R300 Econometrics

Ben Blaker and James Legrand[1]

Michaelmas Term, 2023-2024

Contents

[1] A changelog and archive can be found at github.com/james-legrand/Metrics-Notes.

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, $\mathbb{E}[(x-\mu)^2]$
  - $Var(aX+bY) = a^2VarX + b^2VarY + 2abCov(X,Y)$
  - $\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$ (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,

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

Thus,

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

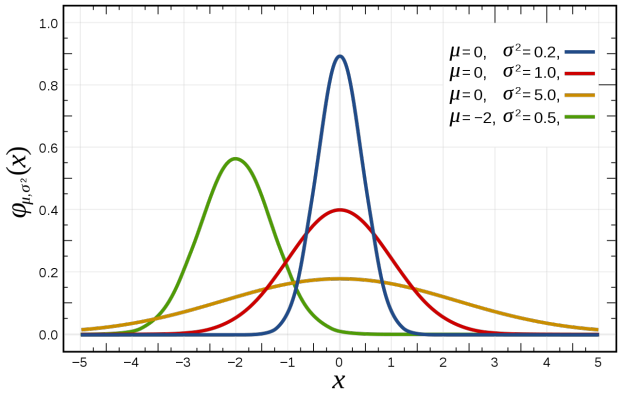
and so on.

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^2_1.
- 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 $\mathbb{E}[g(X)(X-\mu)] = \sigma^2 \mathbb{E}g'(X)$.

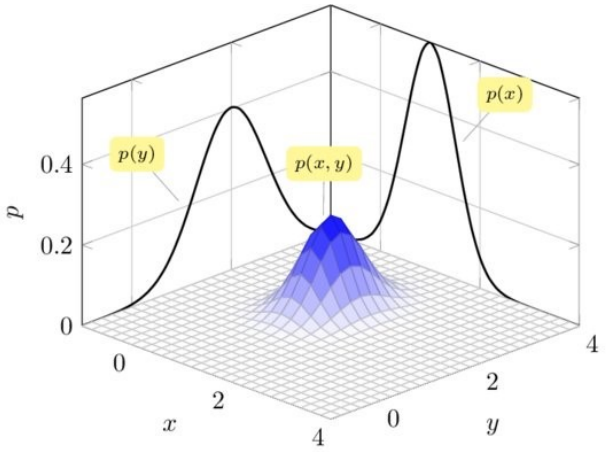
Proof. We shall prove in the case of a standard normal: $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$
Since $\int x \exp(-x^2/2)\,dx = -\exp(-x^2/2)$ we get from integration by parts:
$E[g(X)X] = \frac{1}{\sqrt{2\pi}} \int g(x) x \exp(-x^2/2)\,dx = \frac{1}{\sqrt{2\pi}} \int g'(x) \exp(-x^2/2)\,dx = E[g'(X)]$.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

Multivariate Normal

PDF:

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

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



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, $\text{corr}(X, Y) \approx -1$ and when c is very large, $\text{corr}(X, Y) \approx 1$. 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:

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

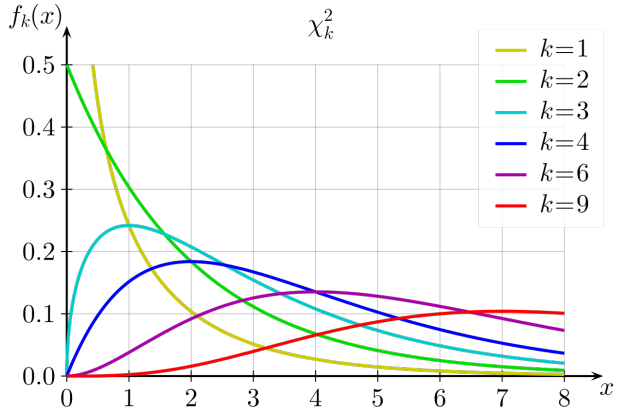
using the symmetry of $|X|$ and $|X| \leq c$. Note that $X - Y$ is not normally distributed due to the non-zero probability of $X - Y = 0$. However, a normal has no discrete part, i.e.: the probability of any point is 0. Thus, X and Y are not jointly normally distributed, even though they are individually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

Chi-Squared (χ^2)

PDF:

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

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

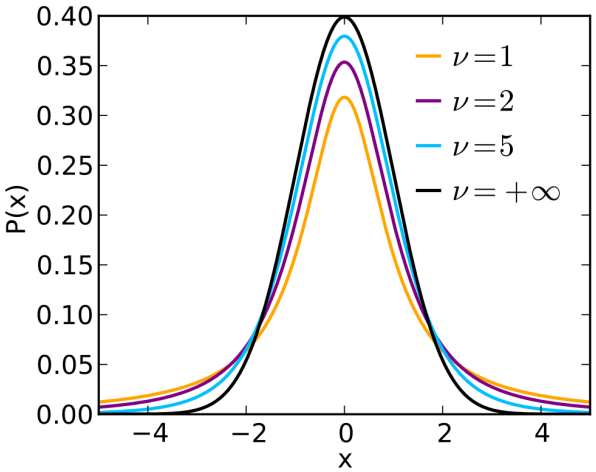
- If $X_1, \ldots, X_n$ are independent with $X_i \sim \chi^2_{p_i}$, then $\sum X_i \sim \chi^2_{\sum p_i}$ (i.e., independent chi squared variables add to a chi squared, and the degrees of freedom add).
- If $\mathbf{X} \sim N_n(\mu, \mathbf{\Sigma})$, then $(\mathbf{X}-\mu)'\mathbf{\Sigma}^{-1}(\mathbf{X}-\mu)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

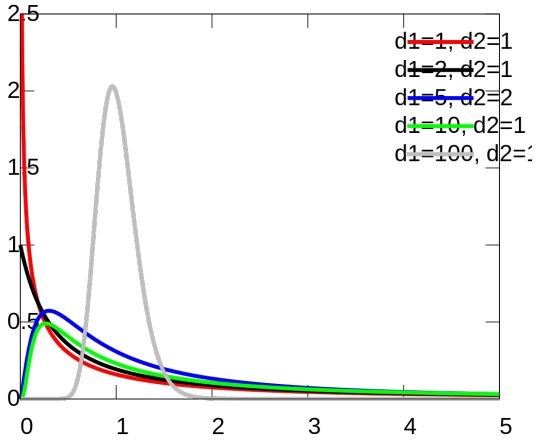
$$\text{var}(X) = E(X^2) - (E(X))^2 = \frac{\nu}{\nu - 2}$$

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.

$$E(Y - g(X))^2 = E[Y - E(Y|X) + E(Y|X) - g(X)]^2$$
$$= E[Y - E(Y|X)]^2 + 2E[(Y - E(Y|X))(E(Y|X) - g(X))] + E[E(Y|X) - g(X)]^2$$

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

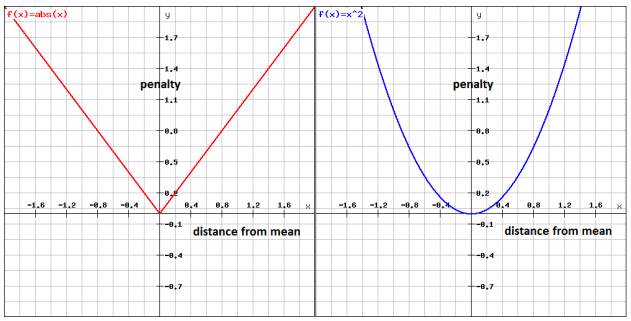
$$\frac{d}{dc}E(|X - c|) = E\left(\frac{d}{dc}|X - c|\right) = E\left(\frac{-(X - c)}{|X - c|}\right)$$

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

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

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.

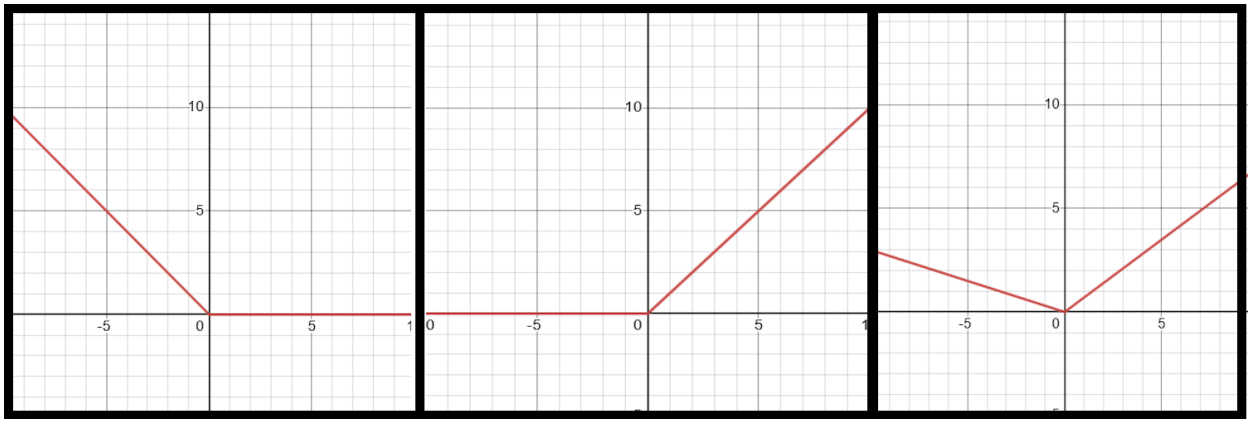
Because the MAE is a more robust estimator of scale than the sample variance or standard deviation, it works better with distributions without a mean or variance, such as the Cauchy distribution.

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

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

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 $y = h(x_1, \mathbf{x_2}, \varepsilon)$ 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 $f(\boldsymbol{\varepsilon} \,|\, x_l, \boldsymbol{x_2})$ 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 $h(x_1, \mathbf{x_2}, \varepsilon)$ 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. $\hat{y}'e = (\mathbf{X}\hat{\boldsymbol{\beta}})'e = \hat{\boldsymbol{\beta}}'\mathbf{X}'e = 0$

□

3 Geometric Interpretation of OLS, Mean Variance of OLS, Partitioned Regression

3.1 Geometric Interpretation

Consider estimation of β in the model:

$$y_i = x_i^{'}\beta + \varepsilon_i,\ i = 1, ..., n$$

This is equivalent in matrix form to: $Y = X\beta + \varepsilon$

The OLS estimator is: $\hat{\beta} = (X^{'}X)^{-1}X^{'}Y$

Definition 3.1.1

The Projection Matrix is defined as:

$$P_X = X(X^{'}X)^{-1}X^{'}$$

The Residual Maker Matrix is defined as:

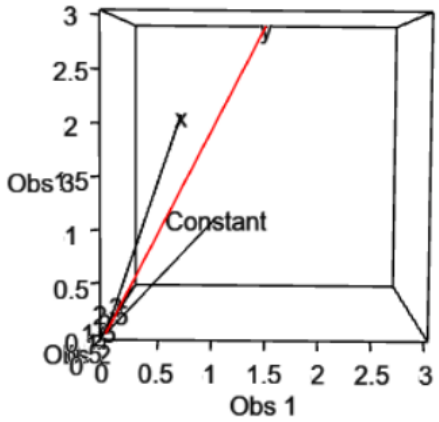
$$M_X = I - P_X$$

Then

$$\hat{Y} = X\hat{\beta} = P_X Y$$

$$\hat{\varepsilon} = Y - \hat{Y} = M_X Y$$

Claim 3.1.1. P_X and M_X are symmetric and idempotent.



Thus, $\hat{Y} = X\hat{\beta}$ is the orthogonal projection of the n-dimensional vector Y onto the subspace spanned by the columns of X. Each column of X represents the n values that each regressor takes for every observation.

The "subspace" spanned by the columns of X is the set of all linear combinations of the columns of X. The orthogonal projection of Y onto this subspace is the closest point in the subspace to Y. This is because we solve:

$$\hat{\beta} = \underset{b}{\operatorname{argmin}} \sum (y_i - x_i'b)^2 = \underset{b}{\operatorname{argmin}} (Y - Xb)'(Y - Xb) = \underset{b}{\operatorname{argmin}} \|Y - Xb\|^2$$

Example. $k = n$

Clearly if we had k=n regressors, then the columns of X would span the entire n-dimensional space and the projection would be the identity matrix. In this case, $\hat{Y} = Y$, and the residuals would be zero.

3.1.1 The Residual Vector

The difference between Y and the projection of Y onto the subspace is the residual vector $\hat{\varepsilon}$.

Claim 3.1.2. The residual vector is orthogonal to the subspace spanned by the columns of X and so is orthogonal to each column of X $X'\hat{\varepsilon} = 0$

Proof. Intuitively: This is because the projection of Y onto the subspace is the closest point in the subspace to Y. If the residual vector were not orthogonal to the subspace, then we could move the projection of Y onto the subspace along the residual vector and get a point that is closer to Y. This would contradict the fact that the projection of Y onto the subspace is the closest point in the subspace to Y.

Algebraically:

$$X'\hat{\varepsilon} = X'(Y - \hat{Y}) = X'(Y - P_X Y) = X'(Y - X(X'X)^{-1}X'Y) = 0$$

□

3.2 Conditional Mean and Variance of OLS

3.2.1 Conditional Mean

Claim 3.2.1. $\hat{\beta}$ is a conditionally unbiased estimator of β

$$\mathbb{E}[\hat{\beta}|X] = \beta$$

Proof.

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

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

1. via strict exogeneity $\mathbb{E}[\varepsilon|X] = 0$, do not need iid (e.g. can have a regressor $x_i = i$)

□

Also only need strict exogeneity for a causal interpretation of β.

Claim 3.2.2. $\hat{\beta}$ is an unconditionally unbiased estimator of β, provided expectations exist

$$\mathbb{E}[\hat{\beta}|X] = \beta$$

Proof. via law of iterated expectations

$$\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|X]] = \mathbb{E}[\beta] = \beta$$

□

3.2.2 Conditional Variance

Theorem 3.2.1.

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

Lemma 3.2.1. Unconditional Variance of a vector:

$$Var(z) = \mathbb{E}[(z - \mathbb{E}[z])(z - \mathbb{E}[z])'] = \mathbb{E}[zz'] - \mathbb{E}[z]\mathbb{E}[z']$$

Corollary 3.2.1. Conditional Variance of a vector:

$$Var(z|X) = \mathbb{E}[zz'|X] - \mathbb{E}[z|X]\mathbb{E}[z'|X]$$

Thus for $z = A(X)w$ where A is a matrix that depends on X we have:

$$\begin{aligned} Var(z|X) &= \mathbb{E}[A(X)ww'A(X)'|X] - \mathbb{E}[A(X)w|X]\mathbb{E}[w'A(X)'|X] \\ &= A(X)\mathbb{E}[ww'|X]A(X)' - A(X)\mathbb{E}[w|X]\mathbb{E}[w'|X]A(X)' \\ &= A(X)Var(w|X)A(X)' \end{aligned}$$

Therefore:

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

Then assuming homoskedasticity and no serial correlation: $Var(\varepsilon|X) = \sigma^2 I_n$

$$= (X'X)^{-1}X'\sigma^2 I_n X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

3.3 Partitioned Regression

To find formulae for conditional variances of component of $\hat{\beta}$ we can partition X and β into two parts:

$$X = \begin{bmatrix} X_1 & X_2 \end{bmatrix}$$

, X_1 is $n \times k_1$, X_2 is $n \times k_2$, $k_1 + k_2 = k$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

Then: $Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$

Theorem 3.3.1.

$$Var(\hat{\beta}_1|X) = \sigma^2(X_1'M_2X_1)^{-1}$$

Proof. Recall that

$$\hat{\beta} = \begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = (X'X)^{-1}X'Y$$

$$\begin{bmatrix}X_1 & X_2\end{bmatrix}'\begin{bmatrix}X_1 & X_2\end{bmatrix}\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}X_1 & X_2\end{bmatrix}'Y$$

thus

$$\begin{bmatrix}X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2\end{bmatrix}\begin{bmatrix}\hat{\beta}_1 \\ \hat{\beta}_2\end{bmatrix} = \begin{bmatrix}X_1'Y \\ X_2'Y\end{bmatrix}$$

this yields two equations in two unknowns:

$$X_1'X_1\hat{\beta}_1 + X_1'X_2\hat{\beta}_2 = X_1'Y$$
$$X_2'X_1\hat{\beta}_1 + X_2'X_2\hat{\beta}_2 = X_2'Y$$

Expressing $\hat{\beta}_1$ in terms of $\hat{\beta}_2$ and substituting into the second equation yields:

$$(X_2'X_1)(X_1'X_1)^{-1}(X_1'Y - (X_1'X_2)\hat{\beta}_2) + (X_2'X_2)\hat{\beta}_2 = X_2'Y$$
$$((X_2'X_2) - (X_2'X_1)(X_1'X_1)^{-1}(X_1'X_2))\hat{\beta}_2 = (X_2' - (X_2'X_1)(X_1'X_1)^{-1}X_1')Y$$
$$X_2'(I - X_1(X_1'X_1)^{-1}X_1')X_2\hat{\beta}_2 = X_2'(I - X_1(X_1'X_1)^{-1}X_1')Y$$

Recalling the definition of the residual maker matrix, M_x, we define M_1 as the residual maker matrix for X_1:

$$M_1 = I - X_1(X_1'X_1)^{-1}X_1'$$

Therefore,

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}X_2'M_1Y$$

and similarly

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y$$
$$Var(\hat{\beta}_1|X) = Var((X_1'M_2X_1)^{-1}X_1'M_2Y|X)$$
$$= (X_1'M_2X_1)^{-1}X_1'M_2Var(Y|X)M_2X_1(X_1'M_2X_1)^{-1}$$
$$= (X_1'M_2X_1)^{-1}X_1'M_2\sigma^2I_nM_2X_1(X_1'M_2X_1)^{-1}$$
$$= \sigma^2(X_1'M_2X_1)^{-1}$$

Similarly,

$$Var(\hat{\beta}_2|X) = \sigma^2(X_2'M_1X_2)^{-1}$$

If X_1 and X_2 are 'almost' colinear, projection of X_1 onto spaces orthogonal to X_2 is almost zero. Thus $X_1'M_2X_1$ is almost zero and so $Var(\hat{\beta}_1|X)$ is very large. This is an example of multicollinearity.

□

3.3.1 FRISCH-WAUGH-LOVELL THEOREM

Theorem 3.3.2. The OLS estimator of β_1 in the regression of Y on X is the same as the OLS estimator of β_1 in the regression of M_2Y on M_2X_1.

This is from a two step procedure:

1. Obtain M_2Y by regressing Y on X_2 and forming residuals. This is the portion of Y not correlated with X_2.

$$\hat{e} = Y - X_2(X_2'X_2)^{-1}X_2'Y = M_2Y$$

Obtain M_2X_1 by regressing X_1 on X_2. This is the portion of X_1 not correlated with X_2.

$$\hat{v} = X_1 - X_2(X_2'X_2)^{-1}X_2'X_1 = M_2X_1$$

2. Then regress M_2Y on M_2X_1, equivalently $\hat{e}$ on $\hat{v}$. This measures the effect of X_1 on Y after controlling for X_2.

Proof. Comparing the OLS estimators:

$$\hat{\beta}_1 = (X_1'M_2X_1)^{-1}X_1'M_2Y = (X_1'M_2'M_2X_1)^{-1}X_1'M_2'M_2Y$$

$$= [(M_2X_1)'(M_2X_1)]^{-1}(M_2X_1)'M_2Y$$

Thus the OLS estimator of β_1 in the regression of Y on X is the same as the OLS estimator of β_1 in the regression of M_2Y on M_2X_1.

Then comparing regression residuals:

$$\hat{\varepsilon} = Y - X\hat{\beta} = Y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2$$

Residual from step 2 of the partitioned regression is:

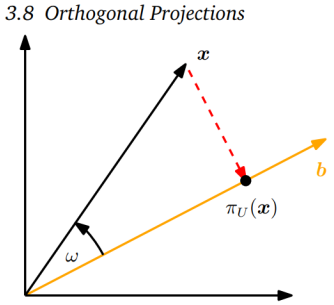
$$\tilde{\varepsilon} = M_2Y - M_2X_1\hat{\beta}_1 = M_2(Y - X_1\hat{\beta}_1) = M_2(Y - X_1\hat{\beta}_1 - X_2\hat{\beta}_2) = M_2\hat{\varepsilon} = \hat{\varepsilon}$$

This third equality holds because $M_2X_2 = 0$. Thus the residuals from the two regressions are the same and so the regression procedures are identical.

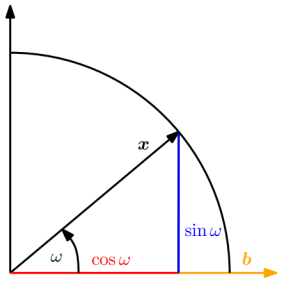
□

3.4 Appendix: Projection Onto a Line

Assume inner product is the dot proucts, defined as $x'y = \sum_{i=1}^{n} x_i y_i$



(a) Projection of $\boldsymbol{x} \in \mathbb{R}^2$ onto a subspace U with basis vector $\boldsymbol{b}$.



(b) Projection of a two-dimensional vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\| = 1$ onto a one-dimensional subspace spanned by $\boldsymbol{b}$.

where x is projected onto a one-dimensional subspace $U \subseteq \mathbb{R}^n$ spanned by basis vector b. This goes through the origin.

When projecting $x \in \mathbb{R}^n$ onto U, we want to find the vector $\pi_U(X) \in U$ that is closest to x.

Proposition 3.4.1. As before we minimise $||x - \pi_U(x)||^2$. This implies that $x - \pi_U(x)$ is orthogonal to U and thus also orthogonal to the basis vector b.

$$\langle x - \pi_U(x), b \rangle = 0$$

Proposition 3.4.2. Further, the projection $\pi_U(x)$ must be an element of U and so is a scalar multiple of b, which spans U. Hence:

$$\pi_U(x) = \lambda b$$

for some $\lambda \in \mathbb{R}$

3.4.1 Finding λ

Substituting Prop 1.4.2 into 1.4.1 we get:

$$\langle x - \lambda b, b \rangle = 0$$

Exploiting the bilinearity of the inner product:

$$\langle x, b \rangle - \lambda \langle b, b \rangle = 0$$

$$\Rightarrow \lambda = \frac{\langle x, b \rangle}{\langle b, b \rangle} = \frac{\langle x, b \rangle}{||b||^2} = \frac{x'b}{b'b}$$

3.4.2 Finding $\pi_U(x)$

Since $\pi_U(x) = \lambda b$, we have:

$$\pi_U(x) = \frac{x'b}{b'b} b$$

The length of $\pi_U(x)$ is:

$$||\pi_U(x)|| = ||\lambda b|| = |\lambda| \, ||b||$$

Thus the projection acts as a coordinate of $\pi_U(x)$ in the direction of b.

