bounded, continuous functions f.

Claim 6.1.1. Convergence in probability implies convergence in distribution but not vice versa. The reverse only holds when the limit in distribution is a constant.

Example ($z_n \xrightarrow{d} z \nRightarrow z_n \xrightarrow{p} z$)**.** Let $z \sim N(0,1)$. Let $z_n = -z$ for $n = 1, 2, 3, \ldots$; hence $z_n \sim N(0,1)$. z_n has the same distribution function as z for all n so, trivially, $\lim_{n\to\infty} F_n(x) = F(x)$ for all x. Therefore, $z_n \xrightarrow{d} z$. But $P(|z_n - z| > \varepsilon) = P(|2z| > \varepsilon) = P(|z| > \varepsilon/2) \neq 0$. So z_n does not tend to z in probability.

The extension to a sequence of random vectors is immediate: $\mathbf{z_n} \xrightarrow{d} \mathbf{z}$ if the joint c.d.f. F_n of the random vector $\mathbf{z_n}$ converges to the joint c.d.f. F of $\mathbf{z}$ at every continuity point of F. However, element-by-element convergence does not necessarily imply convergence for the vector sequence (unlike with convergence in probability). Intuitively this is because different c.d.f.'s can have the same marginals.

A common way to establish the connection between scalar convergence in distribution and vector convergence in distribution is for every linear combination of z_{nk} to converge to the linear combination of z_n. Formally:

Definition 6.1.3: Cramer-Wold device

$\mathbf{z_n} \xrightarrow{d} \mathbf{z}$ if and only if $\lambda'\mathbf{z_n} \xrightarrow{d} \lambda'\mathbf{z}$ for every $\lambda \in \mathbb{R}^k$ with $\lambda'\lambda = 1$.

Note:-

Big O Little o notation

- Roughly speaking, a function is $o(z)$ iff it's of lower asymptotic order than z.
- $f(n) = o(g(n))$ iff $\lim_{n\to\infty} f(n)/g(n) = 0$.
- If $\{f(n)\}$ is a sequence of random variables, then $f(n) = o_p(g(n))$ iff $plim_{n\to\infty} f(n)/g(n) = 0$.
- We write $X_n - X = o_p(n^{-\gamma})$ iff $n^{\gamma}(X_n - X) \xrightarrow{\text{p}} 0$.

- Roughly speaking, a function is $O(z)$ iff it's of the same asymptotic order as z.
- $f(n) = O(g(n))$ iff $|f(n)/g(n)| < K$ for all $n > N$ and some positive integer N and some constant $K > 0$.
- If $\{f(n)\}$ is a sequence of random variables, then $f(n) = o_p(g(n))$ iff $plim_{n\to\infty} f(n)/g(n) = 0$.

Definition 6.1.4: Continuous mapping theorem (CMT)

Let f be continuous at every point $a \in C$ where $P(z \in C) = 1$. Then

1. If $\mathbf{z_n} \xrightarrow{p} \mathbf{z}$, then $f(\mathbf{z_n}) \xrightarrow{p} f(\mathbf{z})$
2. If $\mathbf{z_n} \xrightarrow{d} \mathbf{z}$, then $f(\mathbf{z_n}) \xrightarrow{d} f(\mathbf{z})$

Example. The CMT allows f to be discontinuous only if the probability of being at a discontinuity point is zero.
Consider $f(u) = u^{-1}$ is discontinuous at $u = 0$, but if $z_n \xrightarrow{d} z \sim N(0,1)$ then $P(z = 0) = 0$ so $z_n^{-1} \xrightarrow{d} z^{-1}$

Corollary 6.1.1 (Slutsky's theorem). If $z_n \xrightarrow{d} z$ and $c_n \xrightarrow{p} c$ as $n \to \infty$, then

1. $z_n + c_n \xrightarrow{d} z + c$
2. $z_n c_n \xrightarrow{d} zc$
3. $\frac{z_n}{c_n} \xrightarrow{d} \frac{z}{c}$ if $c \neq 0$.

The requirement that c_n converges to a constant is important. If it were to converge to a non-degenerate random variable, the theorem would be no longer valid. For example, let $z_n \sim \text{Uniform}(0,1)$ and $c_n = -z_n$. The sum $z_n + c_n = 0$ for all values of n. Moreover, $c_n \xrightarrow{d} c$ where $z \sim \text{Uniform}(0,1)$, $c \sim \text{Uniform}(-1,0)$, and z and c are independent.

Note:-

The theorem remains valid if we replace all convergences in distribution with convergences in probability.

Proof. This theorem follows from the fact that if z_n converges in distribution to z and c_n converges in probability to a constant c, then the joint vector (z_n, c_n) converges in distribution to (z, c).
Next we apply the continuous mapping theorem, recognising the functions $g(z,c)$ such as $g(z,c) = z + c$, $g(z,c) = zc$, and $g(z,c) = zc^{-1}$ are continuous (for the last function to be continuous, c has to be invertible). □

Definition 6.1.5: Khinchine's law of large numbers

If Y_i are i.i.d. with finite mean $\mathbb{E}Y_i = m < \infty$ then $\frac{1}{n}\sum_{i=1}^{n} Y_i \xrightarrow{p} m$

Lemma 6.1.1 (Markov's inequality). Let ξ be a non-negative random variable and let $\varepsilon > 0$ be a positive number. Then for any real number $p > 0$, the following inequality holds:

$$P(|\xi| \geq \varepsilon) \leq \frac{E[|\xi|^p]}{\varepsilon^p}.$$

Proof. Let ξ be a non-negative random variable and $\varepsilon > 0$. For any positive integer p:

$$\begin{aligned}
E[|\xi|^p] &= \int_0^\infty x^p f_\xi(x)\,dx && \text{(expectation definition)}\\
&= \int_0^\varepsilon x^p f_\xi(x)\,dx + \int_\varepsilon^\infty x^p f_\xi(x)\,dx && \text{(splitting the integral)}\\
&\geq \int_\varepsilon^\infty \varepsilon^p f_\xi(x)\,dx && \text{(since } x^p \geq \varepsilon^p \text{ for } x \geq \varepsilon)\\
&= \varepsilon^p P(|\xi| \geq \varepsilon) && \text{(definition of probability)}\\
P(|\xi| \geq \varepsilon) &\leq \frac{E[|\xi|^p]}{\varepsilon^p} && \text{(Markov's inequality)}
\end{aligned}$$

□

Lemma 6.1.2 (Chebyshev's inequality). Let η be a random variable with $\mathbb{E}[\eta] = m$ and $\mathrm{Var}(\eta) < \infty$. Then for any $\varepsilon > 0$,

$$P(|\eta - \mathbb{E}[\eta]| \geq \varepsilon) \leq \frac{\mathrm{Var}(\eta)}{\varepsilon^2}.$$