Using the dot product as the inner product we have:

$$= \frac{|x'b|}{||b||^2} ||b|| = |cos(\theta)| \, ||x|| \, ||b|| \frac{||b||}{||b||^2} = |cos(\theta)| \, ||x||$$

3.4.3 The Projection Matrix P_π

As projection is a linear mapping, there exists a matrix P_π such that:

$$\pi_U(x) = P_\pi x$$

With the dot as the inner product and

$$\pi_U(x) = \lambda b = b\lambda = b\frac{b'x}{||b||^2} = \frac{bb'}{||b||^2}x$$

Thus

$$P_\pi = \frac{bb'}{||b||^2}$$

3.5 Projection Onto a General Subspace

We find a projection of $x \in \mathbb{R}^n$ onto a subspace $U \subseteq \mathbb{R}^n$ with $dim(U) = m \geq 1$. Assume that $b_1, ..., b_m$ is an ordered basis for U. Any projection $\pi_U(x)$ onto U can be written as a linear combination of the basis vectors: such that $\pi_U(x) = \sum_{i=1}^m \lambda_i b_i$. We follow the same three step procedure as before:

3.5.1 Finding $\lambda_1, ..., \lambda_m$

We find coordinates $\lambda_1, ..., \lambda_m$ such that the linear combination

$$\pi_U(x) = \sum_{i=1}^m \lambda_i b_i = \mathbf{B}\vec{\lambda}$$

$$\mathbf{B} = \begin{bmatrix} \vec{b_1} & ... & \vec{b_m} \end{bmatrix}, \in \mathbb{R}^{n\times m}, \vec{\lambda} = \begin{bmatrix} \lambda_1 \\ ... \\ \lambda_m \end{bmatrix} \in \mathbb{R}^m$$

is such that $\pi_U(x)$ is the closest point in U to x. This implies that $x - \pi_U(x)$ is orthogonal to U and thus also orthogonal to each basis vector b_i. Thus we obtain simultaneous equations:

$$\langle x - \pi_U(x), b_1 \rangle = b_1'(x - \pi_U(x)) = 0$$

$$\vdots$$

$$\langle x - \pi_U(x), b_m \rangle = b_m'(x - \pi_U(x)) = 0$$

as $\pi_U(x) = \mathbf{B}\vec{\lambda}$ we have:

$$b_1'(x - \mathbf{B}\vec{\lambda}) = 0$$

$$\vdots$$

$$b_m'(x - \mathbf{B}\vec{\lambda}) = 0$$

thus we obtain a homogeneous system of linear equations:

$$\begin{bmatrix} b_1' \\ \vdots \\ b_m' \end{bmatrix} (x - \mathbf{B}\vec{\lambda}) = 0$$

$$\Leftrightarrow \mathbf{B}'(x - \mathbf{B}\vec{\lambda}) = 0$$
$$\Leftrightarrow \mathbf{B}'\mathbf{B}\vec{\lambda} = \mathbf{B}'x$$
$$\Leftrightarrow \vec{\lambda} = (\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'x$$

where we require that $\mathbf{B}'\mathbf{B}$ is invertible, which is true if and only if $\mathbf{B}$ has full column rank, which is true if and only if the basis vectors $b_1, ..., b_m$ are linearly independent.

3.5.2 Finding $\pi_U(x)$

We have that $\pi_U(x) = \mathbf{B}\vec{\lambda}$ and so:

$$\pi_U(x) = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'x$$

3.5.3 The Projection Matrix P_π

As projection is a linear mapping, there exists a matrix P_π such that:

$$\pi_U(x) = P_\pi x$$

Thus

$$P_\pi = \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'$$

3.6 Appendix: OLS Estimator Equivalence

Claim 3.6.1.

$$\hat{\beta} = (X'X)^{-1}X'Y = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$
$$\Leftrightarrow X$$

includes a constant

Let us take the case for $k = 1$, i.e. X is a vector of length n. Then: suppose X includes a constant, i.e. $X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$. Then let $\tilde{x_i} = (1, x_i)'$ Then $X = (\tilde{x_1}, ..., \tilde{x_n})'$ Thus:

$$(X'X)^{-1}X'Y = (\sum_{i=1}^{n} \tilde{x_i}\tilde{x_i}')^{-1} \sum_{i=1}^{n} \tilde{x_i}y_i$$

$$= \left[\sum_{i=1}^{n} \begin{bmatrix} 1 & x_i \\ x_i & x_i^2 \end{bmatrix}\right]^{-1} \sum_{i=1}^{n} \begin{bmatrix} 1 \\ x_i \end{bmatrix} Y_i$$

$$= \left[n \begin{bmatrix} 1 & \bar{x} \\ \bar{x} & \frac{1}{n}\sum_{i=1}^{n} x_i^2 \end{bmatrix}\right]^{-1} n \begin{bmatrix} \bar{y} \\ \frac{1}{n}\sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

=

$$\frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix} \begin{bmatrix} \bar{y} \\ \frac{1}{n}\sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

The second component is the estimate fo the slope coefficient, and the first component is the estimate of the intercept coefficient. Thus we have:

$$\hat{\beta} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

4 Gauss-Markov Theorem. Estimation of σ^2. Distribution of OLS in normal regression

4.1 Gauss-Markov Theorem

Theorem 4.1.1. Consider an $n \times 1$ random vector Y and an $n \times k$ random matrix X.

Assume (no need for iid, large n or normality):

- **GM1** No perfect multicollinearity: $\text{rank}(X) = k$
- **GM2** Strict Exogeneity $E(Y|X) = X\beta$, equivalently $E(\varepsilon|X) = 0$
- **GM3** Homoskedasticity and no serial correlation $Var(Y|X) = \sigma^2 I$, equivalently $Var(\varepsilon|X) = \sigma^2 I$

Then, the OLS estimator $\hat{\beta}_{OLS}$ has the minimum conditional variance in the class of estimators that, conditional on every X, are linear in Y and unbiased. Thus $\hat{\beta}_{OLS}$ is the Best Linear conditionally Unbiased Estimator (BLUE).

A linear estimator of β is any estimator of the form $\tilde{\beta} = A(X)Y$ where $A(X)$ is a $k \times n$ matrix. For OLS $\tilde{\beta}_{OLS} = A(X)Y = (X'X)^{-1}X'Y$

Definition 4.1.1

Minimum conditional variance implies:

$$Var(\tilde{\beta}|X) - Var(\hat{\beta}_{OLS}|X) \quad \text{is positive semi-definite} \quad \forall \tilde{\beta}$$

This $k \times k$ matrix A is positive semi-definite iff $z'Az \geq 0$ for all $k \times 1$ vectors z. Thus for any z:

$$z'Var(\tilde{\beta}|X)z \geq z'Var(\hat{\beta}_{OLS}|X)z$$

Note this is equivalent to:

$$Var(z'\tilde{\beta}|X) \geq Var(z'\hat{\beta}_{OLS}|X)$$

Thus any linear combination of the elements of $\tilde{\beta}$ has a conditional variance that is at least as large as the conditional variance of the corresponding linear combination of the elements of $\hat{\beta}_{OLS}$.
In particular, any component of $\tilde{\beta}$ has a conditional variance that is at least as large as the conditional variance of the corresponding component of $\hat{\beta}_{OLS}$.

4.2 GM PROOF

Lemma 4.2.1. We know

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'Y$$

is conditioanlly (and unconditionally) unbiased under GM2, and is a linear function of Y. We also know under GM3 that

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

Now consider any other linear conditionally unbiased estimator $\tilde{\beta} = A(X)Y$.

$$E(\tilde{\beta}|X) = E(AY|X) = AE(Y|X) = AX\beta \quad \text{under GM3}$$

As we assume conditioanlly unbiased, for any β

$$AX\beta = \beta \Rightarrow AX = I$$

Lemma 4.2.2.

$$Var(\tilde{\beta}|X) = Var(AY|X) = AVar(Y|X)A' = A\sigma^2 I_n A' = \sigma^2 AA'$$

Decomposing A:

$$A = A - (X'X)^{-1}X' + (X'X)^{-1}X' = W + (X'X)^{-1}X'$$

Thus:

$$\begin{aligned} Var(\tilde{\beta}|X) &= \sigma^2(W + (X'X)^{-1}X')(W + (X'X)^{-1}X')' \\ &= \sigma^2(W + (X'X)^{-1}X')(W' + X(X'X)^{-1}) \end{aligned}$$

But

$$WX = AX - (X'X)^{-1}X'X = I - I = 0$$

Therefore,

$$\begin{aligned} Var(\tilde{\beta}|X) &= \sigma^2 WW' + \sigma^2(X'X)^{-1} \\ &= Var(\hat{\beta}_{OLS}|X) + \sigma^2 WW' \end{aligned}$$

Lemma 4.2.3. $\sigma^2 WW'$ is positive semi-definite

For any k-dimensional vector z, denote the k-dimensional vector $W'z$ as $\alpha = (\alpha_1, ..., \alpha_k)'$

$$z'\sigma^2 WW'z = \sigma^2(z'W)(W'z) = \sigma^2\alpha'\alpha = \sigma^2 \sum_{i=1}^{k} \alpha_i^2 \geq 0$$

Thus $Var(\tilde{\beta}|X) - Var(\hat{\beta}_{OLS}|X)$ is psd for any linear conditionally unbiased estimator $\tilde{\beta} = AY$.

□

4.3 Estimation of σ^2

Given the following but how to estimate?:

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

We are given X, thus only need to estimate σ^2.

Note: $\sigma^2 = E(\varepsilon_i^2|X)$ and since trivially $\sigma^2 = E(\sigma^2)$ we have:

$$\sigma^2 = E(E(\varepsilon_i^2|X)) = E(\varepsilon_i^2)$$

This suggests the MOM estimator:

$$\hat{\sigma}^2_{MOM} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$$

except we don't know ε_i as β is unknown.

But with $\hat{\beta} = \hat{\beta}_{OLS}$ we can use the sample analague $\hat{\varepsilon}_i = y_i - x_i'\hat{\beta}_{OLS}$ and thus:

Theorem 4.3.1. The biased (ML) estimator of σ^2 is:

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 \left(= \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i'\hat{\varepsilon}_i\right)$$

4.3.1 Unbiased Estimator of σ^2

We first compute the bias of $E(\hat{\sigma}^2|X)$ to then correct for it:

Lemma 4.3.1.

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

Then

$$E(\hat{\sigma}^2|X) = E(\hat{\varepsilon}'\hat{\varepsilon}|X)$$

$$= E(\varepsilon' M_X' M_X \varepsilon|X) = E(\varepsilon' M_X \varepsilon|X) = E(tr(\varepsilon' M_X \varepsilon)|X), \quad \text{since argument is a scalar}$$

$$\because \text{trace multiplications are commutative if conformations exists} \Rightarrow$$

$$= E(tr(M_X\varepsilon\varepsilon')|X) = tr(E(\varepsilon' M_X \varepsilon|X)) = tr(M_X E(\varepsilon\varepsilon'|X)) = tr(M_X\sigma^2 I_n) = \sigma^2 tr(M_X)$$

Lemma 4.3.2.

$$tr(M_X) = n - k$$

Proof.

$$M_X = (I - X(X'X)^{-1}X')$$

$$tr(M_X) = tr(I_n - X(X'X)^{-1}X') = tr(I_n) - ((X'X)^{-1}X'X)$$

$$tr(I_n) - tr(I_k)$$

$$= n - k$$

□

Thus

$$E(\hat{\sigma}^2|X) = \frac{n-k}{n}\sigma^2$$

Theorem 4.3.2.

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

is an unbiased estimator of σ^2 is an unbiased estimator of σ^2

4.4 Estimation of standard errors

By default STATA computes s.e. of component $\hat{\beta}_j$ of $\hat{\beta}_{OLS}$ as the square roots from the i-th diagonal element of $\hat{\sigma}^2(X'X)^{-1}$ or more explicitly with the partitioned regression formulae:

$$\hat{se}(\hat{\beta}_i) = \frac{\hat{\sigma_u}}{\sqrt{X_i' M_{-i} X_i}}$$

where X_i is the i-th column of X (i-th regressor) and M_{-i} is the residual maker matrix in the regression on all the other explanatory variables but X_i.

4.5 Distribution of OLS

Knowing the mean and variance of OLS is not sufficient to test hypotheses about β. We need to also know the distribution. In small samples this is easy to derive with the following assumption:

$$\textbf{Normal Regression} : \varepsilon | X \ N(0, \sigma^2 I_n)$$

This subsumes GM2 and GM3 and adds normality.

Claim 4.5.1. There are several properties of the multivariate Gaussian which become useful in derivations.

- If $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I_n})$ and $\mathbf{A}$ be any deterministic $r \times n$ matrix, then $\mathbf{AZ} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{AA}')$. In particular any linear combinations of normals is normal.
- The Normal distribution $N(0, \sigma^2 I_n)$ is invariant to rotations/orthogonal transformations. If $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I_n})$ and $\mathbf{Q}$ is any $n \times n$ orthogonal matrix, then $\mathbf{QZ} \sim \mathbf{N}(\mathbf{0}, \sigma^2 \mathbf{I_n})$, i.e. QZ has the same distribution as Z.

Theorem 4.5.1. Using the first property and the normal regression assumption, we obtain:

$$\hat{\beta}_{OLS} | X = \beta + (X'X)^{-1} X' \varepsilon | X \sim N(\beta, \sigma^2 (X'X)^{-1})$$

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 $\hat{\boldsymbol{\sigma}}^{\mathbf{2}}$ and $\hat{\boldsymbol{\beta}}$ to make progress.

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

Recall the definition of the variance estimator:

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

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

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

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 $\mathbf{v}'\mathbf{M_X}\mathbf{v} = \mathbf{v}'\mathbf{M_X^2}\mathbf{v} = \mathbf{v}'\mathbf{M_X}'\mathbf{M_X}\mathbf{v} = (\mathbf{v}'\mathbf{M_X})'(\mathbf{M_X}\mathbf{v}) = \|\mathbf{M_X}\mathbf{v}\|^2$ 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 $\text{Var}(\mathbf{H'}\boldsymbol{e}) = \mathbf{H'I_n}\sigma^2\mathbf{H} = \sigma^2\mathbf{H'H} = \mathbf{I_n}\sigma^2$. Thus $\mathbf{H'}\boldsymbol{\varepsilon} \sim N(0, \mathbf{I_n}\sigma^2)$.

Let $\mathbf{u} = \mathbf{H'}\boldsymbol{\varepsilon}$, and partition $\underset{n\times 1}{\mathbf{u}} = \begin{bmatrix} \mathbf{u_1} \\ {\scriptstyle (n-k)\times 1} \\ \mathbf{u_2} \\ {\scriptstyle k\times 1} \end{bmatrix}$ 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 normals with mean 0 and variance σ^2. We can transform each normal into a standard normal with division by σ; since each normal is squared we divide by σ^2. A sum of j squared standard normals is distributed χ^2_j, thus $\frac{(n-k)\hat{\sigma}^2}{\sigma^2}$ 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 $\hat{\beta} = \hat{\beta}_{OLS} = (X'X)X'Y$

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

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

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

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 ($z_n \xrightarrow{d} z \not\Rightarrow z_n \xrightarrow{p} z$)**.** 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 $n^{\gamma}(X_n - X) \xrightarrow{p} 0$.
- 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 $z_n \sim$ 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 $z \sim$ Uniform$(0,1)$, $c \sim$ 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:

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

□

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 $E(u_i^2) = \mathbf{c}'\mathbf{\Sigma}\mathbf{c} < \infty$. By the univariate CLT,

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

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

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

□

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)	$E(y_i \mid x_i) = x_i'\beta$	(GM2)	$E(\mathbf{Y}\mid\mathbf{X}) = \mathbf{X}'\beta$
(OLS3)	$\mathrm{Var}(y_i \mid x_i) = \sigma^2$	(GM3)	$\mathrm{Var}(\mathbf{Y}\mid\mathbf{X}) = \sigma^2\mathbf{I}$
(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,

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

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

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

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

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

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

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

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

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

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

3.

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

□

[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

$$G(\theta) = \left[\frac{\partial g(\theta_1,\theta_2)}{\partial \theta_1} \quad \frac{\partial g(\theta_1,\theta_2)}{\partial \theta_2}\right] = [\theta_2, \theta_1]$$

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

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

7 Multicollinearity. Ridge and LASSO. Model Selection for Prediction. Mallow's C_P Criterion

7.0.1 Perfect Multicollinearity

Definition 7.0.1: Perfect Multicollinearity

This defines the case where rank $(X) \neq k$ and GM1 is violated

$\Rightarrow \exists$ an exact linear dependence between the columns of X so that

$$X\alpha = 0 \quad \text{for some } \alpha \neq 0$$

Corollary 7.0.1. Models

$$Y = X\beta + \varepsilon \quad \text{and} \quad Y = X\tilde{\beta} + \varepsilon$$

where $\tilde{\beta} = \beta + \lambda\alpha$ for any λ are equivalent. Thus β is not identified as we cannot distinguish between the two.

Corollary 7.0.2. In the OLS case, rank $(X) \neq k$:

$\Rightarrow$ non invertability of $X'X$
$\Rightarrow$ infinite number of solutions to the OLS problem. $min_b||Y - Xb||^2$. Thus OLS estimator is not well defined.

Example. Dummy Variable Trap:

Consider the model $Y = \beta_0 X_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where X_0 is a constant vector $= \vec{1}$ and X_1, X_2 are dummies such that $X_2 = 1 - X_1$.

$\Rightarrow X_2$ is a linear combination of X_1 and X_0 and thus model is not identified. Specifically, $X\alpha = 0, \text{for } \alpha = (1, -1, -1)'$

$\Rightarrow X'X$ is not invertible.

$\Rightarrow$ OLS estimator is not well defined.

We also can have perfect multicollinearity with small sample sizes s.t. $n < k$. Thus rank $(X) \leq n < k$.

7.0.2 Imperfect Multicollinearity

Often there will exist a linear combination of X that is almost but not exactly 0

Definition 7.0.2

Imperfect multicollinearity

$$rank(X) = k, \quad X\alpha \approx 0 \quad \text{for some } \alpha \neq 0$$

But this is not a unit invariant quantity, instead we can look at the relative size of the eigenvalues of $X'X$.

$$\frac{\lambda_{max}}{\lambda_{min}}$$

Corollary 7.0.3. Multicollinearity will result in a large norm of the variance-covariance matrix of the OLS estimator $\sigma^2(X'X)^{-1}$ and thus a very large trace. Trace of the variance of OLS estimator yields:

$$tr(\sigma^2(X'X)^{-1}) = E(||\hat{\beta} - \beta||^2|X)$$

Thus large multicollinearity will mean a large expected squared distance between true and estimated value of parameters.

Note:-

Norm here is defined as $\max_{x\neq 0} \frac{||Ax||}{||x||}$, more details

Also:

$$\frac{\lambda_{max}}{\lambda_{min}}$$

represents the condition number $C = ||A||\,||A^{-1}||$ of matrix $X'X$ (if this matrix is positive definite and (given) symmetric)[a]. Intution behind condition number: For a set of simulateneous equations $Ax = b$ the condition number of A sets a bound of the sensitivity of the relative (i.e. invariant to units) solution error (in x) induced by errors in the problem vector (b). $\frac{||\Delta x||}{||x||} \leq C\frac{||\Delta b||}{||b||}$[b]. Thus if C is large, then the solution can be very sensitive to small changes in the problem. (Worst case when b points in the smallest eigenvector direction of A and Δb points in the largest eigenvector direction of A.)

Applying this to $X'X$, we can see that if the condition number is large, then the "solution error" in β can be very sensitive to small changes in the problem, i.e. small changes in $X'Y$, i.e. ε. Here, $Ax = b \equiv X'X\beta = X'Y$. This shows why it is a good condition to represent the large variance of the OLS estimator in the presence of multicollinearity.

[a] Proof in Appendix
[b] Proof in Appendix

7.1 Ridge Regression

By G-M the resulting large MSE is still best among any other conditionally unbiased estimator. Thus, solutions recommend minimising expected MSE via introducing bias and reducing variance.

Definition 7.1.1

Ridge Regression Estimator:

This solves the OLS problem but where size of $||\beta||^2 = \beta'\beta$ is penalised by λ.

$$\min_{\beta} \left(||Y - X\beta||^2 + \lambda ||\beta||^2 \right) \text{ for some } \lambda > 0$$

We can show this results in the estimate (for $k < n$):

$$\hat{\beta}_r = (X'X + \lambda I)^{-1} X'Y$$

Note:

$$E(\hat{\beta}_r | X) = (X'X + \lambda I)^{-1} X'X\beta$$

and thus $\hat{\beta}_r$ is biased. When λ rises the bias increases and the variance decreases, illustrating the exploitable bias-variance tradeoff.

$$Var(\hat{\beta}_r | X) = \sigma^2 (X'X + \lambda I)^{-1} X'X (X'X + \lambda I)^{-1}$$

We can see this explicitly by looking at the condition number of $X'X + \lambda I$:[c]

$$= \frac{\lambda_{max} + \lambda}{\lambda_{min} + \lambda} < \frac{\lambda_{max}}{\lambda_{min}}$$

As $(X'X + \lambda I)\hat{\beta}_r = X'Y$, we can see that the relative solution error in $\hat{\beta}_r$ is bounded by:

$$\frac{||\Delta \hat{\beta}_r||}{||\hat{\beta}_r||} \leq \frac{\lambda_{max} + \lambda}{\lambda_{min} + \lambda} \frac{||\Delta X'Y||}{||X'Y||}$$

Thus, as λ rises, the solution error in $\hat{\beta}_r$ becomes less sensitive to changes in $X'Y$, i.e. ε. This is the bias-variance tradeoff.

[c] Proof in Appendix

Proof.

$$\equiv \min_{b} (Y'Y - b'X'Y - Y'Xb + b'X'Xb + \lambda b'b)$$

$$FOC : -X'Y - X'Y + 2X'Xb + 2\lambda b = 0$$

$$2(X'X + \lambda I)b = 2X'Y$$

$$b = (X'X + \lambda I)^{-1} X'Y = \hat{\beta}_r$$

□

7.1.1 Cross-Validation

The parameter λ is usually chosen by CV:

Overall idea is to minimise expected squared prediction error $E(y - x'\hat{\beta}_r)^2$, where (y, x) is a new observation from the joint distribution of the dependent and explanatory variables. As we do not have such a new observation we approximate it via the leave one out CV method:

1. Drop the i-th observation (y_i, x_i) from the sample
2. Estimate $\hat{\beta}_{r(-i)}$ on the remaining $n-1$ observations
3. Estimate $E(y - x'\hat{\beta}_{-i,r})^2$ via taking the average of the squared prediction errors on the dropped observation (y_i, x_i) via $\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\beta}_{-i,r})^2$
4. Choose λ that minimises the above estimate.