Proof. Using Markov's inequality, for any random variable η with finite expectation $E[\eta]$ and finite non-zero variance $\mathrm{Var}(\eta)$, and for any $\varepsilon > 0$, we have:

$$\begin{aligned}
P(|\eta - E[\eta]| \geq \varepsilon) &= P((\eta - E[\eta])^2 \geq \varepsilon^2) && \text{(squaring both sides)}\\
&\leq \frac{E[(\eta - E[\eta])^2]}{\varepsilon^2} && \text{(applying Markov's inequality)}\\
&= \frac{\mathrm{Var}(\eta)}{\varepsilon^2}. && \text{(variance definition)}
\end{aligned}$$

□

Definition 6.1.6: Chebyshev's law of large numbers

If Y_i are uncorrelated, and $\mathbb{E}Y_i = m < \infty$, $Var(Y_i) = \sigma_i^2 < \infty$ and $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 \to 0$, then $\frac{1}{n}\sum_{i=1}^n (Y_i - m) \xrightarrow{P} 0$

Proof. Let $Y_1, Y_2, \ldots, Y_n$ be uncorrelated random variables with $E[Y_i] = m$ and $\mathrm{Var}(Y_i) = \sigma_i^2 < \infty$. Assume that $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 \to 0$ as $n \to \infty$. Define $S_n = \frac{1}{n}\sum_{i=1}^n (Y_i - m)$. We want to show that $S_n \to 0$ in probability. By Chebyshev's inequality, for any $\varepsilon > 0$,

$$P(|S_n - E[S_n]| \geq \varepsilon) \leq \frac{\mathrm{Var}(S_n)}{\varepsilon^2}.$$

Since $E[S_n] = 0$ and the Y_i's are uncorrelated, we have

$$\mathrm{Var}(S_n) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^n (Y_i - m)\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(Y_i - m) = \frac{1}{n^2}\sum_{i=1}^n \sigma_i^2.$$

$$\Rightarrow P(|S_n| \geq \varepsilon) \leq \frac{1}{n^2}\frac{\sum_{i=1}^n \sigma_i^2}{\varepsilon^2}.$$

Since $\frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 \to 0$, it follows that for any $\varepsilon > 0$,

$$P(|S_n| \geq \varepsilon) \to 0 \text{ as } n \to \infty.$$

Hence, $S_n \to 0$ in probability. □

Definition 6.1.7: Univariate Lindeberg-Lévy Central Limit Theorem

If Y_i are i.i.d. random variables with finite mean m and variance σ^2, then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - m\right) \xrightarrow{d} N(0, \sigma^2)$$

Definition 6.1.8: Multivariate Lindeberg-Lévy Central Limit Theorem

If Y_i are i.i.d. with mean m and variance-covariance Σ, then

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i - m\right) \xrightarrow{d} N(0, \Sigma).$$

Proof. Set $\mathbf{c} \in \mathbb{R}^k$ with $\mathbf{c}'\mathbf{c} = 1$ and define $u_i = \mathbf{c}'(\mathbf{y}_i - \mathbf{m})$. The u_i are i.i.d. with $E(u_i^2) = \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c} < \infty$. By the univariate CLT,

$$\mathbf{c}'\sqrt{n}(\bar{\mathbf{y}} - \mathbf{m}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i \xrightarrow{d} N(0, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$$

Notice that if $\mathbf{z} \sim N(0, \boldsymbol{\Sigma})$ then $\mathbf{c}'\mathbf{z} \sim N(0, \mathbf{c}'\boldsymbol{\Sigma}\mathbf{c})$. Thus

$$\mathbf{c}'\sqrt{n}(\bar{\mathbf{y}} - \mathbf{m}) \xrightarrow{d} \mathbf{c}'\mathbf{z}.$$

Since this holds for all $\mathbf{c}$, we can use the Cramer-Wold device:

$$\sqrt{n}(\bar{\mathbf{y}} - \mathbf{m}) \xrightarrow{d} \mathbf{z} \sim N(0, \boldsymbol{\Sigma})$$

□

6.2 OLS in large samples

(OLS0)	(y_i, x_i) is an i.i.d. sequence		
(OLS1)	$E(x_i x_i')$ is finite non-singular	(GM1)	rank $\mathbf{X} = k$
(OLS2)	$E(y_i \vert x_i) = x_i'\beta$	(GM2)	$E(\mathbf{Y} \vert \mathbf{X}) = \mathbf{X}'\beta$
(OLS3)	$\text{Var}(y_i \vert x_i) = \sigma^2$	(GM3)	$\text{Var}(\mathbf{Y} \vert \mathbf{X}) = \sigma^2\mathbf{I}$
(OLS4)	$E\varepsilon_i^4 < \infty, \quad E\|x_i\|^4 < \infty$		

Remarks
(OLS0): Equivalent to random sampling, tells us that the pairs (x_i, y_i) are independent across i.
(OLS1): Ensures $\mathbf{X}'\mathbf{X}$ is invertible, or comparatively in sample $\frac{1}{n}\sum_{i=1}^n x_i x_i'$ exists.
(OLS2): Since all other x's are independent, this is equivalent to conditioning on all x's
(OLS3): Homoskedasticity and no serial correlation
(OLS4): Implies the existence of $\mathbb{E}(\varepsilon_i^2 x_i x_i')$ via Cauchy-Schwartz. This is required to use the CLT.

Lemma 6.2.1 (Expectation inequality)**.** For any random vector $Y \in \mathbb{R}^m$ with $\mathbb{E}\|Y\| < \infty$ then

$$\|\mathbb{E}[Y]\| \leq E\|Y\|$$

Lemma 6.2.2 (Holder's inequality)**.** If $p > 1$ and $q > 1$ and $\frac{1}{p} + \frac{1}{q} = 1$, then for any random $m \times n$ matrices X and Y,

$$(\mathbb{E}\|X'Y\|) \leq (\mathbb{E}\|X\|^p)^{1/p} (\mathbb{E}\|Y\|^q)^{1/q}$$

Corollary 6.2.1 (Cauchy-Schwartz inequality)**.** For any random $m \times n$ matrices X and Y,

$$(\mathbb{E}\|X'Y\|) \leq \left(\mathbb{E}\|X\|^2\right)^{1/2} \left(\mathbb{E}\|Y\|^2\right)^{1/2}$$

To see that the elements of $\mathbb{E}(\varepsilon_i^2 x_i x_i')$ are finite:

$$\begin{aligned}
\|\mathbb{E}(\varepsilon_i^2 x_i x_i')\| &\leq \mathbb{E}\|\varepsilon_i^2 x_i x_i'\| && \text{(using Lemma 6.2.1)}\\
&= \mathbb{E}(\varepsilon_i^2 \|x_i\|^2)\\
&\leq \mathbb{E}\left(\varepsilon_i^4\right)^{1/2} \mathbb{E}\left(\|x_i\|^4\right)^{1/2} && \text{(using Corollary 6.2.1)}\\
&< \infty && \text{(using OLS4)}
\end{aligned}$$