Example. Ridge in an Overidentified Model: $k > n$

Clearly here the classical OLS estimator would not be defined, since $X'X$ is not invertible. Its rank is bounded by n but dimensions are $k \times k$. Thus we need to introduce some bias to make the problem well defined. We note that, assuming (X is of rank n), the column space of X will span $\mathbb{R}^n$. Thus we can find exact vectors b such that $Xb = Y$, where multiplicity due to the fact $k > n$. We select among these by minimising the L2 norm:

$$\min_b ||b||^2 \quad \text{subject to} \quad Xb = Y$$

$$\mathbb{L} = b'b + \lambda'(Y - Xb)$$

$$\text{FOC}: \begin{cases} b : 2b - X'\lambda = 0 \\ \lambda : (Y - Xb) = 0 \end{cases}$$

$$\Rightarrow b = \frac{1}{2}X'\lambda \quad \text{and} \quad Y = \frac{1}{2}XX'\lambda$$

$$\Rightarrow \frac{\lambda}{2} = (XX')^{-1}Y$$

$$\Rightarrow b = X'(XX')^{-1}Y$$

Note:-

Exact Identification $n = k$
Provided X is full rank we then simply have a unique solution to the OLS problem and take:

$$\hat{\beta} = X^{-1}Y$$

7.2 Mallow's C_P Criterion

When we allow biased estimators and only care about expected prediction error we then are no longer bound to correctly specified models. Intuitively if some regressors have non-zero, but small coefficients, we may still want to drop them to reduce the variance of prediction at the expense of introducing some, hopefully small, bias.

Proposition 7.2.1. Consider two models, unrestricted and restricted:

$$Y = X\beta + \varepsilon = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

$$Y = X_1\beta_1 + \varepsilon$$

where X is $n \times k$ and X_1 is $n \times p$ and X_2 is $n \times q$, where $p+q = k$. Suppose the true model is the unrestricted one and this regression satisfies GM1-3 assumptions, and $\varepsilon | X_1, X_2 \sim N(0, \sigma^2 I_n)$.

Under the Mallow's criterion we prefer the unrestricted model if:

$$C_p = \frac{SSR_r}{\hat{\sigma}^2} - n + 2p > k$$

Intuitively we may still prefer the restricted model if β_2 could only be estimated very imprecisely.

We show the bounds of this intuition as follows:

Long model: $\hat{Y} = X\hat{\beta}$
Short model: $\tilde{Y} = X_1\tilde{\beta}_1$

Our measure of accuracy for any predictor $\check{Y}$ of Y is the expected scaled sum of squared deviations of $\check{Y}$ from the best (infeasible) predictor $X\beta$.

$$J = E\left(\frac{1}{\sigma^2}(\check{Y} - X\beta)'(\check{Y} - X\beta)\right)$$

n.b. the following expectations are <u>all condtional on X</u>

Lemma 7.2.1. For $\check{Y} = \hat{Y}$, we have $J_u = \frac{(\hat{\beta}-\beta)'X'X(\hat{\beta}-\beta)}{\sigma^2} = k$

Proof.

$$\begin{aligned}
J_u &= E\left[\frac{1}{\sigma^2}(\hat{Y} - X\beta)'(\hat{Y} - X\beta)\right] \\
&= E\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{\sigma^2} \\
&= Etr(\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{\sigma^2}) \\
&= tr(E\frac{(\hat{\beta} - \beta)'X'X(\hat{\beta} - \beta)}{\sigma^2}) \\
&= tr(\frac{X'X}{\sigma^2}E(\hat{\beta} - \beta)(\hat{\beta} - \beta)') \\
&= tr(\frac{X'X}{\sigma^2}Var(\hat{\beta})) \\
&= tr(\frac{X'X}{\sigma^2}\sigma^2(X'X)^{-1}) \\
&= tr(I_k) = k
\end{aligned}$$

□

Lemma 7.2.2. For $\check{Y} = \tilde{Y}$, we have $J_r = p + \frac{1}{\sigma^2}\beta_2'X_2'M_2X_2\beta_2$

Proof.

$$J_r = E\frac{(\begin{pmatrix}\tilde{\beta}_1\\0\end{pmatrix} - \begin{pmatrix}\beta_1\\\beta_2\end{pmatrix})'X'X(\begin{pmatrix}\tilde{\beta}_1\\0\end{pmatrix} - \begin{pmatrix}\beta_1\\\beta_2\end{pmatrix})}{\sigma^2}$$

Omitted Variable Bias:

$$\begin{aligned}
E(\tilde{\beta}_1|X) &= E((X_1'X_1)^{-1}X_1'Y|X) = E((X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + \varepsilon)|X) \\
&= \beta_1 + \underline{(X_1'X_1)^{-1}X_1'X_2\beta_2}
\end{aligned}$$

$\begin{pmatrix}\tilde{\beta}_1\\0\end{pmatrix}-\begin{pmatrix}\beta_1\\\beta_2\end{pmatrix}=\begin{pmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{pmatrix}-\begin{pmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\\beta_2\end{pmatrix}=\begin{pmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{pmatrix}+\begin{pmatrix}(X_1'X_1)^{-1}X_1'X_2\beta_2\\-\beta_2\end{pmatrix}$provided some cross product

Thus we can decompose J_r as follows:

$$J_r = E\frac{\begin{pmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{pmatrix}'X'X\begin{pmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\0\end{pmatrix}}{\sigma^2}+\frac{\begin{pmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\-\beta_2\end{pmatrix}'X'X\begin{pmatrix}\tilde{\beta}_1-E(\tilde{\beta}_1|X)\\-\beta_2\end{pmatrix}}{\sigma^2}(+\dots\text{cross terms}=0)^*$$

$$J_r = E\frac{(\tilde{\beta}_1-E(\tilde{\beta}_1|X))'X'X(\tilde{\beta}_1-E(\tilde{\beta}_1|X))}{\sigma^2}+E\frac{\begin{pmatrix}(X_1'X_1)^{-1}X_1'X_2\beta_2\\-\beta_2\end{pmatrix}'\begin{pmatrix}X_1'X_1 & X_1'X_2\\X_2'X_1 & X_2'X_2\end{pmatrix}\begin{pmatrix}(X_1'X_1)^{-1}X_1'X_2\beta_2\\-\beta_2\end{pmatrix}}{\sigma^2}$$

Simplifying latter term's numerator:

$$\begin{aligned}&=\beta_2'X_2'X_1(X_1'X_1)^{-1}(X_1'X_1)(X_1'X_1)^{-1}X_1'X_2\beta_2-2\beta_2'X_2'X_1(X_1'X_1)^{-1}X_1'X_2\beta_2+\beta_2'X_2'X_2\beta_2)\\&=(-\beta_2'X_2'X_1(X_1'X_1)^{-1}X_1'X_2\beta_2+\beta_2'X_2'X_2\beta_2)\\&=\beta_2'X_2'(I-P_1)X_2\beta_2\\&=\beta_2'X_2'M_1X_2\beta_2\end{aligned}$$

**Evaluating cross terms is equivalent to evaluating the above except substituting one $-\beta_2=0$. Clearly then we have full cancellation and thus we do not need to consider these terms.*

Simplifying the first term:

$$\begin{aligned}&\frac{1}{\sigma^2}E[tr((\tilde{\beta}_1-E(\tilde{\beta}_1|X))'X_1'X_1(\tilde{\beta}_1-E(\tilde{\beta}_1|X))|X)]\\&=\frac{1}{\sigma^2}tr(X'XVar(\tilde{\beta}_1|X))\end{aligned}$$

$$\begin{aligned}\text{But: } Var(\tilde{\beta}_1|X)&=Var((X_1'X_1)^{-1}X_1'Y|X)=(X_1'X_1)^{-1}X_1'Var(Y|X)X_1(X_1'X_1)^{-1}\\&=\sigma^2(X_1'X_1)^{-1}\end{aligned}$$

Thus we have:

$$\begin{aligned}J_r&=tr(I_p)+\frac{1}{\sigma^2}\beta_2'X_2'M_1X_2\beta_2\\&=p+\frac{1}{\sigma^2}\beta_2'X_2'M_1X_2\beta_2\end{aligned}$$

□

Solution:-

Therefore:

$$J_r < J_u \quad \text{if and only if} \quad \frac{1}{\sigma^2}\beta_2'X_2'M_1X_2\beta_2 < q$$

This is likely to occur if

- β_2 is small (small bias when omitted)
- X_2 is highly correlated with X_1, i.e. high multicollinearity (lowers value of $X_2'M_1X_2$, since this is the SSR from regressing X_2 on X_1) ($\hat{\varepsilon}=M_1X_2, SSR=\hat{\varepsilon}'\hat{\varepsilon}=X_2'M_1'M_1X_2$)
- σ^2 is large

We can estimate the LHS via considering the restricted SSR

$$SSR_r = (Y - X_1\tilde{\beta}_1)'(Y - X_1\tilde{\beta}_1) = Y'M_1Y$$

$$\begin{aligned} E(SSR_r) = E((Y'M_1Y)|X) &= E(X_1\beta_1 + X_2\beta_2 + \varepsilon)'M_1(X_1\beta_1 + X_2\beta_2 + \varepsilon) \\ &= \beta_2'X_2'M_1X_2\beta_2 + E(\varepsilon'M_1\varepsilon|X) \quad \text{since} M_1X_1 = 0 \\ &= \beta_2'X_2'M_1X_2\beta_2 + trE(M_1\varepsilon'\varepsilon) = \beta_2'X_2'M_1X_2\beta_2 + \sigma^2 tr(M_1) \\ &= \beta_2'X_2'M_1X_2\beta_2 + \sigma^2(n-p) \end{aligned}$$

Therefore as we can use the MOM sample analogue of $E(SSR_r)$ and σ^2, we can estimate the LHS of the inequality above as:

$$\frac{SSR_r}{\hat{\sigma}^2} - (n-p)$$

$$\Rightarrow \hat{J}_r = p + \frac{SSR_r}{\hat{\sigma}^2} - (n-p)$$

Definition 7.2.1

Mallow's C_p Model Selection Criterion:

$$C_p = \frac{SSR_r}{\hat{\sigma}^2} - n + 2p$$

$$\text{where} \quad \hat{\sigma}^2 = \frac{SSR_u}{n-k} \quad \text{is estimated from the long regression}$$

Minimising this across different sub-models of the 'long' model yields a 'short' model most adequate for 'prediction'. We then only prefer the 'long' model to short if $C_p > J_u = k$.

As C_p only estimates J_r, choosing the minimum is not a guarantee of the best model, especially under small sample sizes and model subsets that have similar predictive power (flat minimum).

7.2.1 F-Test Interpretation

Mallows' C_p can be thought of as providing a guidance for the choice of the 'optimal' critical value of the F-test in testing the hypothesis that $\beta_2 = 0$.

$$\begin{aligned} C_p = \frac{SSR_r}{SSR_u/(n-k)} - n + 2p &= \frac{SSR_r - SSR_u}{SSR_u/(n-k)} + 2p - k \\ &= (k-p)(\text{F-stat}) + 2p - k \end{aligned}$$

Hence, $C_P > k$ (and so choose long) if and only if

$$(k-p)(\text{F-stat}) + 2p - 2k > 0$$

$$\text{F-stat} > 2$$

7.2.2 Penalised Least Squares interpretation

Proposition 7.2.2. The OLS estimator in the restricted model chosen by C_p (not considering the long regression) can also be viewed as the result of the following penalised least squares criterion:

$$\min(||Y - X\beta||^2 + \lambda||\beta||_0), \quad \text{where} \, \lambda = 2\hat{\sigma}^2$$

and the 0-norm of a vector is the number of non-zero elements in it. Note this norm is not a convex function of β, which makes minimisation difficult when k is large (as we would need to consider 2^k possible models with different combinations of included regressors and compare the penalised least squares results).

Proof.

$$\underset{M_j \in M_r}{\operatorname{argmin}} \underset{b \in M_j}{\operatorname{argmin}} C_p = \frac{RSS_r}{\hat{\sigma}^2} - n + 2p,$$

where M_r denotes the set of all possible models under differing restrictions.

$$= \underset{M_j \in M_r}{\operatorname{argmin}} \underset{b \in M_j}{\operatorname{argmin}} RSS_r - (\hat{\sigma}^2)n + 2(\hat{\sigma}^2)p$$

$$= \underset{M_j \in M_r}{\operatorname{argmin}} \underset{b \in M_j}{\operatorname{argmin}} RSS_r + 2\hat{\sigma}^2 p$$

$$= \underset{M_j \in M_r}{\operatorname{argmin}} \underset{b \in M_j}{\operatorname{argmin}} (Y - Xb)'(Y - Xb) + 2\hat{\sigma}^2 ||b||_0$$

□

7.3 Least Absolute Shrinkage and Selection Operator

Definition 7.3.1

LASSO solves the following problem:

$$\min_b (||Y - Xb||^2 + \lambda ||b||_1), \text{ where } ||\beta||_1 = \sum_{j=1}^{k} |\beta_j| \quad \text{for some } \lambda \geq 0$$

Similarly to $\hat{\beta}_r$, $\hat{\beta}_{LASSO}$ estimates are more stable than OLS. In addition many components of $\hat{\beta}_{LASSO}$ are exactly 0. Hence, we can think of LASSO as not only estimating the parameters but also performing model selection, making the model more parsimonious and thus better interpretable.

Example. Orthonormal design

We proceed with the following special case to build intuition:

$$X'X = I_k$$

$$\Rightarrow ||Y - Xb||^2 = ||(X\hat{\beta}_{OLS} + \hat{\varepsilon}) - Xb||^2$$

$$= (\hat{\beta}_{OLS} - b)'X'X(\hat{\beta}_{OLS} - b) + \hat{\varepsilon}'\hat{\varepsilon}$$

$$\Rightarrow \text{ Objective function: } \min_b \left[\sum_{j=1}^{k} (\hat{\beta}_{OLS,j} - b_j)^2 + \lambda \sum_{j=1}^{k} |b_j| \right]$$

(OLS residuals dropped as do not depend on choice of LASSO estimator b)

We can minimise each element of the overall sum separately:

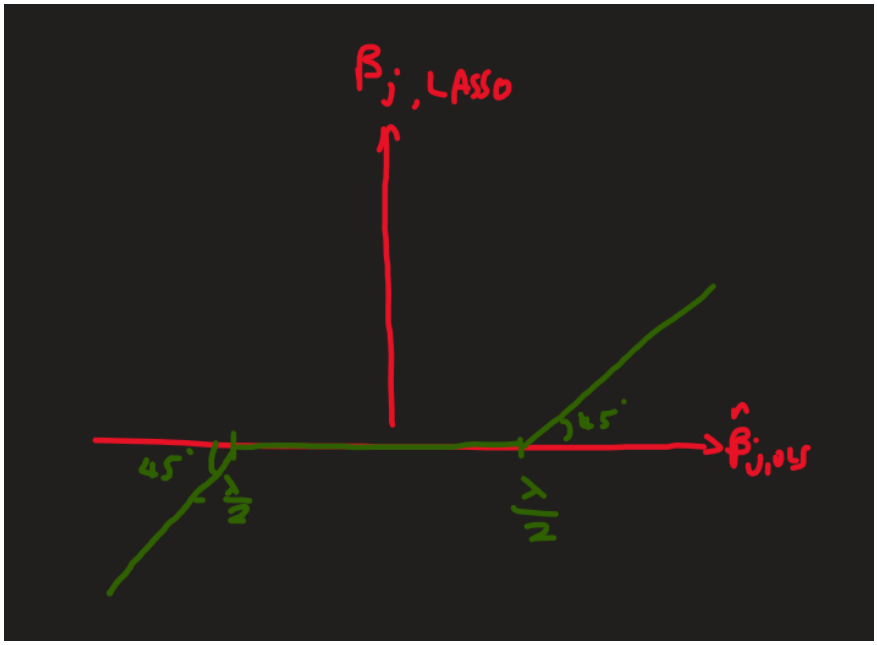
$$\min_{b_j} (b_j - \hat{\beta}_{j,OLS})^2 + \lambda |b_j| \quad \text{for each } j$$

$$\text{But:} \quad (b_j - \hat{\beta}_{j,OLS})^2 + \lambda|b_j| = \begin{cases} (b_j - \hat{\beta}_{j,OLS})^2 + \lambda b_j & \text{for } b_j \geq 0 \\ (b_j - \hat{\beta}_{j,OLS})^2 - \lambda b_j & \text{for } b_j < 0 \end{cases}$$

$$\therefore \hat{b}_{j,LASSO} = \begin{cases} \hat{\beta}_{j,OLS} - \lambda/2 & \text{for } \hat{\beta}_{j,OLS} \geq \lambda/2 \\ \hat{\beta}_{j,OLS} + \lambda/2 & \text{for } \hat{\beta}_{j,OLS} \leq -\lambda/2 \\ 0 & \text{otherwise} \end{cases}$$

7.3.1 Thresholding vs Shrinking

This estimator is an example of a soft thresholding estimator: when $|\hat{\beta}_{j,OLS}|$ is below the threshold $\lambda/2$ we set the estimator $\hat{\beta}_{j,LASSO}$ to zero, and when $|\hat{\beta}_{j,OLS}|$ is above the threshold, set estimator to $sgn\hat{\beta}_{j,OLS}(|\hat{\beta}_{j,OLS}| - \lambda/2)$.



In contrast the ridge estimator in the orthonormal design case has form:

$$\hat{\beta}_{j,ridge} = (I + \lambda I)^{-1} X'Y = \frac{\hat{\beta}_{j,OLS}}{1 + \lambda}$$

Thus it does not set any $\hat{\beta}_{j,r}$ to zero. It just shrinks $\hat{\beta}_{OLS}$.

Note:-

As in practice we do not work inside the orthonormal design case, we need to standardise the regressors to make shrinkage apply fairly. Thus we demean all regressors to avoid shrinking the constant. And we divide all variables by their standard deviation to avoid shrinking variables with larger variance more, as this is unit variant.

7.4 Appendix

7.4.1 Proof Condition Number is Ratio of Eigenvalues

Theorem 7.4.1. For any positive definite symmetric matrix A, the condition number $C == ||A||\,||A^{-1}||$ is equal to the ratio of the largest eigenvalue to the smallest eigenvalue of A.

$$||A||\,||A^{-1}|| = \frac{\lambda_{max}}{\lambda_{min}}$$

Recall that for any symmetric matrix A, we have that A is diagonalisable (by spectral theorem). This means there exists an orthonormal basis of eigenvectors $(\vec{q_1}, ..., \vec{q_n}) = Q$ that spans $\mathbb{R}^n$, with corresponding eigenvalues $(\lambda_1, ..., \lambda_n)$, where wlog $|\lambda_1| \geq |\lambda_2| \geq ... \geq |\lambda_n|$.
Thus we can write any vector $\vec{x} \in \mathbb{R}^n$ as a linear combination of these eigenvectors: $\vec{x} = Q\vec{c}$
Thus we can rewrite:

$$||A|| = \max_{x \neq 0} \frac{||Ax||}{||x||} = \max_{x \neq 0} \frac{||AQc||}{||Qc||}$$

$$= \max_{x \neq 0} \frac{||c_1\lambda_1 q_1 + ...c_n\lambda_n q_n||}{||c_1 q_1 + ...c_n q_n||}$$

Consider:

$$\frac{||Ax||^2}{||x||^2} = \frac{(c_1\lambda_1 q_1 + ...c_n\lambda_n q_n)'(c_1\lambda_1 q_1 + ...c_n\lambda_n q_n)}{(c_1 q_1 + ...c_n q_n)'(c_1 q_1 + ...c_n q_n)}$$

As eigenvectors orthonormal $q_i' q_j = 0$ for $i \neq j$ and $q_i' q_i = 1$:

$$\therefore = \frac{c_1^2\lambda_1^2 + ...c_n^2\lambda_n^2}{c_1^2 + ...c_n^2} \leq \frac{c_1^2\lambda_1^2 + ...c_n^2\lambda_1^2}{c_1^2 + ...c_n^2} = |\lambda_1|^2$$

$$\Rightarrow \frac{||Ax||}{||x||} \leq |\lambda_1|$$

But we can achieve this bound at $x = q_1$.

Lemma 7.4.1. Thus:

$$||A|| = \max_{x \neq 0} \frac{||Ax||}{||x||} = |\lambda_1|$$

For a *positive definite* symmetric matrix we know that all $\lambda_i > 0$ Thus

$$||A|| = \max(\lambda_1, ..., \lambda_n)$$

Lemma 7.4.2. A matrix A has eigenvalue λ if and only if A^{-1} has eigenvalue λ^{-1}.

Proof. Let v be an eigenvector of A with eigenvalue λ. Then:

$$Av = \lambda v \Rightarrow A^{-1}Av = A^{-1}\lambda v \Rightarrow A^{-1}v = \frac{1}{\lambda}v$$

□

Lemma 7.4.3. A positive definite $\Rightarrow A^{-1}$ (exists and is) positive definite.

Proof. A positive definite implies A invertible, as only solution to Ax=0 is x=0 (thus full rank by rank nullity theorem).
Consider for any $x \in \mathbb{R}^n$

$$x'A^{-1}x$$

Define $x = Ay$ as for any x there must exist y where $y = A^{-1}x$. Thus:

$$x'A^{-1}x = (Ay)'A^{-1}(Ay) = y'A'y$$

As $y'A'y$ is a scalar, it is equal to its transpose. Thus:

$$y'A'y = (y'A'y)' = y'Ay > 0 \quad \text{as A positive definite}$$

Thus A^{-1} is positive definite. □

Thus, as A^{-1} positive definite:

$$||A^{-1}|| = \max(\lambda_1^{-1}, ..., \lambda_n^{-1})$$

As $\lambda_i > 0$ for all i,

$$||A^{-1}|| = \frac{1}{\lambda_{min}}$$

Thus:

$$C = ||A||\,||A^{-1}|| = \frac{\lambda_{max}}{\lambda_{min}}$$

Note:-

We can apply this to the regression case by considering $A = X'X$ so long as $X'X$ is positive definite. We prove that $X'X$ is positive definite if X is full rank and $n > k$: This is because for any vector $\alpha \neq 0$ we have:

$$\alpha'X'X\alpha = (X\alpha)'X\alpha = ||X\alpha||^2 \geq 0$$

To show inequality is strict, we must show that the null space of $X'X$ is trivial. We are given $null(X)$ is empty as X is full rank by rank nullity theorem, thus we simply must show $null(X'X) = null(X)$. (only true for real matrices)

$\underline{null(X) \subseteq null(X'X)}$:

$$\text{Let vector } v \in null(X), \text{then } Xv = 0$$
$$\Rightarrow X'Xv = X'0 = 0$$
$$\Rightarrow v \in null(X'X)$$

$\underline{null(X'X) \subseteq null(X)}$:

$$\text{Let vector } v \in null(X'X), \text{then } X'Xv = 0$$
$$\Rightarrow v'X'Xv = 0$$
$$\Rightarrow ||Xv||^2 = 0$$
$$\Rightarrow Xv = 0 \quad \text{provided X is real valued}$$
$$\Rightarrow v \in null(X)$$

Thus $null(X'X) = null(X) = 0$ and $X'X$ is positive definite.