Theorem 6.2.1. Under OLS0-4:

1. $\hat{\beta}_{OLS} \xrightarrow{p} \beta$
2. $\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{d} \mathrm{N}\left(0, \sigma^2[\mathbb{E}(x_i x_i')]^{-1}\right)$

Proof. 1. We only require OLS0-2 for consistency[a]

$$\begin{aligned}
\hat{\beta}_{OLS} &= (X'X)^{-1}X'Y\\
&= \beta + (X'X)^{-1}X'\varepsilon\\
&= \beta + \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i
\end{aligned}$$

Since $x_i\varepsilon_i$ is i.i.d. by OLS0[b] we can use Khinchine's LLN

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} x_i x_i' \xrightarrow{p} \mathbb{E}(x_i x_i') \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i &\xrightarrow{p} \mathbb{E}(x_i \varepsilon_i)\\
&= \mathbb{E}(\mathbb{E}(x_i \varepsilon_i | x_i))\\
&= 0 \quad \text{(using OLS2)}
\end{aligned}$$

By the Continuous Mapping Theorem,

$$\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \xrightarrow{p} [\mathbb{E}(x_i x_i')]^{-1} \quad \text{(exists due to OLS1)}$$

$$\Rightarrow \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i \xrightarrow{p} 0$$

2.

$$\hat{\beta}_{OLS} - \beta = (X'X)^{-1}X'\varepsilon = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i \varepsilon_i$$

$$\Rightarrow \sqrt{n}\left(\hat{\beta}_{OLS} - \beta\right) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i \varepsilon_i$$

Using the CLT:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i \xrightarrow{d} N(0, Var(x_i\varepsilon_i)) = N\left(0, \sigma^2 \mathbb{E}(x_i x_i')\right)$$

Where the second equality follows from:

$$\begin{aligned}
\text{Var}(x_i\varepsilon_i) &= E[x_i\varepsilon_i\varepsilon_i' x_i'] - E[x_i\varepsilon_i]E[x_i\varepsilon_i]' \\
&= E[\varepsilon_i^2 x_i x_i'] - E[x_i\varepsilon_i]E[x_i\varepsilon_i]' \quad \text{(since } \varepsilon_i \text{ scalar)} \\
&= E[E(\varepsilon_i^2 x_i x_i'|x_i)] - E[E(x_i\varepsilon_i|x_i)]E[x_i\varepsilon_i]' \quad \text{(first expectation exists by OLS4)} \\
&= E[E(\varepsilon_i^2|x_i)x_i x_i'] - E[x_i E(\varepsilon_i|x_i)]E[x_i\varepsilon_i]' \\
&= \sigma^2 E[x_i x_i']. \quad \text{(using OLS2)}
\end{aligned}$$

Using the CMT:

$$\begin{aligned}
\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i &\xrightarrow{d} [\mathbb{E}(x_i x_i')]^{-1} N\left(0, \sigma^2\mathbb{E}(x_i x_i')\right) \\
&\sim N\left(0, [\mathbb{E}(x_i x_i')]^{-1}\sigma^2\mathbb{E}(x_i x_i')[\mathbb{E}(x_i x_i')]^{-1}\right) \\
\sqrt{n}\left(\hat{\beta}_{OLS} - \beta\right) &\xrightarrow{d} N\left(0, \sigma^2[\mathbb{E}(x_i x_i')]^{-1}\right)
\end{aligned}$$

□

[a] Strictly we only need OLS0,1,2': $\mathbb{E}(x_i\varepsilon_i) = 0$

[b] $x_i\varepsilon_i = x_i(y_i - x_i'\beta)$ and we know (y_i, x_i) i.i.d.

Theorem 6.2.2. Under OLS0-4:

1. $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$
2. $W \xrightarrow{d} \chi^2(p)$
3. $t \xrightarrow{d} N(0,1)$

Proof. 1.[a]

$$\begin{aligned}
\hat{\sigma}^2 &= \frac{1}{n-k}\varepsilon' M_X \varepsilon \\
&= \frac{1}{n-k}\varepsilon'(I - X(X'X)^{-1}X')\varepsilon \\
&= \frac{1}{n-k}\varepsilon'\varepsilon - \frac{1}{n-k}\varepsilon' X(X'X)^{-1}X'\varepsilon \\
&= \frac{n}{n-k}\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 - \frac{n}{n-k}\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i \left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} x_i'\varepsilon_i
\end{aligned}$$

Using Khinchine's LLN:

$$\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 \xrightarrow{p} \mathbb{E}[\varepsilon_i^2], \quad \frac{1}{n}\sum_{i=1}^{n}x_i\varepsilon_i \xrightarrow{p} \mathbb{E}[x_i\varepsilon_i]=0, \quad \frac{1}{n}\sum_{i=1}^{n}x_ix_i' \xrightarrow{p} \mathbb{E}(x_ix_i'), \quad \frac{1}{n}\sum_{i=1}^{n}x_i'\varepsilon_i \xrightarrow{p} \mathbb{E}[x_i'\varepsilon_i]=0$$

Using CMT and Slutsky:

$$\begin{aligned}\hat{\sigma}^2 &\xrightarrow{p} \frac{n}{n-k}\mathbb{E}[\varepsilon_i^2] + \frac{n}{n-k}\times 0\\ &= \mathbb{E}[\varepsilon_i^2] \quad (\text{as } n\to\infty)\\ &= \sigma^2\end{aligned}$$

2.

$$W = \frac{\sqrt{n}\left(R\hat{\beta}-q\right)'\left(\sigma^2 R\left(\frac{1}{n}X'X\right)^{-1}R'\right)^{-1}\sqrt{n}\left(R\hat{\beta}-q\right)}{\hat{\sigma}^2/\sigma^2}$$

We have seen that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$, and

$$\begin{aligned}\sqrt{n}\left(R\hat{\beta}-q\right) &= \sqrt{n}(\hat{\beta}-\beta) \quad (\text{since } H_0: R\beta = q)\\ &\xrightarrow{d} RN\left(0,\sigma^2[\mathbb{E}(x_ix_i')]^{-1}\right)\\ &= N\left(0,\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)\\ &= \left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{1/2}N(0,I_p)\end{aligned}$$

Since $\frac{1}{n}X'X \xrightarrow{p} \mathbb{E}[x_ix_i']$, by the CMT,

$$\left(\sigma^2R\left(\frac{1}{n}X'X\right)^{-1}R'\right)^{-1} \xrightarrow{p} \left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{-1}$$

$$\begin{aligned}\Rightarrow W &\xrightarrow{d} \frac{\left(\left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{1/2}N(0,I_p)\right)'\left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{-1}\left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{-1/2}N(0,I_p)}{1}\\ &= (N(0,I_p))'\left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{1/2}\left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{-1}\left(\sigma^2R[\mathbb{E}(x_ix_i')]^{-1}R'\right)^{1/2}N(0,I_p)\\ &= (N(0,I_p))'I_pN(0,I_p)\\ &= \chi^2(p)\end{aligned}$$