7.4.2 Proof Condition Number Bounds Solution Sensitivity

Theorem 7.4.2. For a set of simulateneous equations $Ax = b$ the condition number of A sets a bound of the sensitivity of the relative (i.e. invariant to units) solution error (in x) induced by errors in the problem vector (b). $\frac{||\Delta x||}{||x||} \leq C \frac{||\Delta b||}{||b||}$

Proof. Recall the submultiplicity property of the norm:

$$||Ax|| \leq ||A||\,||x||, ||AB|| \leq ||A||\,||B||$$

Suppose there exists a problem error in b such that $\tilde{b} = b + \Delta b$.
Then the solution to the new problem is $\tilde{x} = A^{-1}\tilde{b}$. We seek a bound on the unit invariant relative error in the solution:

Consider:

$$A(x + \Delta x) = b + \Delta b$$

$$Ax = b \Rightarrow A\Delta x = \Delta b \Rightarrow \Delta x = A^{-1}\Delta b$$

$$\Rightarrow ||\Delta x|| \leq ||A^{-1}||\,||\Delta b||$$

But:

$$Ax = b$$

$$\Rightarrow ||b|| \leq ||A||\,||x||$$

Now multiplying inequalities:

$$||\Delta x||\,||b|| \leq ||A^{-1}||\,||\Delta b||\,||A||\,||x||$$

$$\Rightarrow \frac{||\Delta x||}{||x||} \leq ||A^{-1}||\,||A|| \frac{||\Delta b||}{||b||}$$

□

7.4.3 Proof Ridge Regression Reduces Condition Number

Theorem 7.4.3. Let A be a positive definite symmetric matrix with condition number:

$$C = ||A||\,||A^{-1}|| = \frac{\lambda_{max}}{\lambda_{min}}$$

Then the condition number of $(A + \mu I)$ is:

$$\frac{\lambda_{max} + \mu}{\lambda_{min} + \mu} < \frac{\lambda_{max}}{\lambda_{min}}$$

Proof. Let Q be an orthonormal basis of eigenvectors of A, with corresponding eigenvalues $\lambda_1, ..., \lambda_n$. Then:

$$(A + \mu I)Q = AQ + \mu IQ = Q\Lambda + Q\mu I = Q(\Lambda + \mu I)$$

Thus we can diagonalise (A+μI) with eigenvalues $\lambda_1 + \mu, ..., \lambda_n + \mu$ as follows.

$$(A + \mu I) = Q(\tilde{\Lambda} + \mu I)Q'$$

Where Q is the same orthonormal basis, and all eigenvalues are simply augmented by μ.

Thus:

$$\frac{\tilde{\lambda}_{max}}{\tilde{\lambda}_{min}} = \frac{\lambda_{max} + \mu}{\lambda_{min} + \mu}$$

□

Note:-

$$Q' = Q^{-1}$$

$$\because Q'Q = QQ' = I$$

8 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.

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

8.1 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 8.1.1. 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 8.1.1: 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 $\hat{\varepsilon}_i^2$ is a biased estimate of ε^2.

8.2 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.i.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 8.2.1: Strict Stationarity

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

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

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 $(Z_t, ..., Z_{t+m})$ are stable over time.

Theorem 8.2.1. If Z_t is i.i.d., then it is strictly stationary

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

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

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 8.2.2: 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.

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

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 8.2.2. 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 8.2.1.

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.

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

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 8.2.1. 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:

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

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

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

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

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

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 8.2.3. $\hat{V}_{nw}$ 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

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

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:

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

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.

9 Functional CLT. Fixed-b asymptotics.

9.1 Fixed bandwidth approach

The choice of truncation lag G in the Newey-West method is arbitrary. There are many ways of choosing this lag optimally, see Andrews (1991) for an example.
Kiefer, Vogelsang and Bunzel (2000) show that the accuracy of the tests based on the Newey-West variance estimator may be quite poor in finite samples, specifically tests over-reject the null (the estimated variance is 'too small'). They proposed an alternative where G is chosen such that $b \equiv \frac{G+1}{T} \to 0$ as $T \to \infty$. b is known as the bandwidth, and is kept fixed. For example, when G+1=T, b is fixed at 1. Under this approach $\hat{V}$ converges to a limiting random matrix that is proportional to V. The distribution of HAC robust tests based on $\hat{V}$ don't depend on the model's parameters (i.e. the distribution is pivotal), and can be tabulated.

Definition 9.1.1: Long-run variance

Sum of all the variances and covariances of a process, i.e. $\text{Var}(\sum_{t=1}^{T} \varepsilon_t)$.

Consider the simple regression on only a constant term

$$Y_t = \beta + \varepsilon_t.$$

The OLS estimator of β is $\hat{\beta}_{OLS} = \bar{Y}$, and under serial correlation:

$$Var(\hat{\beta}_{OLS}) = \frac{1}{T} V_T = \frac{1}{T}\left(\mathbb{E}\varepsilon_t^2 + \sum_{\ell=1}^{T-1} \frac{T-\ell}{T} 2\mathbb{E}(\varepsilon_t \varepsilon_{t-\ell})\right) \neq \frac{1}{T} Var(\varepsilon_t)$$

where the first equality follows from the previous lecture. As $T \to \infty$, the variance of the OLS estimator converges to the long-run variance of ε_t.
Newey-West

$$\hat{V}_{NW} = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t^2 + \sum_{\ell=1}^{G} \frac{G+1-\ell}{G+1} \frac{2}{T} \sum_{t=1+\ell}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-\ell}$$

KVB
KVB obtains an inconsistent estimator of V_T with $G = T-1$:

$$\hat{V}_{KVB} = \frac{1}{T}\sum_{t=1}^{T} \hat{\varepsilon}_t^2 + \sum_{\ell=1}^{T-1} \frac{T-\ell}{T} \frac{2}{T} \sum_{t=1+\ell}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-\ell}$$

Here $\hat{\varepsilon}_t = Y_t - \bar{Y}$.

We can show that $\hat{V}_{KVB}$ is positive semi-definite as follows:

$$\begin{aligned}
\hat{V}_{KVB} &= \frac{1}{T}\sum_{t=1}^{T}\hat{\varepsilon}_t^2 + \sum_{\ell=1}^{T-1}\frac{T-\ell}{T}\frac{2}{T}\sum_{t=1+\ell}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-\ell} \\
&= \frac{1}{T}\mathbf{1}'\left(\frac{1}{T}\sum_{t=1+|i-j|}^{T}\hat{\varepsilon}_t\hat{\varepsilon}_{t-|i-j|}\right)\mathbf{1} \quad \text{where 1 is a T-vector of ones} \\
&= \frac{1}{T^2}\mathbf{1}'ZZ'\mathbf{1}
\end{aligned}$$

$$\text{where } \underbrace{Z}_{T\times(2T-1)} = \begin{bmatrix}
\hat{\varepsilon}_1 & \hat{\varepsilon}_2 & \cdots & \hat{\varepsilon}_T & 0 & \cdots & 0 \\
0 & \hat{\varepsilon}_1 & \cdots & \hat{\varepsilon}_{T-1} & \hat{\varepsilon}_T & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & \hat{\varepsilon}_1 & \hat{\varepsilon}_2 & \cdots & \hat{\varepsilon}_T
\end{bmatrix}$$

Thus $\hat{V}_{KVB}$ is positive semi-definite. Further, since $\sum_{t=1}^{T}\hat{\varepsilon}_t = 0$ (property of OLS residuals) the sum of elements in the T-th column is zero. Moreover, the sum of elements in the $T+i$-th column gives:

$$\begin{aligned}
\sum_{t=i+1}^{T}\hat{\varepsilon}_t &= \sum_{t=0}^{T}\hat{\varepsilon}_t - \sum_{t=0}^{i}\hat{\varepsilon}_t \\
&= 0 - \sum_{t=0}^{i}\hat{\varepsilon}_t
\end{aligned}$$

Thus,

$$\mathbf{1}'Z = \left(\sum_{i=1}^{1}\hat{\varepsilon}_i, \sum_{i=1}^{2}\hat{\varepsilon}_i, \cdots, \sum_{i=1}^{T-1}\hat{\varepsilon}_i, 0, -\sum_{i=1}^{1}\hat{\varepsilon}_i, -\sum_{i=1}^{2}\hat{\varepsilon}_i, \cdots, -\sum_{i=1}^{T-1}\hat{\varepsilon}_i\right)$$

Hence,

$$\begin{aligned}
\hat{V}_{KVB} &= \frac{1}{T^2}\mathbf{1}'ZZ'\mathbf{1} \\
&= \frac{1}{T^2}\begin{bmatrix}\sum_{i=1}^{1}\hat{\varepsilon}_i & \cdots & \sum_{i=1}^{T-1}\hat{\varepsilon}_i & 0 & -\sum_{i=1}^{1}\hat{\varepsilon}_i & \cdots & -\sum_{i=1}^{T-1}\hat{\varepsilon}_i\end{bmatrix}\begin{bmatrix}\sum_{i=1}^{1}\hat{\varepsilon}_i \\ \vdots \\ \sum_{i=1}^{T-1}\hat{\varepsilon}_i \\ 0 \\ -\sum_{i=1}^{1}\hat{\varepsilon}_i \\ \vdots \\ -\sum_{i=1}^{T-1}\hat{\varepsilon}_i\end{bmatrix} \\
&= \frac{1}{T^2}\left(\left(\sum_{i=1}^{1}\hat{\varepsilon}_i\right)^2 + \cdots + \left(\sum_{i=1}^{T-1}\hat{\varepsilon}_i\right)^2 + 0 + \left(-\sum_{i=1}^{1}\hat{\varepsilon}_i\right)^2 + \cdots + \left(-\sum_{i=1}^{T-1}\hat{\varepsilon}_i\right)^2\right) \\
&= \frac{2}{T^2}\sum_{s=1}^{T-1}\left(\sum_{i=1}^{s}\hat{\varepsilon}_i\right)^2 \\
&= \frac{2}{T}\sum_{s=1}^{T-1}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{s}\hat{\varepsilon}_i\right)^2
\end{aligned}$$

We know that $\hat{\varepsilon}_t = Y_t - \bar{Y} = \beta + \varepsilon_t - (\beta + \bar{\varepsilon}) = \varepsilon_t - \bar{\varepsilon}$.

$$\begin{aligned}
\hat{V}_{KVB} &= \frac{2}{T}\sum_{s=1}^{T-1}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{s}\hat{\varepsilon}_i\right)^2 \\
&= \frac{2}{T}\sum_{s=1}^{T-1}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{s}(\varepsilon_i - \bar{\varepsilon})\right)^2 \\
&= \frac{2}{T}\sum_{s=1}^{T-1}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{s}\varepsilon_i - \frac{s}{\sqrt{T}}\bar{\varepsilon}\right)^2 \\
&= \frac{2}{T}\sum_{s=1}^{T-1}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{s}\varepsilon_i - \frac{s}{T}\frac{1}{\sqrt{T}}\sum_{i=1}^{T}\varepsilon_i\right)^2
\end{aligned}$$

9.2 Functional CLT

We first introduce the concept of Brownian motion (or the Wiener process).

Definition 9.2.1: Brownian motion

The standard Brownian motion $W(\lambda), \lambda \in [0,1]$ is a continuous time stochastic process such that $W(\lambda_1), \ldots, W(\lambda_k)$ are jointly normally distributed for any $k \in [0,1]$ for fixed $\lambda_1, \ldots, \lambda_k$ with:

$$\mathbb{E}W(\lambda_i) = 0, \quad Cov(W(\lambda_i), W(\lambda_j)) = \min(\lambda_i, \lambda_j) \quad \forall i, j \in [0,1]$$

This is to say, it is a set of random variables indexed by λ, or alternatively a random function in C[0,1] the space of continuous functions on [0,1]. Further any interval of indices within $W(\lambda)$ is

jointly normal.

Properties

- $\text{Var}(W(\lambda_i)) = \lambda_i$

  $\text{Var}(W(\lambda_i)) = \text{Cov}(W(\lambda_i), W(\lambda_i)) = \min(\lambda_i, \lambda_i) = \lambda_i$

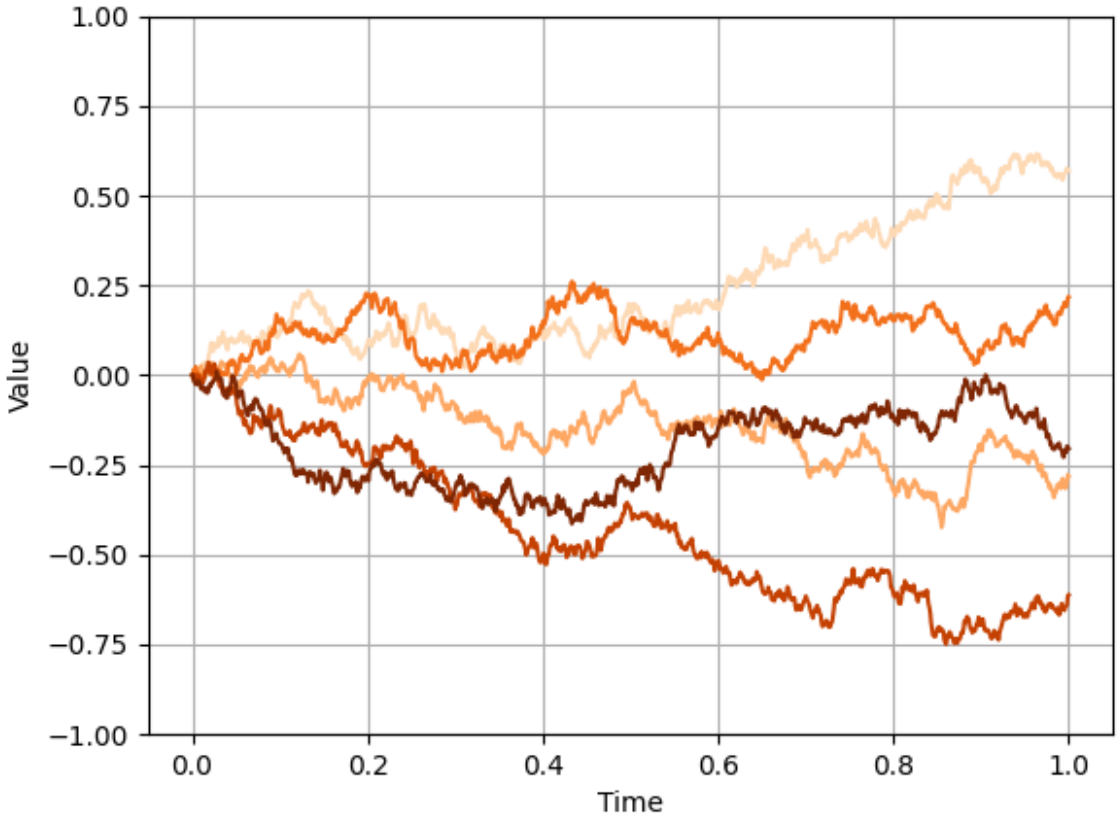
- $W(0) = 0$

  $\mathbb{E}W(0) = 0$ and $\text{Var}(W(0)) = 0$

- $W(\lambda)$ has independent increments, for every $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_k \leq 1$ the random variables $W(\lambda_1), W(\lambda_2) - W(\lambda_1), \ldots, W(\lambda_k) - W(\lambda_{k-1})$ are independent.

- $W(\lambda)$ has gaussian increments, $W(\lambda_{i+u}) - W(\lambda_i) \sim N(0, u)$

- $W(\lambda)$ is nowhere differentiable



The functional central limit theorem (FCLT) is a generalisation of the conventional CLT to function-valued random variables. To understand this we first generalise the standard notions of consistency and convergence in distribution to the space $C[0, 1]$. We define the distance between two functions using the sup-norm:

$$d(f, g) = \sup_{x \in [0,1]} |f(x) - g(x)|$$

This represents the maximum distance between the two functions.

Definition 9.2.2: Convergence in probability

A random element $\xi_T \in C[0, 1]$ converges in probability to f (that i, $\xi_T \xrightarrow{p} f$) if $\Pr[d(\xi_T, f) > \delta] \to 0$ for all $\delta > 0$.

Definition 9.2.3: Convergence in distribution

Let $\{\xi_T\}$ be a sequence of random elements in $C[0,1]$ and let F be a distribution function on $C[0,1]$, with induced probability measure π_T. Then π_T converges weakly to π, or equivalently $\xi_T \xrightarrow{d} \xi$ where ξ has probability measure π, if and only if $\int f d\pi_T \to f d\pi$ for all bounded continuous functions f: $C[0,1] \to \mathbb{R}$.

Definition 9.2.4: Continuous Mapping Theorem

If h is a continuous functional mapping $C[0,1]$ to some metric space and $\xi_T \xrightarrow{d} \xi$ then $h(\xi_T) \xrightarrow{d} h(\xi)$.

We now present some background and intuition for the functional central limit theorem.

Explanation. Let's first consider the partial sum process, defined as $X_T(\lambda) = \frac{1}{T}\sum_{t=1}^{[T\lambda]} \zeta_t$ with $\zeta \sim WN(0,1)$. The square brackets denote the floor function (i.e. the integer part of $T\lambda$). Let's see how this partial sum looks when $T = 10$ and consider $\xi_T(\lambda)$ for $\lambda = 0, 0.01, 0.1, 0.2$:

$$\lambda = 0, \quad [10 \times 0] = 0 : \quad X_{10}(0) = \frac{1}{10}\sum_{t=1}^{0} \zeta_t = 0$$

$$\lambda = 0.01, \quad [10 \times 0.01] = 0 : \quad X_{10}(0.01) = \frac{1}{10}\sum_{t=1}^{0} \zeta_t = 0$$

$$\lambda = 0.1, \quad [10 \times 0.1] = 1 : \quad X_{10}(0.1) = \frac{1}{10}\sum_{t=1}^{1} \zeta_t = \frac{\zeta_1}{10}$$

$$\lambda = 0.2, \quad [10 \times 0.2] = 2 : \quad X_{10}(0.2) = \frac{1}{10}\sum_{t=1}^{2} \zeta_t = \frac{\zeta_1 + \zeta_2}{10}$$

For a sequence of errors ζ_t

1. The function $X_T(\lambda)$ is a random step function defined on [0,1].
2. As T gets bigger the step size gets smaller, and the function becomes smoother (looking more and more like a Wiener process).

Lets consider the following for any fixed $\lambda \in [0,1]$:

$$\begin{aligned}\sqrt{T}X_T(\lambda) &= \sqrt{T}\frac{1}{T}\sum_{t=1}^{[T\lambda]} \zeta_t \\ &= \frac{1}{\sqrt{T}}\sum_{t=1}^{[T\lambda]} \zeta_t \\ &= \frac{\sqrt{[T\lambda]}}{\sqrt{T}}\frac{1}{\sqrt{[T\lambda]}}\sum_{t=1}^{[T\lambda]} \zeta_t\end{aligned}$$

Now, as $T \to \infty$:

$$\frac{\sqrt{[T\lambda]}}{\sqrt{T}} \to \sqrt{\lambda}$$

$$\frac{1}{\sqrt{[T\lambda]}} \sum_{t=1}^{[T\lambda]} \zeta_t \xrightarrow{d} N(0,1) \quad \text{by the CLT}$$

It follows from Slutsky's theorem that:

$$\sqrt{T} X_T(\lambda) \xrightarrow{d} \sqrt{\lambda} N(0,\lambda) = W(\lambda)$$

Since the above holds for any $\lambda \in [0,1]$, we might expect this holds uniformly for $\lambda \in [0,1]$. This is indeed the case, and is known as the functional central limit theorem (or Donsker's theorem for partial sums). □

The above is based on a step-function, however Alexei present a piecewise linear function where we linearly interpolate between points. This is presented below, the substantive results are the same. Let ζ_T, $t = 1, 2, ...$ be zero mean i.i.d. random variables with variance 1. Let $\xi_T(\lambda)$ be the function constructed by linearly interpolating between the partial sums of ζ at the points $\lambda = (0, \frac{1}{T}, \frac{2}{T}, \ldots, \frac{T-1}{T}, 1)$, that is:

$$\xi_T(\lambda) = \frac{1}{\sqrt{T}} \left(\sum_{t=1}^{[T\lambda]} \zeta + (T\lambda - [T\lambda]) \zeta_{[T\lambda]+1} \right)$$

so that ξ_T is a piecewise-linear random element of C[0,1] (between each point we linearly interpolate). The CLT for vector valued processes ensures that $[\xi_T(\lambda_1), \xi_T(\lambda_2), \ldots, \xi_T(\lambda_k)]$ converges in distribution to a k-dimensional normal random variable. The FCLT extends this result to hold not just for finitely many fixed values of λ, but rather for ξ_T treated as a function of λ.

Theorem 9.2.1 (Functional Central Limit Theorem). $\xi_T(\lambda) \xrightarrow{d} W$, where W is a standard Brownian motion on the unit interval.