3.

$$\begin{aligned}t &= \frac{\hat{\beta}_j-\beta}{\sqrt{\hat{\sigma}^2(X'X)^{-1}_{jj}}}\\ &= \frac{(\hat{\beta}_j-\beta)/\sqrt{\sigma^2(X'X)^{-1}_{jj}}}{\sqrt{\hat{\sigma}^2/\sigma^2}}\\ &= \frac{\xrightarrow{d} N(0,1)}{\sqrt{\xrightarrow{p} 1}} \quad (\hat{\sigma}^2 \xrightarrow{p} \sigma^2 \text{ and Theorem 6.2.1-2})\\ &\xrightarrow{d} N(0,1) \quad (\text{by Slutsky})\end{aligned}$$

□

[a] See Lecture 5 for derivation of the first step

The distribution of the Wald statistic is as expected, recall $W/p|x \sim F(p, n-k)$ under normal regression, and thus we see $W|x \sim pF(p, n-k) \xrightarrow{d} \chi^2(p)$. Why?

$$\begin{aligned} p \times F &= p\frac{\chi^2(p)/p}{\chi^2(n-k)/(n-k)} \\ &= \frac{\chi^2(p)}{\chi^2(n-k)/(n-k)} \end{aligned}$$

$$\frac{\chi^2(n-k)}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n-k} Z_i^2 \xrightarrow{p} \mathbb{E}[Z_i^2] = 1$$

$$\Rightarrow pF \xrightarrow{d} \chi^2(p)$$

Asymptotic confidence intervals and sets

Since $t \xrightarrow{d} N(0,1)$ we can build asymptotic confidence intervals for β_j. From the critical values of $N(0,1)$:

$$Pr\left(\left|\frac{\sqrt{n}(\hat{\beta}_j - \beta)}{\sqrt{\hat{\sigma}^2(\frac{1}{n}X'X)^{-1}_{jj}}}\right| \leq 1.96\right) \approx 0.95$$

$$\Rightarrow Pr\left(\left|\hat{\beta}_j - \beta\right| \leq 1.96\sqrt{\hat{\sigma}^2(X'X)^{-1}_{jj}}\right) \approx 0.95 \quad \text{(cancel n's and rearrange)}$$

$$\Rightarrow \left[\hat{\beta}_j - 1.96\sqrt{\hat{\sigma}^2(X'X)^{-1}_{jj}}, \hat{\beta}_j + 1.96\sqrt{\hat{\sigma}^2(X'X)^{-1}_{jj}}\right] \quad \text{Asymptotic confidence interval}$$

This gives us the set all all values of β_j that are not rejected by he t-test with asymptotic size 5%. We say that the confidence interval is obtained by inversion of the test. We can similarly invert the Wald test, consider a test of the entire vector $\beta = b$ (i.e. $R = I_p$):

$$\begin{aligned} W &= (\hat{\beta} - b)'(\hat{\sigma}^2(X'X^{-1}))^{-1}(\hat{\beta} - b) \\ &= \frac{(\hat{\beta} - b)'X'X(\hat{\beta} - b)}{\hat{\sigma}^2} \end{aligned}$$

The asymptotic 95% confidence set for β is the ellipsoid with centre $\hat{\beta}$:

$$\Rightarrow \left\{\beta : \frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{\hat{\sigma}^2} \leq \chi^2_{0.95}(k)\right\}$$

6.3 Delta method

Sometimes we need to know confidence intervals or sets for some (possibly nonlinear) function of regression parameters. We can do this with the delta method.

Definition 6.3.1: Delta method

Suppose $\hat{\theta}$ is a k-dimensional vector where $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \xi$, and suppose $g : \mathbb{R}^k \to \mathbb{R}$ has continuous first dervatives. Denote by $G(\theta)$ the $r \times k$ matrix of first derivatives evaluated at θ: $G(\theta) \equiv \frac{\partial g(\theta)}{\partial \theta'}$ then as $n \to \infty$

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} G(\theta)\xi$$

. In particular, if $\xi \sim N(0, V)$ then as $n \to \infty$

$$\sqrt{n}(g(\hat{\theta}) - g(\theta)) \xrightarrow{d} N(0, GVG')$$

Proof. By the mean value theorem, there exists a k-dimensional vector $\bar{\theta}$ between $\hat{\theta}$ and θ such that

$$g(\hat{\theta}) - g(\theta) = \underset{r\times k}{G(\bar{\theta})}\underset{k\times 1}{(\hat{\theta}-\theta)}$$

$$\Rightarrow \sqrt{n}(g(\hat{\theta}) - g(\theta)) = G(\bar{\theta})\sqrt{n}(\hat{\theta}-\theta)$$

Since $\bar{\theta}$ is between $\hat{\theta}$ and θ and since $\hat{\theta} \xrightarrow{p} \theta$ we know $\bar{\theta} \xrightarrow{p} \theta$. $G()$ is assumed continuous, so by CMT:

$$G(\bar{\theta}) \xrightarrow{p} G(\theta)$$

$$\Rightarrow \sqrt{n}(g(\hat{\theta}) - g(\theta)) = G(\bar{\theta})\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{p} G(\theta)\xi$$

□

Exercise 6.3.1. Let $\{\hat{\theta}_n\}$ be a sequence of 2x1 random vectors satisfying $\sqrt{n}(\hat{\theta}_0-\theta_0) \xrightarrow{d} N(0,V)$ where the asymptotic mean is $\theta_0 = [0,1]'$ and the asymptotic covariance matrix is I_2. Denote the two entries of $\hat{\theta}_n$ by $\hat{\theta}_{n,1}$ and $\hat{\theta}_{n,2}$. Derive the asymptotic distribution of the sequence of products $\{\hat{\theta}_{n,1}\hat{\theta}_{n,2}\}$

Solution:-

We can apply the delta method because the function

$$g(\theta) = g(\theta_1,\theta_2) = \theta_1\theta_2$$

is continuously differentiable. The asymptotic mean of the transformed sequence is

$$g(\theta_0) = \theta_{0,1}\theta_{0,2} = 0 \times 1 = 0$$

The Jacobian of the function is

$$G(\theta) = \begin{bmatrix} \frac{\partial g(\theta_1,\theta_2)}{\partial\theta_1} & \frac{\partial g(\theta_1,\theta_2)}{\partial\theta_2} \end{bmatrix} = [\theta_2, \theta_1]$$

By evaluating at θ_0 we obtain $G(\theta_0) = [1,0]$.
Therefore the asymptotic covariance matrix is

$$G(\theta_0)VG(\theta_0)' = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 1$$

And we can write $\sqrt{n}\hat{\theta}_{n,1}\hat{\theta}_{n,2} \xrightarrow{d} N(0,1)$

Example (Nerlove's returns to scale).

$$\log TC_i = \beta_1 + \beta_2 \log Q_i + \beta_3 \log p_{C_i} + \beta_4 \log p_{L_i} + \beta_5 \log p_{F_i} + \varepsilon_i$$

Suppose we want to study the asymptotic confidence region of the normalised regression with coefficients $\alpha = (\beta_3/\beta_2, \beta_4/\beta_2, \beta_5/\beta_2)'$ (i.e. the powers of the Cobb-Douglas production function). Define

$$g(\beta) = \begin{bmatrix} \beta_3/\beta_2 \\ \beta_4/\beta_2 \\ \beta_5/\beta_2 \end{bmatrix}$$

$$G(\beta) = \frac{\partial g(\beta)}{\partial \theta'} = \begin{bmatrix} 0 & -\beta_3/\beta_2^2 & 1/\beta_2 & 0 & 0 \\ 0 & -\beta_4/\beta_2^2 & 0 & 1/\beta_2 & 0 \\ 0 & -\beta_5/\beta_2^2 & 0 & 0 & 1/\beta_2 \end{bmatrix}$$

Thus considering the Wald statistic with $H_0 : \hat{\alpha} = \alpha$, i.e.: $R = I_3, q = \alpha$:

$$\begin{aligned} W &= \frac{\left(R\hat{\beta} - q\right)' \left(\sigma^2 R \left(X'X\right)^{-1} R'\right)^{-1} \left(R\hat{\beta} - q\right)}{\hat{\sigma}^2/\sigma^2} \\ &= \frac{\sqrt{n}\left(\hat{\alpha} - \alpha\right)' \left(\sigma^2 R \left(\frac{1}{n} X'X\right)^{-1} R'\right)^{-1} \sqrt{n}\left(\hat{\alpha} - \alpha\right)}{\hat{\sigma}^2/\sigma^2} \\ &\xrightarrow{d} [N(0, I_3)]' I_3 N(0, I_3) \quad \text{(using theorem 6.2.2)} \\ &= \chi^2(3) \end{aligned}$$

Hence the asymptotic 95% confidence set for α is the ellipsoid

$$\left\{\alpha : (\hat{\alpha} - \alpha)' \left(G(\hat{\beta})\hat{\sigma}^2 (X'X)^{-1} G(\hat{\beta})'\right) (\hat{\alpha} - \alpha) \leq \chi^2_{0.95}(3)\right\}$$

6 Heteroskedasticity and serial correlation. HAC standard errors.

The homoskedasticity and no serial correlation assumption (GM3) can be violated in three ways:

- Heteroskedasticity only (B)- $Var(\varepsilon|X)$ is diagonal with unequal elements along the diagonal.
- Serial correlation only (C) - $Var(\varepsilon|X)$ has non-zero off-diagonal elements, but all diagonal elements are the same.
- Heteroskedasticity and serial correlation (D) - $Var(\varepsilon|X)$ is a general non-diagonal matrix with unequal elements along the diagonal.

$$A=\sigma^2\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}\quad B=\sigma^2\begin{bmatrix}1&0&0\\0&2&0\\0&0&3\end{bmatrix}\quad C=\sigma^2\begin{bmatrix}1&\rho&\rho^2\\\rho&1&\rho\\\rho^2&\rho&1\end{bmatrix}\quad D=\sigma^2\begin{bmatrix}1&\rho&\rho^2\\\rho&2&\rho\\\rho^2&\rho&3\end{bmatrix}$$

6.1 Heteroskedasticity

Under heteroskedasticity OLS is still consistent and asymptotically normal, although is no longer efficient and has a different asymptotic covariance matrix. Thus the default standard errors will be wrong. Recall the large sample OLS assumptions, now consider the weaker assumptions OLS2' and OLS3'

(OLS0) (y_i, x_i) is an i.i.d. sequence
(OLS1) $E(x_i x_i')$ is finite non-singular
(OLS2) $E(y_i|x_i) = x_i'\beta$
(OLS3) $\text{Var}(y_i|x_i) = \sigma^2$
(OLS4) $E\varepsilon_i^4 < \infty, \quad E\|x_i\|^4 < \infty$

(OLS2') $E(\varepsilon_i x_i) = 0$
(OLS3') $\text{Var}(\varepsilon_i x_i) = V < \infty$ and is non-singular

Theorem 6.1.1. Under OLS0,1,2',3',4

1. $\hat\beta_{OLS} \xrightarrow{p} \beta$ (OLS is consistent)
2. $\sqrt{n}\left(\hat\beta_{OLS} - \beta\right) \xrightarrow{d} N\left(0, (\mathbb{E}(x_i x_i'))^{-1} V (\mathbb{E}(x_i x_i'))^{-1}\right)$

Proof. 1. We only require OLS0,1,2' for consistency

$$\begin{aligned}\hat\beta_{OLS} &= \beta + (X'X)^{-1}X'\varepsilon\\ &= \beta + \left(\frac{1}{n}\sum_{i=1}^n x_i x_i'\right)^{-1}\frac{1}{n}\sum_{i=1}^n x_i\varepsilon_i\\ &\xrightarrow{p} \beta + [\mathbb{E}(x_i x_i')]^{-1}\mathbb{E}(\varepsilon_i x_i)\\ &= \beta\end{aligned}$$

2.

$$\sqrt{n}\left(\hat{\beta}_{OLS}-\beta\right)=\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\varepsilon_i$$

Using the CLT:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\varepsilon_i\xrightarrow{d}N(0,V)$$

Using the CMT:

$$\left(\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}x_i\varepsilon_i\xrightarrow{d}[\mathbb{E}(x_ix_i')]^{-1}N(0,V)$$
$$\Rightarrow\sqrt{n}\left(\hat{\beta}_{OLS}-\beta\right)\xrightarrow{d}N\left(0,(\mathbb{E}(x_ix_i'))^{-1}V(\mathbb{E}(x_ix_i'))^{-1}\right)$$

□

When the errors are homoskedastic the variance is as in previous lectures:

$$\mathbb{E}[X'X]^{-1}\mathbb{E}[X'X\varepsilon_i^2]\mathbb{E}[X'X]^{-1}=\mathbb{E}[X'X]^{-1}\sigma^2\mathbb{E}[X'X]\mathbb{E}[X'X]^{-1}=\sigma^2\mathbb{E}[X'X]^{-1}$$

The classic covariance matrix estimator can be highly biased if homoskedasticity fails, we now consider how to construct covariance matrix estimators which do not require homoskedasticity. If ε_i were known, we could have estimated V as follows:

$$\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\hat{\varepsilon}_i^2\xrightarrow{p}V$$

Of course ε_i is unknown, but since $\hat{\beta}_{OLS}$ remains consistent we can use the observed residuals $\hat{\varepsilon}_i=Y_i-x_i'\hat{\beta}_{OLS}$:

$$\hat{V}=\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\hat{\varepsilon}_i^2$$