Lemma 9.2.1 (Beveridge-Nelson decomposition). Let $u_t \sim I(1)$, where $\Delta u_t = \varepsilon_t = C(L)\zeta_t$. Then

$$u_t = C(1) \sum_{s=1}^{t} \zeta_s + C^*(L)\zeta_t + (u_0 - C^*(L)\zeta_0)$$

Proof.

$$C(L) = C(1) + [C(L) - C(1)] = C(1) + C^*(L)(1 - L)$$

where $c_j^* = -\sum_{i=j+1}^{\infty} c_i$. Why can we do this? Define A(L) = C(L) - C(1). Clearly A(1) = 0 and is thus a root of A(L). This justifies the factorisation C(L)-C(1) = (1-L)C*(L).
Thus we can write

$$\varepsilon_t = C(L)\zeta_t = C(1)\zeta_t + C^*(L)\Delta\zeta_t$$

Then because $u_t = \sum_{s=1}^{t} \varepsilon_s + u_0$ we get the result:

$$\begin{aligned} u_t = \sum_{s=1}^{t} \varepsilon_s + u_0 &= \sum_{s=1}^{t} C(1)\zeta_s + C^*(L)\Delta\zeta_s + u_0 \\ &= C(1)\sum_{s=1}^{t} \zeta_s + C^*(L)\zeta_t + u_0 - C^*(L)\zeta_0 \end{aligned}$$

□

We now consider some arbitrary linear process $\varepsilon_t = C(L)\zeta_t$ where $\zeta_t \sim iid(0,1)$ as before. By the B-N decomposition we can write $\varepsilon_t = C(1)\zeta_t + C^*(L)\Delta\zeta_t$. Consider the following:

$$\begin{aligned} \nu_T(\lambda) &= \frac{1}{\sqrt{T}}\left(\sum_{t=1}^{[T\lambda]} \varepsilon_t + (T\lambda - [T\lambda])\varepsilon_{[T\lambda]+1}\right) \\ &= \frac{1}{\sqrt{T}}\left(\sum_{t=1}^{[T\lambda]} (C(1)\zeta_t + C^*(L)\Delta\zeta_t) + (T\lambda - [T\lambda])C(1)(\zeta_{[T\lambda]+1} + C^*(L)\Delta\zeta_{[T\lambda]+1})\right) \\ &= \frac{1}{\sqrt{T}}\left(C(1)\sum_{t=1}^{[T\lambda]} \zeta_t + C^*(L)\zeta_{[T\lambda]} - C^*(L)\zeta_0 + (T\lambda - [T\lambda])C(1)(\zeta_{[T\lambda]+1} + C^*(L)\Delta\zeta_{[T\lambda]+1})\right) \\ &= C(1)\frac{1}{\sqrt{T}}\left(\sum_{t=1}^{[T\lambda]} \zeta_t + (T\lambda - [T\lambda])\zeta_{[T\lambda]+1}\right) + C^*(L)\frac{1}{\sqrt{T}}\left(\zeta_{[T\lambda]} - \zeta_0 + (T\lambda - [T\lambda])\Delta\zeta_{[T\lambda]+1}\right) \\ &= C(1)\xi_T(\lambda) + \frac{1}{\sqrt{T}}I(0) \end{aligned}$$

Since the second term is $T^{-\frac{1}{2}}$ multiplied by an I(0) process, it converges to zero in probability. Further, since $\xi_T(\lambda) \xrightarrow{d} W(\lambda)$, this suggests that $C(1)\xi_T \xrightarrow{d} C(1)W(\lambda)$ and $\nu_T(\lambda) \xrightarrow{d} C(1)W(\lambda)$.

Theorem 9.2.2. $\nu_T(\lambda) \xrightarrow{d} C(1)W(\lambda)$

For a more rigorous proof of convergence see Stock (1994) pg 2750. [1]

9.3 Fixed-b asymptotics

Recall the definition of Brownian motion (ignoring the smoothing terms):

$$\xi_T(\lambda) = \frac{1}{\sqrt{T}}\sum_{t=1}^{[T\lambda]} \varepsilon_t.$$

Thus we can see that

$$\begin{aligned} \frac{1}{\sqrt{T}}\sum_{i=1}^{s} \varepsilon_i &= \frac{1}{\sqrt{T}}\sum_{i=1}^{T\times\frac{s}{T}} \varepsilon_i = \xi_T(\frac{s}{T}) \\ \frac{1}{\sqrt{T}}\sum_{i=1}^{T} \varepsilon_i &= \xi_T(1) \end{aligned}$$

[1] This topic is such a fucking rabbit hole, there is no chance this is understandable to our tiny reg monkey brains. This shit is so convoluted don't even bother going further.

Consider our representation from earlier:

$$\begin{aligned}\hat{V}_{KVB} &= \frac{2}{T}\sum_{s=1}^{T-1}\left(\frac{1}{\sqrt{T}}\sum_{i=1}^{s}\varepsilon_i - \frac{s}{T}\frac{1}{\sqrt{T}}\sum_{i=1}^{T}\varepsilon_i\right)^2 \\ &= \frac{2}{T}\sum_{s=1}^{T-1}\left(\xi_T(\frac{s}{T}) - \frac{s}{T}\xi_T(1)\right)^2 \\ &\approx 2\int_0^1 \left(\xi_T(\lambda) - \lambda\xi_T(1)\right)^2 d\lambda \quad \lambda \coloneqq \frac{s}{T}\end{aligned}$$

The approximation follows from the fact that the second line is a Riemann sum, where as $T \to \infty$ the approximation error converges to zero.
We know that $\xi_T(\lambda) \xrightarrow{d} c(1)W(\lambda)$, thus by the continuous mapping theorem:

$$\begin{aligned}\hat{V}_{KVB} &\xrightarrow{d} 2\int_0^1 \left(c(1)W(\lambda) - \lambda c(1)W(1)\right)^2 d\lambda \\ &= 2[c(1)]^2\int_0^1 \left(W(\lambda) - \lambda W(1)\right)^2 d\lambda\end{aligned}$$

The right hand side is proportional to $[c(1)]^2$, which is the long-run variance of ε_t.

Example (Long-run variance). $\varepsilon_t = C(L)\zeta_t = c_0\zeta_t + c_1\zeta_{t-1} + \ldots$
Long run variance is defined differently to before, here it is $\text{Var}(\varepsilon_t)$.

$$\begin{aligned}Var(\varepsilon_t) &= Var(C(L)\zeta_t) \\ &= Var(C(1)\zeta_t) \quad \text{since } \zeta_t \text{ is i.i.d. the lags don't matter} \\ &= C(1)^2 Var(\zeta_t) \\ &= C(1)^2 \quad \text{since } \zeta_t \sim iid(0,1)\end{aligned}$$

If we now consider the t-statistic (based on $\hat{V}_{KVB}$) for testing $H_0 : \beta = 0$:

$$\begin{aligned}t &= \frac{\hat{\beta}}{\sqrt{Var(\hat{\beta})}} = \frac{\bar{Y}}{\sqrt{\frac{1}{T}\hat{V}_{KVB}}} = \frac{\beta + \bar{\varepsilon}}{\frac{1}{\sqrt{T}}\sqrt{\hat{V}_{KVB}}} \\ &\overset{H_0}{=} \frac{\sqrt{T}\bar{\varepsilon}}{\sqrt{\hat{V}_{KVB}}} = \frac{\frac{1}{\sqrt{T}}\sum_{j=1}^{T}\varepsilon_j}{\sqrt{\hat{V}_{KVB}}} = \frac{\frac{1}{\sqrt{T}}\xi_T(1)}{\sqrt{\hat{V}_{KVB}}} \\ &\xrightarrow{d} \frac{c(1)W(1)}{\sqrt{2[c(1)]^2\int_0^1 \left(W(\lambda) - \lambda W(1)\right)^2 d\lambda}} \\ &= \frac{c(1)W(1)}{c(1)\sqrt{2\int_0^1 \left(W(\lambda) - \lambda W(1)\right)^2 d\lambda}} \\ &= \frac{W(1)}{\sqrt{2\int_0^1 \left(W(\lambda) - \lambda W(1)\right)^2 d\lambda}}\end{aligned}$$

This doesn't depend on $c(1)$ (the model parameters), meaning the distribution is pivotal. Thus it can be simulated and critical values recorded. The pdf is given below, note how the (normalised) KVB distribution has fatter tails than the normal distribution.

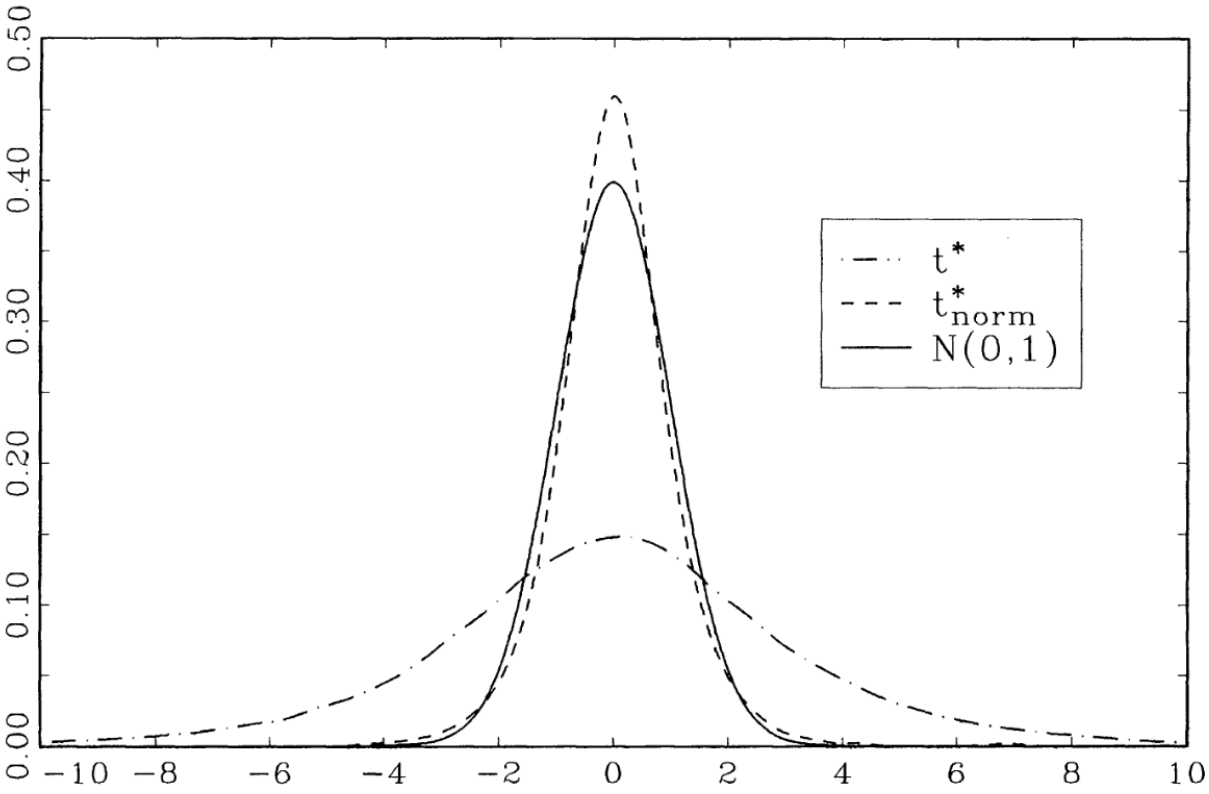


FIGURE 1.—Densities of t^*, t^*_{norm}, and $N(0,1)$.

KVB show that in finite samples these tests may outperform tests based on Newey-West standard errors. At a high level, if there is lots of serial correlation KVB is much better, whereas if it is only minor NW is probably fine. NW suffers when samples are small and serial correlation is large. KVB, like HAC estimator tests, suffer from serious size distortions (although less so) if the data have highly persistent serial correlation and are close to being non-stationary. KVB also show the finite sample power of their test dominates finite sample power of HAC tests.

10 Probit. Maximum Likelihood.

10.1 Binary choice

Suppose we don't have a continuous dependent variable, rather it is binary: $y_i = \{0, 1\}$. We could still use OLS here, let's check out the assumptions:

(OLS0) (y_i, x_i) is an i.i.d. sequence

✓ Binary y_i doesn't break this, we can still have an i.i.d. sequence

(OLS1) $E(x_i x_i')$ is finite non-singular

✓ Binary y_i doesn't affect this

(OLS2) $E(y_i|x_i) = x_i'\beta$

? $E(y_i|x_i) = 1 \times P(y_i = 1|x_i) + 0 \times P(y_i = 0|x_i) = P(y_i = 1|x_i) \stackrel{?}{=} x_i'\beta$

Hence, for OLS2 to hold we need use the linear probability model.

(OLS3) Var $(y_i|x_i) = \sigma^2$

$\times$ $Var(y_i|x_i) = E(y_i^2|x_i) - E(y_i|x_i)^2 = E(y_i|x_i) - E(y_i|x_i)^2 = x_i'\beta(1 - x_i'\beta)$

using $y^2 = y$. Hence OLS3 cannot hold, we do have heteroskedasticity.

(OLS4) $E\varepsilon_i^4 < \infty, \quad E\|x_i\|^4 < \infty$

✓ May still hold

We can fix the heteroskedasticity with GLS or White standard errors, but the linear probability model is more of a problem. This model does not restrict predicted probabilities to be between 0 and 1, and the use of any other model will violate OLS2 meaning OLS will not be consistent. The standard alternative is to use a function of the form

$$P(y_i = 1|x_i) = F(x_i'\beta)$$

where F$(\cdot)$ is a known CDF, typically assumed to be symmetric about zero, so that $F(u) = 1 - F(-u)$. The standard choices for F are

- Logistic: $F(u) = \frac{e^u}{1+e^u}$, known as the **logit** model
- Normal: $F(u) = \Phi(u)$, known as the **probit** model

This is identical to the latent variable model

$$\begin{aligned} y_i* &= x_i'\beta + \varepsilon_i \\ \varepsilon_i &\sim F(\cdot) \\ y_i &= \begin{cases} 1 & \text{if } y_i* > 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

Since then

$$\begin{aligned} P(y_i = 1|x_i) &= P(y_i* > 0|x_i) \\ &= P(x_i'\beta + \varepsilon_i > 0|x_i) \\ &= P(\varepsilon_i > -x_i\beta|x_i) \\ &= 1 - F(-x_i'\beta) \\ &= F(x_i'\beta) \end{aligned}$$

10.2 Maximum likelihood estimation

The probit model is typically estimated by the method of maximum likelihood (ML). Consider the typical setup:

$$\begin{aligned} z_1, \dots, z_n \overset{i.i.d.}{\sim} f(\cdot|\theta) \quad &\rightarrow \quad L(\theta) = \prod_{i=1}^{n} f(z_i|\theta) \\ \log L(\theta) = \ell(\theta) \quad &= \quad \sum_{i=1}^{n} \log f(z_i|\theta) \\ \hat{\theta}_{ML} \quad &= \quad \arg\max_{\theta} \ell(\theta) \end{aligned}$$

This is known as a *parametric model*, it requires the specification of the distribution of the data up to an unknown parameter θ.
A key property is that the expected log-likelihood is maximised at the true value of the parameter vector θ_0. Set $Z = (z_1, \dots, z_n)$.

Theorem 10.2.1. $\theta_0 = \arg\max_{\theta} \mathbb{E}(\log L(\theta)|Z)$

The proof is presented in Lecture 11 using KL divergence. This motivates estimating θ by finding the value which maximises log-likelihood.

Example (OLS using MLE).

$$f(Y_1, \dots, Y_n|X, \beta, \sigma^2) \quad : \quad L = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - X_i'\beta)^2}{2\sigma^2}}$$

$$\begin{aligned} \Rightarrow \ell = \log L &= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{(Y_i - X_i'\beta)^2}{2\sigma^2} \\ &= n \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{\sum_{i=1}^{n}(Y_i - X_i'\beta)^2}{2\sigma^2} \\ &= \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^{n}(Y_i - X_i'\beta)^2}{2\sigma^2} \end{aligned}$$

Hence, the FOCs are:

$$\frac{\partial \ell}{\partial \beta} = -\frac{\sum_{i=1}^{n}(-X_i)(Y_i - X_i'\beta)}{\sigma^2} = 0 \tag{10.1}$$

$$\frac{\partial ell}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(Y_i - X_i'\beta)^2}{2\sigma^4} = 0 \tag{10.2}$$

(10.1) $\Rightarrow \sum_{i=1}^{n} X_i Y_i - X_i X_i' \hat{\beta}_{ML} = 0$
$\Rightarrow X'Y - X'X\hat{\beta}_{ML} = 0$
$\Rightarrow \hat{\beta}_{ML} = (X'X)^{-1}X'Y = \hat{\beta}_{OLS}$

(10.2) $\Rightarrow n\sigma^2 = \sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_{ML})^2$
$\Rightarrow \sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i'\hat{\beta}_{ML})^2$

Thus, $\hat{\beta}_{OLS}$ is actually the MLE for β, so it has the desirable properties discussed in Lecture 11. However, the ML estimator for the variance is biased due to not correcting for the loss in degrees of freedom from estimating $\hat{\beta}_{ML}$.

Consider the problem of estimating θ if you have a vector of data Z with the joint density of its elements given by $f(z|\theta)$.

Definition 10.2.1: Score

The score of the likelihood function is the vector of partial derivatives with respect to the parameters.

$$\frac{\partial}{\partial\theta}\log f(Z|\theta)$$

Theorem 10.2.2. If $\log f(Z|\theta)$ is second differentiable and the support of Z doesn't depend on θ then the score has mean zero:

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(Z|\theta)\right] = 0$$

Proof.

$$\begin{aligned}\mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(Z|\theta)\right] &= \int_{\mathbb{R}} \frac{\frac{\partial}{\partial\theta}f(z|\theta)}{f(z|\theta)} f(z|\theta)dz \\ &= \frac{\partial}{\partial\theta}\int_{\mathbb{R}} f(z|\theta)dz \\ &= \frac{\partial}{\partial\theta}1 \\ &= 0\end{aligned}$$

□

Definition 10.2.2: Fisher information

The covariance matrix of the score is known as the Fisher information

$$I(\theta) = Var\left(\frac{\partial}{\partial\theta}\log f(Z|\theta)\right) = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log f(Z|\theta)\right)^2 |\theta\right]$$

Note:-

Because the likelihood of θ given Z is always proportional to the probability $f(Z|\theta)$; their logarithms necessarily differ by a constant that is independent of θ, and the derivatives are necessarily equal. Thus one can substitute in $\log L(\theta) = \ell(\theta)$ for $\log f(Z|\theta)$ in the above definitions.

The Fisher information is a way of measuring the amount of information that an observable Z carries about the unknown parameter θ. If f is sharply peaked with respect to changes in θ, it is easy to indicate the "correct" value of θ from the data, or equivalently, that the data Z provides a lot of information about the parameter θ. If f is flat and spread-out, then it would take many samples of Z to estimate the true value of θ. Note that $I(\theta) \geq 0$. Near the ML estimate, low Fisher information suggests the maximum appears flat, that is, there are many nearby values with similar log-likelihood. Conversely, high Fisher information indicates the maximum is sharp.

Claim 10.2.1. If we have n i.i.d. distributions (from n samples) then the Fisher information will be n times the Fisher information of a single sample from the common distribution.

$$I_n(\theta) = nI_1(\theta)$$

Lemma 10.2.1 (Information equality). The variance of the score is equal to the negative expected value of the Hessian matrix of the log-likelihood.

$$I(\theta) = Var\left(\frac{\partial}{\partial\theta}\log f(Z|\theta)\right) = -\mathbb{E}\left(\frac{\partial^2}{\partial\theta\partial\theta'}\log f(Z|\theta)\right)$$

Proof. Let $\boldsymbol{Z}$ be an m-component column vector of random variables, not necessarily i.i.d. To ease notation, we denote their joint density as $f(\boldsymbol{Z}|\boldsymbol{\theta}) \equiv f$. Also note all expectations are conditional on $\boldsymbol{\theta}$, and integrals are multiple integrals over $z_1, \ldots, z_n$.

$$\begin{aligned}
\mathbb{E}\frac{\partial^2 \log f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial \log f}{\partial\boldsymbol{\theta}'}\right)\right] \\
&= \mathbb{E}\left[\frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{1}{f}\frac{\partial f}{\partial\boldsymbol{\theta}'}\right)\right] \\
&= \mathbb{E}\left[-\frac{1}{f^2}\frac{\partial f}{\partial\boldsymbol{\theta}}\frac{\partial f}{\partial\boldsymbol{\theta}'} + \frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \\
&= -\mathbb{E}\left[\left(\frac{1}{f}\frac{\partial f}{\partial\boldsymbol{\theta}}\right)\left(\frac{1}{f}\frac{\partial f}{\partial\boldsymbol{\theta}'}\right)\right] + \mathbb{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]
\end{aligned}$$

To obtain the information equality, we need to show the second term is zero.

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] &= \int_{\mathbb{R}} f\frac{1}{f}\frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}d\boldsymbol{Z} \\
&= \int_{\mathbb{R}} \frac{\partial^2 f}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}d\boldsymbol{Z} \\
&= \int_{\mathbb{R}} \frac{\partial}{\partial\boldsymbol{\theta}}\left(\frac{\partial f}{\partial\boldsymbol{\theta}'}\right)d\boldsymbol{Z} \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\int_{\mathbb{R}} \frac{\partial f}{\partial\boldsymbol{\theta}'}d\boldsymbol{Z} \quad \text{(we can interchange these because we are economists)} \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\frac{\partial}{\partial\boldsymbol{\theta}'}\int_{\mathbb{R}} f d\boldsymbol{Z} \quad \text{(what even is a regularity condition)} \\
&= \frac{\partial}{\partial\boldsymbol{\theta}}\frac{\partial}{\partial\boldsymbol{\theta}'}1 \\
&= 0
\end{aligned}$$

□

Note:-

All we are assuming here is that we can interchange the order of differentiation and integration; a set of sufficient conditions for this are:

1. The function $\frac{\partial}{\partial\boldsymbol{\theta}}f(\boldsymbol{Z}|\boldsymbol{\theta})$ is continuous in $\boldsymbol{Z}$ and in $\boldsymbol{\theta} \in \Theta$ where Θ is an open set.
2. The integral $\int f(\boldsymbol{Z}|\boldsymbol{\theta})d\boldsymbol{Z}$ exists.
3. $\int \left|\frac{\partial}{\partial\boldsymbol{\theta}}f(\boldsymbol{Z}|\boldsymbol{\theta})\right| d\boldsymbol{Z} < M < \infty$ for all $\boldsymbol{\theta} \in \Theta$

Misspecification and the information equality

Suppose our random variables have joint density f as before, but we specify that they have joint density g instead. As before

$$\mathbb{E}_f\frac{\partial^2 \log g}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = -\mathbb{E}_f\left[\left(\frac{1}{g}\frac{\partial g}{\partial\boldsymbol{\theta}}\right)\left(\frac{1}{g}\frac{\partial g}{\partial\boldsymbol{\theta}'}\right)\right] + \mathbb{E}_f\left[\frac{1}{g}\frac{\partial^2 g}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]$$

where the f subscript denotes the fact that we are taking the expectation with respect to the true distribution. Previously we made progress because the integrand contained $f\frac{1}{f} = 1$, however we now have $f\frac{1}{g}$ which doesn't simplify. In general, under misspecification

$$\mathbb{E}_f\left[\frac{1}{g}\frac{\partial^2 g}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] \neq 0$$

and the IE doesn't hold. Note: this does not exclude the possibility that this expected value is after all zero and the IE holds, it just generally isn't.

Theorem 10.2.3 (Cramer-Rao lower bound)**.** If $\tilde{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$, then we have the following bound on its variance

$$Var(\tilde{\boldsymbol{\theta}}|\boldsymbol{Z}) \geq [I(\boldsymbol{\theta})]^{-1}$$

These are both matrices, meaning this inequality tells us the difference between the left and right hand sides is positive semi-definite.
This result is similar to the Gauss-Markov theorem which established a lower bound for unbiased estimators in homoskedastic linear regression.

Example (Information bound for normal regression)**.** We will apply the CRLB conditionally on X. Define the expected Hessian

$$\mathbb{E}(H) = \begin{bmatrix} \mathbb{E}\left(\frac{\partial^2\ell}{\partial\beta\partial\beta'}|X\right) & \mathbb{E}\left(\frac{\partial^2\ell}{\partial\beta\partial\sigma^2}|X\right) \\ \mathbb{E}\left(\frac{\partial^2\ell}{\partial\sigma^2\partial\beta'}|X\right) & \mathbb{E}\left(\frac{\partial^2\ell}{\partial\sigma^2\partial\sigma^2}|X\right) \end{bmatrix}$$

Recall the log likelihood

$$\ell = \frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{\sum_{i=1}^n (Y_i - X_i'\beta)^2}{2\sigma^2}$$

Thus we have second derivatives

$$\begin{aligned}
\frac{\partial^2\ell}{\partial\beta\partial\beta'} &= \frac{\partial}{\partial\beta'}\frac{\sum_{i=1}^n X_i(Y_i - X_i'\beta)}{\sigma^2} &&= -\frac{1}{\sigma^2}\sum_{i=1}^n X_iX_i' = \frac{1}{\sigma^2}X'X \\
\frac{\partial^2\ell}{\partial\beta\partial\sigma^2} &= \frac{\partial}{\partial\sigma^2}\frac{\sum_{i=1}^n X_i(Y_i - X_i'\beta)}{\sigma^2} &&= -\frac{\sum_{i=1}^n X_i(Y_i - X_i'\beta)}{\sigma^4} = -\frac{1}{\sigma^4}X'(Y - X\beta) \\
\frac{\partial^2\ell}{\partial\sigma^2\partial\sigma^2} &= \frac{n}{2}\frac{1}{\sigma^4} - \frac{\sum_{i=1}^n (Y_i - X_i'\beta)^2}{\sigma^6} &&= \frac{n}{2}\frac{1}{\sigma^4} - \frac{1}{\sigma^6}(Y - X\beta)'(Y - X\beta)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow \mathbb{E}(H) &= \begin{bmatrix} \mathbb{E}\left[\frac{1}{\sigma^2}X'X|X\right] & \mathbb{E}\left[-\frac{1}{\sigma^4}X'(Y - X\beta)|X\right] \\ \mathbb{E}\left[-\frac{1}{\sigma^4}X'(Y - X\beta)|X\right] & \mathbb{E}\left[\frac{n}{2}\frac{1}{\sigma^4} - \frac{1}{\sigma^6}(Y - X\beta)'(Y - X\beta)|X\right] \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & \frac{n}{2}\frac{1}{\sigma^4} - \frac{n\sigma^2}{\sigma^6} \end{bmatrix} \\
&= \begin{bmatrix} \frac{1}{\sigma^2}X'X & 0 \\ 0 & -\frac{n}{2}\frac{1}{\sigma^4} \end{bmatrix}
\end{aligned}$$

The block diagonal matrix can be inverted to find the lower bound on asymptotic conditional variance

$$[I(\theta)]^{-1} = \begin{bmatrix} \sigma^2(X'X)^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

The variance of $\hat{\beta}_{OLS} = \hat{\beta}_{ML}$ meets the CRLB. Thus we have the following theorem

Theorem 10.2.4. In the normal regression, OLS is the Best Unbiased Estimator (BUE).

This result should be distinguished from the Gauss-Markov Theorem that $\hat{\beta}_{OLS}$ is minimum variance among those estimators that are unbiased and linear in y. Theorem 10.2.4 says that $\hat{\beta}_{OLS}$ is minimum variance in a larger class of estimators that includes non-linear unbiased estimators. This stronger statement is obtained under the normality assumption which is not assumed in the Gauss-Markov Theorem. Put differently, the Gauss-Markov Theorem does not exclude the possibility of some non-linear estimator beating OLS, but this possibility is ruled out by the normality assumption.
As we have already seen, the ML estimator of σ^2 is biased, so the CRLB does not apply. But the OLS estimator $\hat{\sigma}^2$ of σ^2 is unbiased, does it achieve the bound? We know $\frac{(n-k)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-k)$, and $Var(\chi^2(p) = 2p)$. Thus

$$\begin{aligned} & Var\left(\frac{(n-k)\hat{\sigma}^2}{\sigma^2}\right) = 2(n-k) \\ \Rightarrow & \frac{(n-k)^2}{\sigma^4} Var(\hat{\sigma}^2) = 2(n-k) \\ \Rightarrow & Var(\hat{\sigma}^2) = \frac{2\sigma^4}{n-k} \end{aligned}$$

Therefore $\hat{\sigma}^2$ does not attain the CRLB $2\sigma^4/n$. However it can be shown that an unbiased estimator with variance lower than $\hat{\sigma}^2$ does not exist.

11 ML Asymptotics. Likelihood Ratio Test.

More rigour: Amemiya (1985)

11.0.1 Consistency of ML

Let z_i be iid with density $f(z;\theta_0)$ for $i = 1,\dots,n$.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(z_i;\theta)$$

By Khinchine's LLN for any θ

$$\frac{1}{n}\sum_{i=1}^{n} \log f(z_i;\theta) \xrightarrow{p} \mathbb{E}_{\theta_0}[\log f(z;\theta)]$$

We can invoke KLLN as given z_i iid $\Rightarrow$ any function of z_i is also iid. We also need to assume the expectation exists. This is taken over the value of the true parameter, but the conditioned θ runs across the real line.

Proposition 11.0.1.

$$\hat{\theta}_{ML} \xrightarrow{p} \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{\theta_0}[\log f(z;\theta)]$$

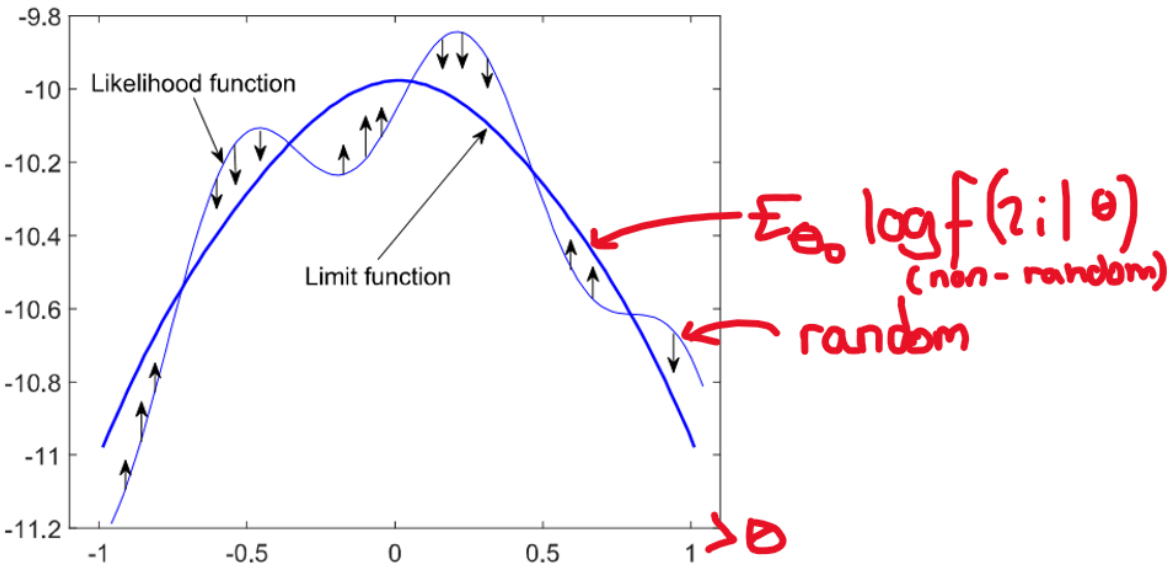
Proof.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log f(z_i;\theta) = \underset{\theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log f(z_i;\theta)$$

But:

$$\frac{1}{n}\sum_{i=1}^{n} \log f(z_i;\theta) \xrightarrow{p} \mathbb{E}_{\theta_0}[\log f(z;\theta)] \text{ uniformly} \Rightarrow \underset{\theta}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^{n} \log f(z_i;\theta) \xrightarrow{p} \underset{\theta}{\operatorname{argmax}}\, \mathbb{E}_{\theta_0}[\log f(z;\theta)]$$

□



Proposition 11.0.2. $E_{\theta_0} \log f(z_i;\theta)$ is maximised at the true value of parameter θ_0

Proof. Consider the KL divergence between $f(z;\theta)$ and $f(z;\theta_0)$:

$$E_{\theta_0} \log \frac{f(z;\theta_0)}{f(z;\theta)}$$

By construction the minimiser of the KL divergence must be the maximiser of $E_{\theta_0} \log f(z;\theta)$. By Jensen's inequality:

$$= -E_{\theta_0} \frac{\log f(z;\theta)}{\log f(z;\theta_0)} \geq -\log E_{\theta_0} \frac{f(z;\theta)}{f(z;\theta_0)} = -\log \int \frac{f(z;\theta)}{f(z;\theta_0)} f(z;\theta_0) dz = -\log 1 = 0$$

But we can achieve this bound by setting $\theta = \theta_0$ is a maximiser of $E_{\theta_0} \log f(z;\theta)$. □

Note:-

If there exists another maximiser θ_1, we must have $f(z;\theta_0) = f(z;\theta_1)$ for all z. In such a case, we say that a case, we say that the parameter is non-identified.
In the linear regression example , $\theta = (\beta', \sigma^2)$, would not be identified if $X'X$ has rank lower than k (perfect multicollinearity).
Pointwise convergence is not enough for consistency of the θ_{ML} estimator. Sufficient conditions are given by uniform convergence and "enough" curvature of $E_{\theta_0} \log f(z;\theta)$ around θ_0.

11.0.2 Asymptotic Normality of ML

Proposition 11.0.3.

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0))$$

where $I(\theta_0)$ is the Fisher information matrix:

$$I_1(\theta_0) = Var\left[\frac{\partial}{\partial\theta} \log f(z;\theta_0)\right] = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2} \log f(z;\theta_0)\right] = E_{\theta_0}(H_1) = \mathbb{E}_{\theta_0}(\frac{H}{n})$$

Note $I_1(\theta_0)$ is the Fisher information for a single observation.
Define $I(\theta_0)$ as the Fisher information matrix for the sample.

This is the sum of the Fisher information for each observation $I(\theta_0) = nI_1(\theta_0)$, since $\log(z_i;\theta)$ is a function of iid z_i, and so is iid.

$$Var\left[\frac{\partial}{\partial\theta}L(\theta_0)\right] = Var\left[\frac{\partial}{\partial\theta}\sum_{i=1}^{n}\log f(z_i;\theta_0)\right] = nVar\left[\frac{\partial}{\partial\theta}\log f(z;\theta_0)\right] \text{ since iid}$$

Proof. Let $\Psi(\theta) = \frac{\partial}{\partial\theta}\frac{1}{n}L(\theta;Z)$, where

$$L(\theta;Z) = \sum_{i=1}^{n}\log f(z_i;\theta)$$

$\hat{\theta}_{ML}$ can be obtained as a solution to the likelihood equation: $\Psi(\hat{\theta}_{ML}) = \frac{\partial}{\partial\theta}\frac{1}{n}L(\hat{\theta}_{ML};Z) = 0$
Assuming consistency, $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$, it makes sense to expand $\Psi(\hat{\theta}_{ML})$ around θ_0:

$$\Psi(\hat{\theta}_{ML}) = 0 = \Psi(\theta_0) + (\hat{\theta}_{ML} - \theta_0)\Psi'(\theta_0) + \frac{1}{2}(\hat{\theta}_{ML} - \theta_0)^2\Psi''(\tilde{\theta})$$

where $\tilde{\theta}$ is between $\hat{\theta}_{ML}$ and θ_0, such that the Taylor expansion is exact by the MVT.
Therefore when θ is scalar,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{-\sqrt{n}\Psi(\theta_0)}{\Psi'(\theta_0) + (\hat{\theta}_{ML} - \theta_0)\Psi''(\tilde{\theta})/2}$$

But under the random sampling assumption:

$$\Psi(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\theta}\log f(z_i;\theta_0)\bigg|_{\theta=\theta_0} \xrightarrow{p} \frac{\partial}{\partial\theta}\mathbb{E}_{\theta_0}\log f(z;\theta_0)\bigg|_{\theta=\theta_0}$$

And with the Lindeberg-Levy CLT:

$$-\sqrt{n}\Psi(\theta_0) \xrightarrow{d} N\left(0, Var\left(\frac{\partial}{\partial\theta}\log f(z_i,\theta_0)\right)\right) = N\left(0, \frac{1}{n}I(\theta_0)\right)$$

Next, by Khinchine's LLN:

$$\Psi'(\theta_0) \xrightarrow{p} \frac{\partial^2}{\partial\theta^2}\mathbb{E}_{\theta_0}\log f(z;\theta_0)\bigg|_{\theta=\theta_0}$$

Finally, $(\hat{\theta}_ML - \theta_0)\Psi''(\tilde{\theta}) \xrightarrow{p} 0$ i.e. is $o_p(1)$, since $\hat{\theta}_{ML} - \theta_0 = o_p(1)$ and $\Psi''(\tilde{\theta})$ converges to a finite constant (Amemiya 1985, p. 67,ch 4).
Therefore by Slutsky's theorem:

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{-\sqrt{n}\Psi(\theta_0)}{\Psi'(\theta_0) + (\hat{\theta}_{ML} - \theta_0)\Psi''(\tilde{\theta})/2} \xrightarrow{d} \frac{N(0, \frac{1}{n}I(\theta_0))}{\mathbb{E}_{\theta_0}\frac{\partial^2}{\partial\theta^2}\log f(z;\theta_0)}$$

$$\therefore \sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} \frac{N(0, \frac{1}{n}I(\theta_0))}{-\frac{1}{n}I(\theta_0)} = N(0, nI^{-1}(\theta_0))$$

In other words in large samples, $\hat{\theta}_{ML}$ is approximately normally distributed with mean θ_0 and variance $I^{-1}(\theta_0)$. □

This generalises straightforwardly to the case of a vector θ.

NOTE: $I(\theta_0)$ refers to the sample Fisher information matrix, which is $n \times I_1(\theta)$ - the finite infor-

mation matrix of one observation. Thus saying $\hat{\theta}_{ML}$ is approximately normally distributed with mean θ_0 and variance $I^{-1}(\theta_0)$, means its variance is in fact $(1/n)I_1^{-1}(\theta_0)$, which goes to zero for large n and thus we have $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$ as we found earlier.

11.1 Asymptotic efficiency of the maximum likelihood estimator

Proposition 11.1.1. θ_{ML} **is asymptotically efficient:**
Lowest asymptotic variance among all estimators that are

- asymptotically normal
- asymptotically unbiased
- regular

Recall the Cramér-Rao result:
Any unbiased estimator of θ_0 has variance no smaller than the inverse of the Fisher information.
While suggestive of asymptotic efficiency here, it is a *finite* sample result and thus does not imply this.

11.1.1 Irregular Estimators

Hodges' Estimator

$$\theta_H = \begin{cases} \hat{\theta}_{ML} & \text{if } |\hat{\theta}_{ML}| \geq n^{-1/4} \\ 0 & \text{if } |\hat{\theta}_{ML}| < n^{-1/4} \end{cases}$$

Case 1: $\theta_0 \neq 0$
$\hat{\theta}_H$ is asymptotically equivalent to $\hat{\theta}_{ML}$. This is because $\hat{\theta}_{ML} \xrightarrow{p} \theta_0 \neq 0$, and $n^{-1/4} \to 0$, thus $|\hat{\theta}_{ML}| \geq n^{-1/4}$ will be true asymptotically, so $\hat{\theta}_H = \hat{\theta}_{ML}$ asymptotically.

Note:-

Big O, Little O Notation
$f(x) \in O(g(x))$ if $\exists\ K > 0$ and x_0 such that $|f(x)| \leq Kg(x)$ for all $x > x_0$.
$f(x) \in o(g(x))$ if $\forall\ K > 0\ \exists\ x_0$ such that $|f(x)| < Kg(x)$ for all $x > x_0$.

Product Rule: $f(x) = O(g(x))$ and $h(x) = O(k(x)) \Rightarrow f(x)h(x) = O(g(x)k(x))$
Little O $\Rightarrow$ Big O: $f(x) = o(g(x)) \Rightarrow f(x) = O(g(x))$

In probability:
$X_n \in O_P(\alpha_n)$ if $\forall \varepsilon > 0\ \exists\ K > 0$ and x_0 such that $\Pr(|f(x)| \leq Kg(x)) > 1 - \varepsilon$ for all $x > x_0$.
i.e.X_n/α_n is bounded up to an exceptional event of arbitrarily small (but fixed) positive probability, i.e. the ratio is 'bounded in probability.
$f(x) \in o_p(\alpha_n)$ if $\forall \varepsilon > 0\ \forall\ K > 0\ \exists\ x_0$ such that $\Pr(|f(x)| < Kg(x)) > 1 - \varepsilon$ for all $x > x_0$.

Case 2: $\theta_0 = 0$

Proposition 11.1.2. $|\hat{\theta}_{ML}| = O_p(n^{-1/2})$

Since $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, I_1^{-1}(\theta_0))$, we know $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \in O_p(1)$, since its variance (and expectation) is finite and constant wrt n and so must be bounded in probability.
$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \frac{\hat{\theta}_{ML} - \theta_0}{1\sqrt{n}} = O_p(1)$

$\Rightarrow \hat{\theta}_{ML} - \theta_0 = O_p(n^{-1/2})$*
$\therefore |\hat{\theta}_{ML}| = O_p(n^{-1/2})$

*(also loose intution from the product rule of normal big O, $\sqrt{n} = O_p(\sqrt{n})$)
Where let $\hat{\theta}_{ML} - \theta_0 \in O_p(\alpha_n)$
$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \in O_p(1) \Rightarrow O_P(\sqrt{n})O_P(\alpha_n) = O_P(\sqrt{n}\alpha_n) = O_P(1)$
$\Rightarrow \alpha_n = 1/\sqrt{n}$

Proposition 11.1.3. $|\hat{\theta}_{ML}| = o_p(n^{-1/4})$
$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, I_1^{-1}(\theta_0))$ and $n^{-1/4} \xrightarrow{p} 0$ Thus by Slutsky's theorem: $n^{1/4}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} 0 \Rightarrow n^{1/4}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{p} 0$
$\Rightarrow \frac{(\hat{\theta}_{ML} - \theta_0)}{1/n^{1/4}} \xrightarrow{p} 0$
$\Rightarrow \lim_{n\to\infty} \mathbb{P}(|\frac{(\hat{\theta}_{ML} - \theta_0)}{1/n^{1/4}} - 0| > \varepsilon) = 0 \, \forall \varepsilon > 0$
$\Rightarrow \hat{\theta}_{ML} - \theta_0 = o_p(n^{-1/4})$ with the defintion of o_p
Intuitively as $n^{-1/4} > n^{-1/2}$, it makes sense that dividing by $n^{-1/4}$ binds more strictly (sends to zero) than dividing by $n^{-1/2}$, which already binds in probability (sends to a constant variance distribution).

When $\theta_0 = 0$ Hodges' estimator clearly imporves over $\hat{\theta}_{ML}$ because $|\hat{\theta}_{ML}| = o_p(n^{-1/4})$, which implies $\hat{\theta}_H = 0$ exactly asymptotically (with zero variance) for sufficiently large n.

But in finite samples, Hodge's estimator behaves poorly for $\theta \approx 0$. Asymptotically, this is reflected in its erratic behaviour when true value of parameter is drifting towards zero so that $\theta = h/\sqrt{n}$ for some $h \in \mathbb{R}$. For such sequences of θ, $\hat{\theta}_H$ is inconsistent. we have:

$$\sqrt{n}(\hat{\theta}_H - \theta_0) = \sqrt{n}(\hat{\theta}_H - h/\sqrt{n}) \to -h$$

Regular estimators would have the same asymptotic distribution for any value of $h/\sqrt{n}$ (a small change in parameter should not change the distribution of the estimator too much)

11.2 Likelihood Ratio Test

Suppose that the likelihood function is in general given by $L(\theta; Z) \equiv f(Z, \theta)$, where Z is a vector of data and θ is a vector of parameters. Consider testing the null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset$.

The likelihood ratio test is defined by the following procedure:
Reject H_0 if

$$LR(Z) = \frac{\sup_{\theta\in\Theta_0} L(\theta; Z)}{\sup_{\theta\in\Theta_0\cup\Theta_1} L(\theta; Z)} > c$$

. where c is chosen as a critical value so as to satisfy $max_{\theta\in\Theta_0} \Pr(LR(Z) > c) = \alpha$, where α is the significance level of the test (probability of Type 1 error).

Theorem 11.2.1. Neyman-Pearson Lemma:
When $\Theta_0 = \theta_0$ and $\Theta_1 = \theta_1$ (i.e. single values of the parameter vector), the likelihood ratio test is the most powerful test of size α.

11.2.1 Likelihood Ratio Test of linear restrictions in normal regression

Proposition 11.2.1. We show the LR test to be equivalent to the F test, as the LR statistic is a monotone transformation of the F statistic.

Consider a hypothesis $R\beta = r$ about coefficients of linear regression with normal errors:

$$Y = X\beta + \varepsilon, \varepsilon|X \sim N(0, \sigma^2 I)$$

The uncoonstrained ML estimates of β and σ^2 are in such a model $\hat{\beta}_{OLS}$ and $\hat{\sigma^2}_{Ml} = RSS_u/n$.