To show this is a consistent estimator:

$$\begin{aligned}\hat{V}&=\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\hat{\varepsilon}_i^2\\&=\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\left(\varepsilon_i-x_i'\left(\hat{\beta}_{OLS}-\beta\right)\right)^2\\&=\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\left(\varepsilon_i^2-2\varepsilon_ix_i'\left(\hat{\beta}_{OLS}-\beta\right)+\left(x_i'\left(\hat{\beta}_{OLS}-\beta\right)\right)^2\right)\\&=\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\varepsilon_i^2-\frac{2}{n}\sum_{i=1}^{n}(x_ix_i')\varepsilon_ix_i'\left(\hat{\beta}_{OLS}-\beta\right)+\frac{1}{n}\sum_{i=1}^{n}x_ix_i'\left(x_i'\left(\hat{\beta}_{OLS}-\beta\right)\right)^2\\&\xrightarrow{p}V\quad\text{since }\hat{\beta}_{OLS}\xrightarrow{p}\beta\end{aligned}$$

Definition 6.1.1: White's heteroskedasticity robust covaraince matrix

$$\widehat{Var}(\hat{\beta}_{OLS})=(X'X)^{-1}\left(\sum_{i=1}^{n}x_ix_i'\hat{\varepsilon}_i^2\right)(X'X)^{-1}$$

Note:-

Whilst this estimator is consistent, it is biased in finite samples. To see this, suppose the actual covariance matrix of the population regression residuals is given by $\mathbb{E}[\varepsilon\varepsilon'|X] = \Phi = diag(\phi_i)$. The covariance matrix of the OLS estimator is then

$$V = (X'X)^{-1}(X'\Phi X)(X'X)^{-1}$$

Denote the i-th column of the residual maker matrix M by m_i then $\hat{\varepsilon}_i = m_i'\varepsilon$.

$$\Rightarrow \mathbb{E}[\hat{\varepsilon}_i^2] = \mathbb{E}[m_i'\varepsilon\varepsilon' m_i] = m_i'\Phi m_i$$

Notice that m_i is the i-th column of the identity matrix (denoted as e_i) minus the i-th column of the projection matrix $X(X'X)^{-1}X'$ (p_i). Hence $m_i = e_i - p_i$ and

$$\mathbb{E}[\hat{\varepsilon}_i^2] = (e_i - h_i)'\Phi(e_i - h_i) = \phi_i - 2\phi_i h_{ii} + h_i'\Phi h_i$$

where h_{ii} is the i-th diagonal element of the projection matrix. Because this matrix is symmetric and idempotent, $h_{ii} = h_i'h_i$ so:

$$\begin{aligned}\mathbb{E}\left(\hat{V} - V\right) &= (X'X)^{-1}(X'\Phi X)(X'X)^{-1} - (X'X)^{-1}(X'\hat{\Phi}X)(X'X)^{-1} \\ &= (X'X)^{-1}(X'(\Phi - \hat{\Phi})X)(X'X)^{-1} \\ &= (X'X)^{-1}(X'diag(\phi_i - (\phi_i - 2\phi_i h_{ii} + h_i'\Phi h_i))X)(X'X)^{-1} \\ &= (X'X)^{-1}(X'diag(h_i'(\Phi - 2\phi_i I)h_i)X)(X'X)^{-1}\end{aligned}$$

Whilst $\hat{V}$ is biased, here we can see that it is also consistent. Notice that $\hat{\Phi}$ is not consistent for Φ, since there are more elements to estimate as the sample gets large. However, $\hat{\varepsilon}_i$ is consistent for ε_i. We know

$$X'\hat{\Phi}X = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' \hat{\varepsilon}_i^2$$

and since plim $\hat{\varepsilon}_i^2 = \phi_i$ we get plim $X'\hat{\Phi}X = X'\Phi X$.
In summary, $\hat{V}$ is biased since $\mathbb{E}(\hat{\varepsilon}_i^2)$ is a biased estimate of ε.

6.2 Serial correlation (and heteroskedasticity)

As with heteroskedasticity, OLS remains consistent and asymptotically normal, but the default standard errors are wrong. This cannot happen if the data are i.d.d. - if OLS0 holds it must be the case that $\Omega = Var(\varepsilon|X)$ is diagonal. If the data are dependent, then Ω is typically no longer diagonal.

Definition 6.2.1: Strict Stationarity

A sequence of random variables $\{Z_t\}_{t=-\infty}^{\infty}$ is strictly stationary if, for any finite nonnegative integer m,

$$f_{Z_t,Z_{t+1},\ldots,Z_{t+m}}(x_0, x_1, \ldots, x_m) = f_{Z_s,Z_{s+1},\ldots,Z_{s+m}}(x_0, x_1, \ldots, x_m)$$

which is to say that the joint distribution , f, does not depend on the index, t.

Strict stationarity implies that the (marginal) distribution of Z_t does not vary over time. It also implies that the bivariate distributions of(Z_t, Z_{t+1}) and multivariate distributions of $(Z_t, \ldots, Z_{t+m})$ are stable over time.

Theorem 6.2.1. If Z_t is i.i.d., then it is strictly stationary

Proof. Let F denote the joint distribution function, then:

$$\begin{aligned} F(x_{n+1},...,x_{n+m}) &= F(x_{n+1}) \cdot \dots \cdot F(x_{n+m}) \\ &= F(x_{n+k+1}) \cdot \dots \cdot F(x_{n+k+m}) \\ &= F(x_{n+k+1},\dots,x_{n+k+m}) \end{aligned}$$

Lines 1 and 3 follow from the fact that the joint distribution function of a set of mutually independent variables is equal to the product of their marginal distribution functions. On line 2 we have used the fact that all the terms of the sequence have the same distribution. □

Definition 6.2.2: Covarariance stationarity

A sequence of random variables $\{Z_t\}_{t=-\infty}^{\infty}$ is covariance (weakly) stationary if just the first two moments do not depend on t, e.g.

$$\begin{gathered} \mathbb{E}Z_1 = \mathbb{E}Z_2 = \dots \\ Var(Z_1) = Var(Z_2) = \dots \\ Cov(Z_1, Z_{1+m}) = Cov(Z_2, Z_{2+m}) = \dots \end{gathered}$$

A strictly stationary process is covariance-stationary as long as the variance and covariances are finite.
Consider a new set of OLS assumptions:

(SC0) $\{(y_t, x_t)\}_{t=1}^{T}$ is strictly stationary

(SC1) $\{(x_t x_t')\}$ satisfies LLN: $\frac{1}{T}\sum x_t x_t' \xrightarrow{p} \mathbb{E}(x_t x_t') < \infty$, positive definite

(SC2) $\{(x_t \varepsilon_t)\}$ satisfies LLN: $\frac{1}{T}\sum x_t \varepsilon_t \xrightarrow{p} \mathbb{E}(x_t \varepsilon_t) = 0$

(SC3) $\{(x_t \varepsilon_t)\}$ satisfies CLT: $\frac{1}{\sqrt{T}}\sum x_t \varepsilon_t \xrightarrow{d} N(0, V)$, where

$$V = \mathbb{E}(\varepsilon_t^2 x_t x_t') + \sum_{t=1}^{\infty} \left(\mathbb{E}(\varepsilon_t \varepsilon_{t-l} x_t x_{t-l}') + \mathbb{E}(\varepsilon_t \varepsilon_{t-l} x_{t-l} x_t') \right)$$

These assumptions further generalise our GM/OLS conditions, such that if the data were independent, we would have $V = \mathbb{E}(\varepsilon_t^2 x_t x_t')$ as in OLS3'.