We have $\log(max_\theta L(Y, \theta|X))$ (unrestricted)

$$= \log\left[\left(\frac{1}{\sqrt{2\pi}|\sigma^2 I|^{-1/2}}\right)^n \exp(-\frac{1}{2\sigma^2}(Y - X\beta)'(Y - X\beta))\right]\Big|_{\theta=\hat{\theta}_{ML}}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\hat{\sigma}^2_{ML}) - \frac{1}{2\hat{\sigma}^2_{ML}}(Y - X\hat{\beta}_{ML})'(Y - X\hat{\beta}_{ML})$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\frac{RSS_u}{n}) - \frac{1}{2}\frac{RSS_u}{RSS_u/n}$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(RSS_u) - \frac{n}{2}$$

Similarly under the restrictions we can show that:

$$\log(max_{\theta\in\Theta_0} L(Y, \theta|X)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(RSS_r) - \frac{n}{2}$$

where RSS_r is the restricted residual sum of squares.

Therefore the log likelihood ratio statistic for the test of $R\beta = r$ against $R\beta \neq r$ is:

$$LR = -2\left[-\frac{n}{2}\log(\frac{RSS_r}{n}) + \frac{n}{2}\log(\frac{RSS_u}{n})\right] = n\log(\frac{RSS_r}{RSS_u})$$

$$= n\left[\log\left(\frac{p}{n-k}\frac{(RSS - r - RSS_u)/p}{RSS_u/(n-k)} + 1\right)\right]$$

$$= n\left[\log\left(\frac{p}{n-k}\frac{W}{p} + 1\right)\right]$$

Thus LR statistic is a monotone transformation of the F statistic $= W/p$ so that LR test and F test must be equivalnet in the context of testing the linear restrictions in normal regression model. But unlike F test, LR test provides a formidable tool for testing hyportheses in much broader contexts.

Finding c:

$$P(LR > c) = P(n\log(1 + \frac{p}{n-k}F) > c)$$

$$= P(F > \frac{n-k}{p}(e^{c/n} - 1)) = \alpha$$

Thus as we know the F distribution:

$$\frac{n-k}{p}(e^{c/n} - 1) = F_{1-\alpha}(p, n-k)$$

$$\Rightarrow c = n\log(F_{1-\alpha}(p, n-k)\frac{p}{n-k} + 1)$$

12 Probit Asymptotics. Testing Inequality Restrictions.

12.1 Asymptotics of Probit

The conditional likelihood for the probit model is

$$L(\beta) = \prod_{i=1}^{n} f(y_i|x_i,\beta)$$

$$= \prod_{i=1}^{n} P(y_i = 1|x_i,\beta)^{y_i} P(y_i = 0|x_i,\beta)^{1-y_i}$$

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

and the conditional log-likelihood is

$$l(\beta) = \sum_{i=1}^{n} y_i \ln \Phi(x_i'\beta) + (1-y_i)\ln(1-\Phi(x_i'\beta))$$

Given x_i, y_i we can find $\hat{(}\beta_{ML})$ that maximises this function.

Theorem 12.1.1. The Probit Estimator
(i) $(\hat{\beta}_{prob}) \xrightarrow{p} \beta_0$
(ii) $\sqrt{n}(\hat{\beta}_{prob} - \beta_0) \xrightarrow{d} N(0, I_1^{-1})$ where,

$$I_1 = E\left(\frac{\phi^2 x_i x_i'}{\Phi(1-\Phi)}\right)$$

where $\Phi = \Phi(x_i'\beta_0)$ and $\phi = \frac{d}{dt}\Phi(t)\big|_{t=(x_i'\beta_0)}$
Under the following assumptions:

- (Prob 0) $\{y_i, x_i\}_{i=1}^n$ is an iid sequence with binary y_i
- (Prob 1) $E(x_i x_i')$ is finite nonsingular
- (Prob 2) $Pr(y_i = 1|x_i) = \Phi(x_i'\beta)$

Proof. The theorem follows from the fact that $\hat{\beta}_{prob}$ is an MLE estimator. Indeed, the consistency statement is implied (as we have assumed correct distribution).

For the asymptotic normality, recall:

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \xrightarrow{d} N(0, n(I^{-1}(\theta_0)))$$

where $\hat{\theta}_{ML} = \hat{\beta}_{prob}$ and $\theta_0 = \beta_0$

$$I(\beta) = Var\left(\frac{d}{d\beta}L(\beta)\right)\left(\overset{\beta=\beta_0}{=} E\left((\frac{d}{d\beta}L(\beta)\frac{d}{d\beta'}L(\beta))\right)\right)$$

Since $E(\frac{d}{d\beta}L(\beta))\big|_{\beta=\beta_0} = 0$, as we have seen previously that this maximises the likelihood function.

$$\frac{dL(\beta)}{d\beta} = \sum_{i=1}^{n}\left(\frac{y_i}{\Phi(x_i'\beta)}\phi(x_i'\beta)x_i - \frac{(1-y_i)}{1-\Phi(x_i'\beta)}\phi(x_i'\beta)x_i\right)$$

$$= \sum_{i=1}^{n}\frac{y_i - \Phi(x_i'\beta)}{\Phi(x_i'\beta)(1-\Phi(x_i'\beta))}\phi(x_i'\beta)x_i$$

$$Var\frac{dL}{d\beta} = E(Var\frac{dL}{d\beta}|x_i) + Var(E(\frac{dL}{d\beta}|x_i))$$

$$Var(E(\frac{dL}{d\beta}|x_i))\big|_{\beta=\beta_0} = Var(0|x_i) = 0$$

$$E(Var\frac{dL}{d\beta}|x_i)\big|_{\beta=\beta_0} = EVar\left(\sum_{i=1}^{n}\frac{y_i - \Phi(x_i'\beta)}{\Phi(x_i'\beta)(1-\Phi(x_i'\beta))}\phi(x_i'\beta)x_i|x_i\right)$$

Because iid,

$$= nEVar\left(\frac{y_i - \Phi(x_i'\beta)}{\Phi(x_i'\beta)(1-\Phi(x_i'\beta))}\phi(x_i'\beta)x_i|x_i\right)$$

since $\Phi(x_i'\beta)$ is constant conditioning on x_i:

$$= nEVar\left(\frac{y_i}{\Phi(x_i'\beta)(1-\Phi(x_i'\beta))}\phi(x_i'\beta)x_i|x_i\right)$$

$$= nE(\frac{\phi}{\Phi(1-\Phi)}x_iVar(y_i|x_i)\frac{\phi}{\Phi(1-\Phi)}x_i')$$

$Var((y_i|x_i))$ is given by:

$$E(y_i^2|x_i) - E(y_i|x_i)^2$$

by Prob 2:

$$= (\Phi(1^2)) - (\Phi(1))^2 = \Phi - \Phi^2$$

Thus:

$$I(\beta_0) = nE\left(\frac{\phi^2 x_i x_i'}{\Phi(1-\Phi)}\right)$$

$$I_1(\beta_0) = E\left(\frac{\phi^2 x_i x_i'}{\Phi(1-\Phi)}\right)$$

□

12.1.1 Interpreting coefficients in Probit

Unlike for linear regression, $\beta = \beta_0$ cannot be interpreted as the marginal effect of x on y. Here the coefficient measures the drect impact of regressors only on the (unobserved and scaled) underlying index. We are not typically interested in this slope but actually:

Definition 12.1.1

Probit Marginal Effects

$$\frac{\partial Pr(y=1|x)}{\partial x_j} = \frac{\partial}{\partial x_j}(x'\beta) = \phi(x'\beta)\beta_j$$

here x_j refers to the j-th component of vector x as opposed to the j-th observation of this vector.
Note that the effect depends not only on the value of x_j but also other variables.
Marginal Effect at the Average:

$$\frac{\partial Pr(y=1|x)}{\partial x_j}\Big|_{x=\bar{x}} = \phi(\bar{x}'\beta)\beta_j$$

Standard errors, use the delta method, where $\theta = \beta$

$$\frac{\partial Pr(y=1|x)}{\partial x_j}\Big|_{x=\bar{x}} = \phi(\bar{x}'\beta)\beta_j = g(\beta)$$

Thus we have

$$\hat{Var}(\phi(\bar{x}'\beta)\beta_j) = \frac{1}{n}\frac{\partial g(\hat{\beta})}{\partial \beta'}\hat{I}_1^{-1}\frac{\partial g(\hat{\beta})}{\partial \beta}$$

Average Marginal Effect:

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\partial Pr(y=1|x)}{\partial x_j}\Big|_{x=x_i} = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i'\beta)\beta_j$$

Note:-

The ratios of the effects of two variables are equal to the ratio of their coefficients, and are therefore comparable for probit and logit models.

$$= \frac{\frac{\partial Pr(y=1|x)}{\partial x_j}}{\frac{\partial Pr(y=1|x)}{\partial x_k}} = \frac{\phi(x'\beta)\beta_j}{\phi(x'\beta)\beta_k} = \frac{\beta_j}{\beta_k}$$

Thus, while $\hat{\beta}_j$ and $\hat{\beta}_k$ are not directly interpretable as the absolute marginal effects, their ratio can be interpreted as the ratio of the marginal effects.

12.2 Testing Inequality Constraints

Instead of simply deleting or attempting to explain inconsistent signs of parameters in the estimating equation, we can statistically test whether or not the signs of the true values of these estimates are consistent with researcher beliefs.

Consider the stylised framework, of a normal regression model with two explanatory variables:

$$Y = \beta_1 X1 + \beta_2 X2 + \varepsilon$$

where $\varepsilon|X_1, X_2 \sim N(0, \sigma^2 I_n)$ In addition, assume $\sigma^2 = 1$ known and X_1, X_2 orthonormal. That is, $X'X = I_2$ where $X = [X_1, X_2]$

Suppose that we would like to test:

$$H_0 : \beta_1 \geq 0 \text{ and } \beta_1 \geq 0 \text{ vs. } H_1 : \beta_1 < 0 \text{ or } \beta_1 < 0$$

Let's derive the LR statistic. Recall that we assumed that it is known that $\sigma^2 = 1$. Thus, the log-likelihood function is: $\log(maxL(Y, \theta|X))$ without the restrictions:

$$= -\frac{n}{2}log(2\pi) - \frac{1}{2}(Y - X\hat{\beta}_{OLS})'(Y - X\hat{\beta}_{OLS})$$

$\log(maxL(Y, \theta|X))$ with the restrictions:

$$= -\frac{n}{2}log(2\pi) - \frac{1}{2}\min_{b_1, b_2 \geq 0}(Y - Xb)'(Y - Xb)$$

Hence,

$$LR = \min_{b_1, b_2 \geq 0}(Y - Xb)'(Y - Xb) - \frac{1}{2}(Y - X\hat{\beta}_{OLS})'(Y - X\hat{\beta}_{OLS})$$

Note that

$$\begin{aligned}(Y - Xb)'(Y - Xb) &= Y'Y - 2b'X'Y + b'X'Xb \\ &= Y'Y - 2b'X'(X\hat{\beta}_{OLS} + \hat{\varepsilon}) + b'X'Xb \\ &= Y'Y - 2b'\hat{\beta}_{OLS} + b'b\end{aligned}$$

and, similarly

$$(Y - X\hat{\beta}_{OLS})'(Y - X\hat{\beta}_{OLS}) = Y'Y - \hat{\beta}'_{OLS}\hat{\beta}_{OLS}$$

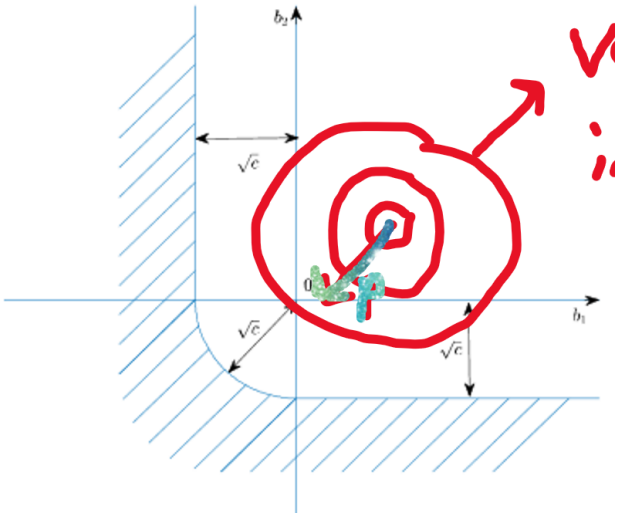
Thus,

$$LR = \min_{b_1, b_2 \geq 0}(\hat{\beta}_{OLS} - b)'(\hat{\beta}_{OLS} - b)$$

so that

$$LR = \begin{cases} 0 & \text{if } \hat{\beta}_{1,OLS} \geq 0 \text{ and } \hat{\beta}_{2,OLS} \geq 0 \\ \hat{\beta}^2_{1,OLS} & \text{if } \hat{\beta}_{1,OLS} < 0 \text{ and } \hat{\beta}_{2,OLS} \geq 0 \\ \hat{\beta}^2_{2,OLS} & \text{if } \hat{\beta}_{1,OLS} \geq 0 \text{ and } \hat{\beta}_{2,OLS} < 0 \\ \hat{\beta}^2_{1,OLS} + \hat{\beta}^2_{2,OLS} & \text{if } \hat{\beta}_{1,OLS} < 0 \text{ and } \hat{\beta}_{2,OLS} < 0 \end{cases}$$

Thus the LR statistic is the squared distance from $\hat{\beta}_{OLS}$ to the positive quadrant in $\mathbb{R}^2$:



Finding c:
If β_{OLS} ends up in the striped region (Ω), LR test rejects. For the test with 5% significance level, we need to choose the critical value c so that

$$\max_{\beta_1,\beta_2\geq 0} Pr(LR > c) = 0.05$$

We can think of $Pr(LR > c)|\beta = \beta_0$ as the volume of the probability density of $\hat{\beta}_{OLS}$ that lies inside the critical region Ω.

Recall that $\hat{\beta}_{OLS}|X \sim N(\beta_0, \sigma^2(X'X)^{-1})$ In this special case:

$$\hat{\beta}_{OLS}|X \sim N(\beta_0, I_2)$$

Thus with this geometric interpretation

$$Pr(LR > c) = \int_\Omega \frac{1}{2\pi} exp\{-\frac{(z-\beta)'(z-\beta)}{2}\}dz$$

$$= \int_{\Omega-\beta} \frac{1}{2\pi} exp\{-\frac{z'z}{2}\}dz$$

where $\Omega - \beta$ a linear shift of Ω. For any β from the positive quadrant (consistent with H_0), $\Omega - \beta \subseteq \Omega$, with equality only at $\beta = 0$. Therefore

$$\max_{\beta_1,\beta_2\geq 0} Pr(LR > c) = Pr(LR > c)|_{\beta_1=0,\beta_2=0}$$

This corresponds to the 'worst case' null, i.e. the case where it is hardest to differentiate between the null and alternative hypotheses.

On the other hand, if $\beta_1 = \beta_2 = 0$, then $\hat{\beta}_{OLS}$ is distributed as $N(0, I_2)$, thus $\hat{\beta}^2_{1,OLS} \sim \chi^2(1)$ and $\hat{\beta}^2_{2,OLS} \sim \chi^2(1)$, and $\hat{\beta}^2_{1,OLS} + \hat{\beta}^2_{2,OLS} \sim \chi^2(2)$

Thus,

$$LR = \begin{cases} 0 & \text{with probability } 1/4 \\ \chi^2(1) & \text{with probability } 1/2 \\ \chi^2(2) & \text{with probability } 1/4 \end{cases}$$

Thus c is the 0.95 quantile of the mixture of chi-squared distributions. This can be found numerically.

$$F_{LR}(x) = 1/4F_0(x) + 1/2F_{\chi^2(1)}(x) + 1/4F_{\chi^2(2)}(x)$$

13 Errors in variables. Endogeneity. IV

13.0.1 Classical measurement error

Suppose we observe noisy versions of the variables we would like to observe. We obtain data y_i and x_i for $i = 1, ..., n$ while the true values are y_i^* and x_i^*. Also assume:

$$x_i = x_i^* + \nu_i$$

$$y_i = y_i^* + \eta_i$$

where:

$$E(\nu_i) = E(\eta_i) = 0$$
$$E(x_i^* \nu_i) = 0, E(y_i^* \eta_i) = 0$$
$$E(x_i^* \eta_i) = 0, E(y_i^* \nu_i) = 0$$
$$E(\nu_i \eta_i) = 0$$

Given that $E(y_i^*|x_i^*) = x_i^* \beta$, if we proceeded as if there were no measurement error we would estimate the following by OLS:

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y$$
$$= (\frac{1}{n} \sum_{i=1}^{n} x_i x_i')^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i'$$
$$= (\frac{1}{n} \sum_{i=1}^{n} (x_i^* + \nu_i)(x_i^* + \nu_i)')^{-1} \frac{1}{n} \sum_{i=1}^{n} (x_i^* + \nu_i)(y_i^* + \eta_i)'$$
$$= \left(\frac{1}{n} \sum_{i=1}^{n} \underset{\downarrow p}{x_i^* x_i'^*} + \underset{\downarrow p}{x_i^* \nu_i'} + \underset{\downarrow p}{\nu_i x_i'^*} + \underset{\downarrow p}{\nu_i \nu_i'} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \underset{\downarrow p}{x_i^* y_i'^*} + \underset{\downarrow p}{x_i^* \eta_i'} + \underset{\downarrow p}{y_i^* \nu_i'} + \underset{\downarrow p}{\nu_i \eta_i'}$$
$$E(x_i^* x_i'^*)\ E(x_i^* \nu_i')\ E(\nu_i x_i'^*)\ E(\nu_i \nu_i') \qquad E(x_i^* y_i'^*)\ E(x_i^* \eta_i')\ E(y_i^* \nu_i')\ E(\nu_i \eta_i')$$

Lemma 13.0.1. Suppose we have sequences of random variables X_n and Y_n such that $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then $X_n Y_n \xrightarrow{p} XY$.

Proof in Appendix

Thus as plim of a sum is the sum of plims, and $1/x$ is a continuous function, we can use the CMT to argue:

$$\xrightarrow{p} (E x_i^* x_i'^* + E \nu_i \nu_i')^{-1} (E(x_i^* (x_i'^* \beta + \varepsilon_i)))$$
$$= (E x_i^* x_i'^* + E \nu_i \nu_i')^{-1} (E x_i^* x_i'^*) \beta \neq \beta$$

We have here an asymmetric bias. In the multiple regression case, the bias will depend on the interaction between explanatory variables.

In the univariate case the above reduces to :

$$\hat{\beta}_{OLS} \xrightarrow{p} \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_{\nu}}\beta$$

This represents an attenuation bias of $\hat{\beta}_{OLS}$ toward zero.

13.0.2 Endogeneity with Errors

Any measurement error in the dependent variable is subsumed in the error term as follows:

$$y_i = y_i^* + \eta_i = x_i^*\beta + (\varepsilon_i + \eta_i)$$

The new error term $(\varepsilon_i + \eta_i)$ creates no problem for estimation as η_i is uncorrelated with x_i^*. However if there were measurement errors in x_i^* then we have:

$$y_i = x_i\beta + \underbrace{(\varepsilon_i + \eta_i - \nu_i'\beta)}_{u_i}$$

In this case, u_i is correlated with x_i

$$E(x_i u_i') = E(x_i u_i) = E(x_i(\varepsilon_i + \eta_i - \nu_i'\beta)) = E(x_i\varepsilon_i) + E(x_i\eta_i) - E(x_i\nu_i'\beta_i)$$

$$= -E(x_i\nu_i')\beta = -E(x_i^* + \nu_i)\nu_i'\beta$$

$$= \cancel{E(x_i^*\nu_i')\beta} + E(\nu_i\nu_i')\beta \neq \vec{0}$$

Thus as error term is correlated with x_i, (OLS2') $Ex_iu_i = 0$ does not hold, and OLS is inconsistent. Two solutions:

Solution 1:
If we can estimate $E(\nu_i\nu_i')$ we can undo the error in estimation by using the following:

$$E[x_i x_i'] = E[x_i^* x_i'^*] + E[\nu_i\nu_i']$$

But this is not usually possible.
Solution 2:
Suppose we get another independent measure of $x*$ such that

$$w_i = x_i^* + \tau_i$$

where τ_i is uncorrelated with any of $y_i^*, x_i^*, \eta_i, \nu_i$.

$$E[w_i x_i'] = E[(x_i^* + \tau_i)(x_i^* + \nu_i)'] = E[x_i^* x_i'^*]$$

$$E[w_i y_i] = E[(x_i^* + \tau_i)(y_i^* + \eta_i)] = E[x_i^* y_i^*]$$

Then if $E[w_i x_i']$ is invertible:

$$E[w_i x_i']^{-1}E[w_i y_i] = E[x_i^* x_i'^*]^{-1}E[x_i^* y_i^*] = [x_i^* x_i'^*]^{-1}E[x_i^* x_i'^*]\beta = \beta$$

So $\hat{\beta}_{IV} = (W'X)^{-1}W'Y$ is consistent for β.

This can also be derived directly by multiplying w_i to the estimating equation and taking expectations:

$$w_i y_i = w_i x_i'\beta + w_i e_i$$

$$E[w_i y_i] = E[w_i x_i']\beta + \cancel{E[(x_i^* + \tau_i)(\varepsilon_i + \eta_i - \nu_i'\beta)]}$$

$$\beta = E[w_i x_i']^{-1} E[w_i y_i]$$

$$\Rightarrow \hat{\beta}_{IV} = (\sum_{i=1}^{n} w_i x_i')^{-1} \sum_{i=1}^{n} w_i y_i = (W'X)^{-1} W'Y$$

where

$$w = \begin{pmatrix} w_1' \\ \vdots \\ w_n' \end{pmatrix}, X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}$$

13.1 Endogeneity

Consider the following model

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

where $E(\varepsilon_i|x_i) \neq 0$ in violation of OLS2 and OLS2'. To distinguish this from the *regression*, we call the above equation a structural equation and β a structural parameter. A structural equation represents a causal link, rather than just an empirical association.

When $E(\varepsilon_i|x_i) \neq 0$, we say that x_i is endogenous. When this occurs, usually this is only by a few components of x_i being correlated with ε_i. The components causing this are referred to as endogenous and the rest exogenous. We then can partition x_i into the exogenous part x_{1i} and the endogenous part x_{2i} by rearranging.

The endogeneity problem may not only be caused through measurement error, but also joint determination, reverse causality or omitted variable bias.