Theorem 6.2.2. Under SC0,1,2,3

1. $\hat{\beta}_{OLS} \xrightarrow{p} \beta$ (OLS is consistent)
2. $\sqrt{T}\left(\hat{\beta}_{OLS} - \beta\right) \xrightarrow{d} N\left(0, (\mathbb{E}(x_t x_t'))^{-1} V (\mathbb{E}(x_t x_t'))^{-1}\right)$

The proof is identical to the heteroskedastic case in Theorem 6.2.1.

Newey-West Method
Under the SC assumptions, the conventional covariance matrix estimators are inconsistent as they do not capture the serial dependence in $x_t e_t$. To consistently estimate the covariance matrix, we need a different estimator. The appropriate class of estimators are called Heteroskedasticity and Autocorrelation Consistent (HAC) covariance matrix estimators.

Define V_T as follows:

$$
\begin{aligned}
V_T &\equiv Var\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} x_t\varepsilon_t\right)\\
&= \mathbb{E}\left[\frac{1}{T}\left(\sum_{t=1}^{T} x_t\varepsilon_t\right)\left(\sum_{t=1}^{T} x_t\varepsilon_t\right)'\right]\\
&= \mathbb{E}\left[\frac{1}{T}(x_1\varepsilon_1 + x_2\varepsilon_2 + \cdots + x_T\varepsilon_T)(x_1'\varepsilon_1 + x_2'\varepsilon_2 + \cdots + x_T'\varepsilon_T)'\right]\\
&= \mathbb{E}\left[\underbrace{\frac{1}{T}\sum_{t=1}^{T}\varepsilon_t^2 x_t x_t' +}_{\text{variance at t}} \underbrace{\frac{1}{T}\sum_{\ell=1}^{T-1}\sum_{t=\ell+1}^{T}\left(\varepsilon_t\varepsilon_{t-\ell}x_t x_{t-\ell}' + \varepsilon_t\varepsilon_{t-\ell}x_{t-\ell}x_t'\right)}_{\text{all covariances with all other time periods}}\right]\\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\varepsilon_t^2 x_t x_t'] + \frac{1}{T}\sum_{\ell=1}^{T-1}\sum_{t=\ell+1}^{T}\left(\mathbb{E}(\varepsilon_t\varepsilon_{t-\ell}x_t x_{t-\ell}') + \mathbb{E}(\varepsilon_t\varepsilon_{t-\ell}x_{t-\ell}x_t')\right)\\
&= \mathbb{E}[\varepsilon_t^2 x_t x_t'] + \sum_{\ell=1}^{T-1}\frac{T-\ell}{T}\left(\mathbb{E}(\varepsilon_t\varepsilon_{t-\ell}x_t x_{t-\ell}') + \mathbb{E}(\varepsilon_t\varepsilon_{t-\ell}x_{t-\ell}x_t')\right) \quad \text{Using SC0}
\end{aligned}
$$

As T get large, $V_T \approx V$. Since we have T data points, we can only estimate $G < T$ autocovariances of $x_t\varepsilon_t$, where G is the truncation lag. Newey and West propose the following procedure:

1. Choose G such that: $G = O(T^\alpha)$ for $0 < \alpha < 1/4$
2. Estimate autocovariances of $x_t\varepsilon_t$ of order ℓ by
$$\hat{\Gamma}_\ell = \frac{1}{T}\sum_{t=\ell+1}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-\ell}x_t x_{t-\ell}'$$
3. Estimate V by
$$\hat{V}_{nw} = \hat{\Gamma}_0 + \sum_{\ell=1}^{G}\frac{G+1-\ell}{G+1}\left(\hat{\Gamma}_\ell + \hat{\Gamma}_\ell'\right)$$

If we know a priori that autocovariances are zero in population beyond a certain finite lag q, we can consistently estimate V with

$$\hat{V} = \hat{\Gamma}_0 + \sum_{\ell=1}^{q}\left(\hat{\Gamma}_\ell + \hat{\Gamma}_\ell'\right)$$

However in the case where we do not know q (which is potentially infinite), we can use the weighted sum suggested by Newey and West. For example, for $q(n) = 3$

$$\hat{V}_{NW} = \hat{\Gamma}_0 + \frac{2}{3}(\hat{\Gamma}_1 + \hat{\Gamma}_1') + \frac{1}{3}(\hat{\Gamma}_2 + \hat{\Gamma}_2')$$

The weighting term ensures $\hat{V}_{nw}$ is positive semi-definite. We can see the similarities between this and our expression for V_T earlier, giving some intuition for its consistency.

$$
\begin{aligned}
V_T &= \mathbb{E}[\varepsilon_t^2 x_t x_t'] \quad + \sum_{\ell=1}^{T-1}\frac{T-\ell}{T}\left[\quad \mathbb{E}(\varepsilon_t\varepsilon_{t-\ell}x_t x_{t-\ell}') \quad + \quad \mathbb{E}(\varepsilon_t\varepsilon_{t-\ell}x_{t-\ell}x_t') \quad\right]\\
\hat{V}_{nw} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2 x_t x_t' + \sum_{\ell=1}^{G}\frac{G+1-\ell}{G+1}\left[\frac{1}{T}\sum_{t=\ell+1}^{T}\left(\varepsilon_t\varepsilon_{t-\ell}x_t x_{t-\ell}'\right) + \frac{1}{T}\sum_{t=\ell+1}^{T}\left(\varepsilon_t\varepsilon_{t-\ell}x_{t-\ell}x_t'\right)\right]
\end{aligned}
$$

Now we can estimate the covariance matrix of $\hat{\beta}_{OLS}$ as

$$\frac{1}{T}\left[\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right]^{-1} \hat{V}_{nw} \left[\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right]^{-1}$$

Lemma 6.2.1. The matrix of sample covariances for any process is positive semi-definite.

Proof. Let $z_1, \ldots, z_T$ be any sequence of T numbers, and let P be a $m \times m$ matrix of sample covariances:

$$P = \begin{bmatrix} \frac{1}{T}\sum_{l=1}^{T} z_t^2 & \frac{1}{T}\sum_{l=2}^{T} z_t z_{t-1} & \ddots & \sum_{l=m+1}^{T} z_t z_{t-m} \\ \frac{1}{T}\sum_{l=2}^{T} z_t z_{t-1} & \frac{1}{T}\sum_{l=1}^{T} z_t^2 & \ddots & \ddots \\ \ddots & \ddots & \ddots & \frac{1}{T}\sum_{l=2}^{T} z_t z_{t-1} \\ \sum_{l=m+1}^{T} z_t z_{t-m} & \ddots & \frac{1}{T}\sum_{l=2}^{T} z_t z_{t-1} & \frac{1}{T}\sum_{l=1}^{T} z_t^2 \end{bmatrix}$$

Consider the $m \times (2T-1)$ matrix:

$$Z = \begin{bmatrix} z_1 & z_2 & \cdots & z_m & \cdots & z_T & 0 & \cdots & 0 \\ 0 & z_1 & z_2 & \cdots & z_m & \cdots & z_T & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & z_1 & z_2 & \cdots & \cdots & \cdots & z_T \end{bmatrix}$$

$$\frac{1}{T}ZZ' = \frac{1}{T} \underbrace{\begin{bmatrix} z_1 & z_2 & \cdots & z_m & \cdots & z_T & 0 & \cdots & 0 \\ 0 & z_1 & z_2 & \cdots & z_m & \cdots & z_T & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & z_1 & z_2 & \cdots & \cdots & \cdots & z_T \end{bmatrix}}_{m\times(2T-1)} \underbrace{\begin{bmatrix} z_1 & 0 & \cdots & 0 \\ z_2 & z_1 & \cdots & \vdots \\ \vdots & z_2 & \ddots & 0 \\ z_m & \vdots & \ddots & z_1 \\ \vdots & z_m & \ddots & z_2 \\ z_T & \vdots & \ddots & \vdots \\ 0 & z_T & \cdots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & z_T \end{bmatrix}}_{(2T-1)\times m}$$

$$= \frac{1}{T}\begin{bmatrix} \sum_{l=1}^{T} z_t^2 & \sum_{l=2}^{T} z_t z_{t-1} & \ddots & \sum_{l=m+1}^{T} z_t z_{t-m} \\ \sum_{l=2}^{T} z_t z_{t-1} & \sum_{l=1}^{T} z_t^2 & \ddots & \ddots \\ \ddots & \ddots & \ddots & \sum_{l=2}^{T} z_t z_{t-1} \\ \sum_{l=m+1}^{T} z_t z_{t-m} & \ddots & \sum_{l=2}^{T} z_t z_{t-1} & \sum_{l=1}^{T} z_t^2 \end{bmatrix} = P_{m\times m}$$

Thus P is p.s.d. since for any vector $v, v'ZZ'v = u'u = \sum_{i=1}^{m} u_i^2 \geq 0$ □

Theorem 6.2.3. $\hat{V}_{nw}$ is positive semi-definite

Proof. Let c be any deterministic k-dimensional vector, we aim to show $c'\hat{V}_{nw}c \geq 0$. Consider the $G+1$ matrix

$$P = \begin{bmatrix} c'\hat{\Gamma}_0 c & c'\hat{\Gamma}_1 c & \ddots & c'\hat{\Gamma}_G c \\ c'\hat{\Gamma}_1 c & c'\hat{\Gamma}_0 c & \ddots & \ddots \\ \ddots & \ddots & \ddots & c'\hat{\Gamma}_1 c \\ c'\hat{\Gamma}_G c & \ddots & c'\hat{\Gamma}_1 c & c'\hat{\Gamma}_0 c \end{bmatrix}$$

If i is a $G+1$-dimensional vector of ones, then we have

$$\begin{aligned} i'Pi &= \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} c'\hat{\Gamma}_0 c & c'\hat{\Gamma}_1 c & \ddots & c'\hat{\Gamma}_G c \\ c'\hat{\Gamma}'_1 c & c'\hat{\Gamma}_0 c & \ddots & \ddots \\ \ddots & \ddots & \ddots & c'\hat{\Gamma}_1 c \\ c'\hat{\Gamma}'_G c & \ddots & c'\hat{\Gamma}'_1 c & c'\hat{\Gamma}_0 c \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \sum_{\ell=0}^{G} c'\hat{\Gamma}_\ell c \\ \sum_{\ell=1}^{1} c'\hat{\Gamma}'_\ell c + \sum_{\ell=0}^{G-1} c'\hat{\Gamma}_\ell c \\ \vdots \\ \sum_{\ell=0}^{G} c'\hat{\Gamma}'_\ell c \end{bmatrix} \qquad m\text{-th row} = \sum_{\ell=1}^{m} c'\hat{\Gamma}'_\ell c + \sum_{\ell=0}^{G-m} c'\hat{\Gamma}_\ell c \\ &= \sum_{\ell=0}^{G} c'\hat{\Gamma}_\ell c + \sum_{\ell=1}^{1} c'\hat{\Gamma}'_\ell c + \sum_{\ell=0}^{G-1} c'\hat{\Gamma}_\ell c + \dots + \sum_{\ell=0}^{G} c'\hat{\Gamma}'_\ell c \\ &= (G+1)c'\hat{\Gamma}_0 c + G(c'\hat{\Gamma}'_1 c + c'\hat{\Gamma}_1 c) + (G-1)(c'\hat{\Gamma}'_2 c + c'\hat{\Gamma}_2 c) + \dots \\ &= (G+1)c'\hat{\Gamma}_0 c + \sum_{\ell=1}^{G} (G+1-\ell)(c'\hat{\Gamma}'_\ell c + c'\hat{\Gamma}_\ell c) \end{aligned}$$

$$\begin{aligned} \Rightarrow \frac{1}{G+1} i'Pi &= c'\hat{\Gamma}_0 c + \sum_{\ell=1}^{G} \frac{G+1-\ell}{G+1}(c'\hat{\Gamma}'_\ell c + c'\hat{\Gamma}_\ell c) \\ &= c'\hat{V}_{nw}c \end{aligned}$$

Hence, it is sufficient to show that P is positive semi-definite. However, P is the matrix of sample covariances of the process $z_t = c'x_t\hat{\varepsilon}_t$ with autocovariances:

$$\mathbb{E}[c'x_t\varepsilon_t\varepsilon_{t-j}x'_{t-j}c] = c'\mathbb{E}[\varepsilon_t\varepsilon_{t-j}x_tx'_{t-j}]c = c'\Gamma_j c \quad \forall j \in \mathbb{Z}$$

The matrix of sample covariances for any process is positive semi-definite, thus $\hat{V}_{nw}$ is p.s.d. □

Note:-

The population covariance matrix is always positive semi-definite, so it's desirable for its estimate to also be positive semi-definite. Thus in a time series context we define sample covariances as:

$$\frac{1}{T}\sum_{t=|i-j|+1}^{T} z_t z_{t-|i-j|} \quad \text{rather than as} \quad \frac{1}{T-|i-j|}\sum_{t=|i-j|+1}^{T} z_t z_{t-|i-j|}$$

Even though the former is biased and the latter unbiased, had we used the latter we might get an estimate that is not positive semi-definite.