13.1.1 Joint Determination: Supply and Demand

Consider the following model where q_i and p_i are determined jointly by the demand equation

$$\text{Demand: } q_i = -\beta_d p_i + \varepsilon_{di}$$

$$\text{Supply: } q_i = -\beta_s p_i + \varepsilon_{si}$$

In matrix notation:

$$\begin{pmatrix} 1 & \beta_d \\ 1 & \beta_s \end{pmatrix} \begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} \varepsilon_{di} \\ \varepsilon_{si} \end{pmatrix}$$

$$\begin{pmatrix} q_i \\ p_i \end{pmatrix} = \frac{1}{-\beta_s - \beta_d} \begin{pmatrix} -\beta_s & -\beta_d \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{di} \\ \varepsilon_{si} \end{pmatrix}$$

$$= \begin{pmatrix} (\beta_s \varepsilon_{di} + \beta_d \varepsilon_{si})/(\beta_s + \beta_d) \\ (\varepsilon_{si} - \varepsilon_{di})/(\beta_s + \beta_d) \end{pmatrix}$$

Thus neither $E[p_i \varepsilon_{di}]$ nor $E[p_i \varepsilon_{si}]$ are zero, so p_i is endogenous.

Running an OLS of q_i on p_i will give a biased estimate of β_d. We estimate $Cov(q_i, p_i)/Var(p_i)$, and assuming demanad and supply shocks uncorrelated:

$$Cov(q_i, p_i)/Var(p_i) = \frac{\frac{\beta_s}{(\beta_s+\beta_d)^2} Var(\varepsilon_{di}) - \frac{\beta_d}{(\beta_s+\beta_d)^2} Var(\varepsilon_{si})}{\frac{1}{(\beta_s+\beta_d)^2} Var(\varepsilon_{di}) - \frac{1}{(\beta_s+\beta_d)^2} Var(\varepsilon_{si})}$$

$$= \beta_s \frac{Var(\varepsilon_{di})}{Var(\varepsilon_{di}) - Var(\varepsilon_{si})} - \beta_d \frac{Var(\varepsilon_{si})}{Var(\varepsilon_{di}) - Var(\varepsilon_{si})}$$

That is, some linear combination of the slopes of the demand and supply curves.

13.1.2 Omitted Variables

Another example of endogeneity would be a structural equation connecting two variables that are both chosen by economics agents, say, wage and educaiton

$$wage_i = \beta_1 + \beta_2 educ_i + \varepsilon_i$$

Both $wage_i$ and $educ_i$ may be affected by person i's ability or some other factor belonging to ε_i. Here the structural equation is thought of reflecting a causal relationship, which would be observed if we could randomly assign education levels to people independent of ability or anything else. In reality as this does not occur we cannot rule out that this choice may have been affected by other factors influencing wage.

13.2 Instrumental Variables

We formalise the device used in Solution 2 of the measurement error. Consider a linear regression model:

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

where $E(\varepsilon_i|x_i) = 0$ and so x_i is endogenous. Suppose we have a variable w_i such that:

$$E[w_i\varepsilon_i] = 0 \text{ (exogeneity)}$$

$$E[w_i w_i'] > 0 \text{ (no redundant instruments, =non singular?)}$$

$$E[w_i x_i'] \text{ has full column rank (relevance)}$$

Then

$$0 = E[w_i\varepsilon_i] = E[w_i(y_i - x_i'\beta)] = E[w_i y_i] - E[w_i x_i']\beta$$
$$\Rightarrow \beta = E[w_i x_i']^{-1} E[w_i y_i]$$

This motivates $\hat{\beta}_{IV}$

$$\hat{\beta}_{IV} = (W'X)^{-1}W'Y = \left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} w_i y_i$$

$$\xrightarrow{p} E[w_i x_i']^{-1} E[w_i y_i] = E[w_i x_i']^{-1} E[w_i x_i']\beta + E[w_i\varepsilon_i] = \beta$$

Note:-

Understanding the Relevance Condition
Suppose we start with the regression model:

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

where x_{2i} is a scalar endogenous variable, and x_{1i} is a $(k-1)\times 1$ vector of exogenous variables.

We define w_i as a vector of instruments, where exogenous variables instrument themselves and z_i is a scalar IV for x_{2i}

Given $E[w_i w_i']$ nonsingular

$$E[w_i x_i'] \text{ full column rank} \Leftrightarrow E(w_i w_i')^{-1} E[w_i x_i'] \text{ full column rank}$$

But this represents the set of population coefficients in a regression model of x_i on w_i.

$$\because E(w_i w_i')^{-1} E[w_i x_i'] = E(w_i w_i')^{-1} E[w_i'(x_{(1,1i)} \dots x_{(k-1,1i)}\ x_{2i})]$$

$$= (\vec{\beta}_{1,1} \dots \vec{\beta}_{k-1,1} \dots \vec{\beta}_2)$$

Consider $\vec{\beta}_{1,1}$:
This represents the coefficient of $x_{1,1i}$ in the regression of $x_{1,1i}$ on w_i. But clearly as $x_{1,1i}$ is included as a regressor in w_i, $\vec{\beta}_{1,1} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.
For $\vec{\beta}_2$:
This represents β in the regression model:

$$x_{2i} = w_i'\vec{\beta} + \nu_i = \pi_1 x_{1,1i} + \ \dots\ + \pi_{k-1} x_{k-1,1i} \ + \pi_k z_i + \nu_i$$

Therefore:

$$E(w_i w_i')^{-1} E[w_i x_i'] = \begin{pmatrix} I_{k-1} & \vec{\pi} \\ \vec{0}' & \pi_k \end{pmatrix}$$

where $\vec{\pi}$ is $(\pi_1 \dots \pi_{k-1})'$

Clearly then for this to be full rank (implying the relevance condition), **we need** $\pi_k \neq 0$.
This means that the instrument z_i is correlated with x_{2i}, even after the effects of all the other exogenous variables have been controlled for.

Example. Demand and Supply shifters
Suppose the market is a local fish market as in Graddy(1985). We may think supply would be affected by weather offshore w_i, so that:

$$q_i = -\beta_s p_i + \gamma w_i + \varepsilon_{si}$$

whereas demand will not be directly affected by w_i so that:

$$q_i = \beta_d p_i + \varepsilon_{di}$$

Then we can use w_i as an instrument for p_i in the estimation of the demand (but not supply) equation. As these two equations need to equal, we can guarantee the relevance condition for $\gamma \neq 0$, as we can express p_i as a function of w_i.

Example. Education and Wage
Angrist and Krueger (1991) propose the quarter of birth indicator as the instrument for education. Due to compulsory education laws in the United States, you cannot drop out from school until you are 16, so people who are born in the first quarter of the year, being oldest in their class, may drop out more often than those born in the other quarters. This insures that $E[w_i x_i] \neq 0$, whereas, arguably, the quarter of birth should not be related to any other determinants of your wage, so that $E[w_i \varepsilon_i] = 0$.

13.3 Appendix

Proof. Approach 1 (CMT only):

$$X_n \xrightarrow{p} X \text{ and } Y_n \xrightarrow{p} Y \Leftrightarrow (X_n, Y_n) \xrightarrow{p} (X, Y)$$

Define a continuous function:

$$f(x, y) = xy$$

Then by the CMT (applied to vector):

$$f(X_n, Y_n) \xrightarrow{p} f(X, Y)$$

$$\Rightarrow X_n Y_n \xrightarrow{p} XY$$

Approach 2 (Slutsky + CMT):

$$X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$$

$$Y_n \xrightarrow{p} X$$

By Slutsky's theorem:

$$X_n Y_n \xrightarrow{d} XY$$

$$\Rightarrow X_n Y_n \xrightarrow{p} XY$$

(this implication only holds for RHS constant) □

14 2SLS. Control Function. Endogeneity and overidentification tests.

14.1 Under, just and overidentification

Consider again the linear regression model, with $\vec{x}_{1i}$ exogenous and $\vec{x}_{2i}$ endogenous.

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

Then take instrument:

$$w_i = \begin{pmatrix} x_{1i} \\ z_i \end{pmatrix}$$

with x_{1i} instrumenting for themselves (included exogenous variables) and z_i instrumenting for x_{2i} (excluded exogenous variables).

If w_i l-dimensional and x_i k-dimensional:

$$\underbrace{E[w_i y_i]}_{l\times 1} = \underbrace{E[w_i x_i']}_{l\times k}\underbrace{\beta}_{k\times 1}$$

- If $l < k$, then we have **underidentification**
- If $l = k$, then we have **just identification**
- If $l > k$, then we have **overidentification**

The relevance condition, $E[w_i x_i']$ full column rank, rules out underidentification. This is because now l rows will be fewer than k columns, and since column rank = row rank, we must have deficient column rank.
If $l < k$ we have more equations than unknowns and $E[w_i x_i']$ is no longer invertible. We could throw away extra variables but better instead to use 2SLS, since we want to extract as much exogenous variation from our endogenous variables as possible.

14.2 2SLS

For now assume $E[\varepsilon_i|w_i] = 0$. Then:

$$0 = E[\varepsilon_i|w_i] = E[y_i - x_i'\beta|w_i] = E[y_i|w_i] - E[x_i'|w_i]\beta$$

$$\Rightarrow E[y_i|w_i] = E[x_i'|w_i]\beta$$

Suppose we also know

$$E[x_i'|w_i] = w_i'\pi$$

Then we have:

$$E[y_i|w_i] = (w_i'\pi)\beta$$

This suggest the following procedure:

Definition 14.2.1

2SLS

Stage 1:

- Regress $X_{n\times k}$ on $W_{n\times l}$ to get $\hat{\pi} = (W'W)^{-1}W'X$
- Use the results to form $\hat{X} = W\hat{\pi}$

Note: $\hat{X} = W\hat{\pi} = W(W'W)^{-1}W'X = P_W X$
For the exogenous variables columns in $\hat{X}$ this will correspond exactly to the original values, but for the endogenous variables columns, they will be formed as a linear combination of both the relevant instruments and exogenous variables.

Stage 2:

Regress $Y_{n\times 1}$ on $\hat{X}_{n\times k}$ to find:

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y = (X'P_W'P_W X)^{-1}X'P_W'Y$$
$$= (X'P_W X)^{-1}X'P_W Y$$

Consider the following **IV assumptions** for the model $y_i = x_i'\beta + \varepsilon_i$:

- (IV0) y_i, x_i, w_i is an i.i.d sequence
- (IV1) $E[w_i w_i'] < \infty$ non-singular; $E[w_i x_i']$ has full column rank (relevance)
- (IV2) $E[\varepsilon_i | w_i] = 0$ ($\Rightarrow$) (IV2') $E(w_i\varepsilon_i) = 0$ (exogeneity)
- (IV3) $E[\varepsilon_i^2 | w_i] = \sigma^2$ (homoskedasticity) or (IV3') $V = Var(w_i\varepsilon_i)$ is finite non singular (Under IV(3): $V = E[w_i w_i'\varepsilon_i^2] - 0 = E[E[w_i w_i'\varepsilon_i^2 | w_i]] = \sigma^2 E[w_i w_i']$)

Theorem 14.2.1. 2SLS consistency

Under IV(0) IV(1) IV(2')

$$\hat{\beta}_{2SLS} \xrightarrow{p} \beta$$

Proof.

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}(\hat{X}Y) = (X'P_W X)^{-1}X'P_W Y$$
$$= \beta + (X'P_W X)^{-1}X'P_W\varepsilon$$
$$\hat{\beta}_{2SLS} - \beta = [X'W(W'W)^{-1}W'X]^{-1}X'W(W'W)^{-1}W'\varepsilon$$
$$= \left[\frac{1}{n}\sum x_i w_i'(\frac{1}{n}\sum w_i w_i')^{-1}\frac{1}{n}\sum w_i x_i'\right]^{-1}\frac{1}{n}\sum x_i w_i'(\frac{1}{n}\sum w_i w_i')^{-1}(\frac{1}{n}\sum w_i\varepsilon_i)$$
$$\xrightarrow{p} [E(x_i w_i')E(w_i w_i')^{-1}E(w_i x_i')]^{-1}E(x_i w_i')E(w_i w_i')^{-1}E(w_i\varepsilon_i)$$

By IV(2'), $E(w_i\varepsilon_i) = 0$ and by IV(1) $E(w_i w_i')$ is non-singular to a finite constant matrix (also assume $E(x_i w_i') < \infty$). Thus

$$\hat{\beta}_{2SLS} - \beta \xrightarrow{p} 0$$

□

In general $dimW \neq dimX$. In the case where they do: $\hat{\beta}_{2SLS} \equiv \hat{\beta}_{IV}$, since $W'X$ now invertible. The 2SLS procedure ensures that $dim\hat{X} = dimX$, so that $\hat{\beta}_{2SLS} \equiv \hat{\beta}_{IV}$, using $\hat{X}$ as an instrument. Explicitly: $(X'P_WX)^{-1}X'P_WY = (X'P'_WX)^{-1}X'P'_WY = (\hat{X}'X)^{-1}(\hat{X}'Y) = \hat{\beta}_{IV}$

Theorem 14.2.2. 2SLS asymptotic distribution

Under IV0-1-2'-3':

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, (D'C^{-1}D)^{-1}D'C^{-1}VC^{-1}D(D'C^{-1}D)^{-1})$$

where $V = Var(w_i\varepsilon_i)$, $C = E[w_iw_i']$ and $D = E[w_ix_i']$.

Proof.

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) =$$

$$\left[\frac{1}{n}\sum x_iw_i'(\frac{1}{n}\sum w_iw_i')^{-1}\frac{1}{n}\sum w_ix_i'\right]^{-1}\frac{1}{n}\sum x_iw_i'(\frac{1}{n}\sum w_iw_i')^{-1}(\frac{1}{\sqrt{n}}\sum w_i\varepsilon_i)$$

By Lindeberg-Levy CLT:

$$\frac{1}{\sqrt{n}}\sum w_i\varepsilon_i \xrightarrow{d} N(0, V)$$

By Slutsky's theorem:

$$\xrightarrow{d} [D'C^{-1}D]^{-1}D'C^{-1}N(0, V)$$

$$= N(0, (D'C^{-1}D)^{-1}D'C^{-1}VC^{-1}D(D'C^{-1}D)^{-1})$$

Under (IV3) (homoskedasticity):

$$V = Var(w_i\varepsilon_i) = E[w_iw_i'\varepsilon_i^2] - 0 = \sigma^2E[w_iw_i'] = \sigma^2C$$

Thus much of the asymptotic variance cancels, leaving

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \sigma^2(D'C^{-1}D)^{-1})$$

□

Note:-

In general for two full column rank conformable matrices A, B:
We have AB full column rank.

Proof: Suppose AB not full column rank.
Then $\exists\ x \neq 0$ such that $ABx = 0$ (by the rank-nullity theorem).
$\Rightarrow Bx \neq 0$ as B full rank implies its null space is only $\{0\}$.
$\Rightarrow A(Bx) \neq 0$ as A also full rank with only trivial null space.
Contradiction

We apply this proof to argue $D'C^{-1}D$ is full column rank, and hence invertible.

14.2.1 Linear Hypothesis Testing with β_{2SLS}

We can estimate the asymptotic variance of $\sqrt{n}(\hat{\beta}_{2SLS} - \beta)$ by:

$$\hat{V} = \hat{\sigma}^2(\tfrac{1}{n}\hat{X}'\hat{X})^{-1}$$

where $\hat{\sigma}^2 = \frac{1}{n}\hat{\varepsilon}'\hat{\varepsilon}$ and $\hat{\varepsilon} = Y - \hat{X}'\hat{\beta}_{2SLS}$

Homoskedasticity or robust variance estimates of $\hat{\beta}_{2SLS}$ can be used to form F-statistics for testing linear hypotheses in the usual way. Asymptotically, such F-statistics would be distributed as $\chi^2(p)/p$, where p is the number of restrictions. However, finite sample distribution of the F-statistics would not be $F(p, n-k)$ even if ε_i is normally distributed.

Asymptotically the Wald statistic for testing $H_0 : R\beta = r$ is:

$$W = (R\hat{\beta}_{2SLS} - r)'[R\hat{V}_{2SLS}R']^{-1}(R\hat{\beta}_{2SLS} - r) \xrightarrow{d} \chi^2(p)$$

where p is the number of restrictions.

Note:-

Under (IV3') (hetereoskedasticity) we can use White's estimate as in earlier discussions:

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

14.3 Control function approach

This is an alternative approach to 2SLS, which is useful when we have multiple endogenous variables.

Consider again the model:

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

where x_{1i} is exogenous and x_{2i} is endogenous.

Instead of extracting the *exogenous* part $w_i'\pi$ of x_i to use in the second stage, we could instead extract the **endogenous part** of x_i (the control function) and add it to the regression as an additional regressor.

Theorem 14.3.1. The two approaches are equivalent in a linear model (but not in non-linear) $\hat{\beta}_{CF} \equiv \hat{\beta}_{2SLS}$

Proof. The exogenous part $w_i'\pi$ of x_i is simply the best linear predictor of x_i given w_i.

The first stage regression:

$$x_i' = w_i'\pi + u_i', \text{ where } \pi \text{ is } l \times k$$

is called a *reduced form* regression, because it does not have any structural interpretation. We just want to predict x_i by a linear function of w_i in the best possible way (thus exogeneity is not required). Recall w_i contains components of both included exogenous variables x_{1i} and excluded exogenous variables z_i.

Thus we partition the reduced form equations into:

$$x_{1i}' = x_{1i}'\pi_{11} + z_i'\pi_{12} + u_{1i}'$$

$$x_{2i}' = x_{1i}'\pi_{21} + z_i'\pi_{22} + u_{2i}'$$

where π_{ij} is a $k_j \times k_i$ matrix.

Of course the BLP of x_{1i} given x_{1i} and z_{1i} is just x_{1i} so the first of the above equations is trivial $x_{1i}' = x_{1i}'$ For the second equation we drop the first subscript and rewrite as:

$$x_{2i}' = x_{1i}'\pi_1 + z_i'\pi_2 + u_i'$$

In 2SLS this regression would be estimated, obtain $\hat{x}'_{2i}$, form $\hat{x}_i$ by combining x_{1i}, with $\hat{x}_{2i}$ and proceeding to second stage.

But note x_{2i} can only be endogenous if $E(u_i\varepsilon_i) \neq 0$, that is, the error of the first stage u_i is correlated with the structural error ε_i. Alternatively, note $x_2 i$ can only be endogenous if $E(\vec{u}_i\varepsilon_i) \neq \vec{0}$*

That is, the error of the first stage regression, u_i is correlated with the structural error ε_i. The error u_i has *soaked up* the endogeneity in x_{2i} thus adding it to the structural equation would control for the endogeneity and so get consistent estimates for the other structural parameters.

Consdier the BLP of ε_i given u_i:

$$\varepsilon_i = u'_i\alpha + e_i$$

By definition the error of the BLP is uncorrelated to the dependent ε_i, else it would have been taken into account in the regression.

Substituting this into the strucutural equation, we obtain

$$y_i = x'_{1i}\beta_1 + x'_{2i}\beta_2 + u'_i\alpha + e_i$$

where:

$$E(u_i e_i) = 0$$

$$E(x_{1i}e_i) = E(x_{1i}(\varepsilon_i - u'_i\alpha)) = 0$$

$$E(x_{2i}e_i) = E((\pi'_1 x_{1i} + \pi'_2 z_i + u_i)e_i) = E(\pi'_2 z_i e_i) = \pi_2 E(z_i(\varepsilon - u'_i\alpha)) = 0$$

Thus OLS2' satisfied and the OLS estimates of β_1, β_2, and α should be consistent. But we do not observe u_i so it must first be estimated from the first stage regression before insertion.

Let $\hat{U}$ be the matrix with rows $\hat{u}'_i$. Then by the partitioned regression formula (FW - theorem):

$$\hat{\beta}_{CF} \equiv (X' M_{\hat{U}} X)^{-1} X' M_{\hat{U}} X$$

But $\hat{U} = M_W X_2$ so that:

$$M_{\hat{U}} = I - \hat{U}(\hat{U}'\hat{U})^{-1}\hat{U}' = I - M_W X_2 (X'_2 M_W X_2)^{-1} X'_2 M_W$$

Since X_1 is a part of W, $M_W X_1 = 0$, and

$$M_{\hat{U}} X_1 = X_1 = P_W X_1$$

Further

$$M_{\hat{U}} X_2 = X_2 - M_W X_2 (X'_2 M_W X_2)^{-1} X'_2 M_W X_2 = P_W X_2$$

Therefore

$$M_{\hat{U}} X = P_W X$$

and so

$$\hat{\beta}_{CF} \equiv (X' M_{\hat{U}} X)^{-1} X' M_{\hat{U}} Y = (X' P_W X)^{-1} X' P_W Y = \hat{\beta}_{2SLS}$$

*$E(x_{2i}\varepsilon_i) \neq \vec{0} \Rightarrow E[(w'_i\pi + u'_i)'\varepsilon] \neq 0 \Rightarrow \pi \cancel{E[w_i\varepsilon]} + E[u_i\varepsilon_i] \neq 0$

□

14.4 Endogeneity and Overidentification test

Endogeneity test: If x_{2i} is not endogenous, then OLS is efficient (BLUE) and 2SLS is not.

Test

$$H_0 : E(x_{2i}\varepsilon_i) = 0 \text{ against } H_1 : E(x_{2i}\varepsilon_i) \neq 0$$

Recall the CF regression:

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + u_i'\alpha + e_i$$

where

$$\alpha = E(u_i u_i')^{-1} E(u_i \varepsilon_i)(\text{ the coefficient of BLP} for \varepsilon_i given u_i)$$

We have $E(x_{2i}\varepsilon_i \neq 0)$ if and only if $E(u_i\varepsilon_i) \neq 0$. Therefore hypothesis test equivalent to:

$$H_0 : \alpha = 0 \text{ against } H_1 : \alpha \neq 0$$

Therefore a natural test would be the Wald statistic for testing linear restrictions $\alpha = 0$ in the control function regression, with u_i replaced with $\hat{u}_i$. It turns out this replacement des not affect the asymptotic distribution of the test stastistic under the null, and remains $\chi^2(k_2)$ where k_2 is the $dim(\alpha) = dim(x_{2i})$ This follows from a general result on the asymptotic distribution of the OLS estimates of regression coefficients with 'generated' regressions (i.e. the hats consistently estimating the true) H(12-26,12-27) In stata this occurs after estat endoggy WUFF WUFF after ivregress. Het robust s.e. then reported as 'robust regression F' otherwise if default daniel homoskedasticity then reported as 'Wu-Hausman F'

Overidentification test: With $l > k$ (instruments ¿ endoggy regressors) we can test the hypothesis that instruments are exogenous, that is

$$H_0 : E(w_i\varepsilon_i) = 0$$

Let us assume the homoskedasticity, so that $E(\varepsilon_i^2|w_i) = \sigma^2$. Then consider a reduced form regression:

$$\varepsilon_i = w_i'\alpha + e_i$$

, where

$$\alpha = (E(w_i w_i'))^{-1} E(w_i\varepsilon_i)$$

We see that $E(w_i\varepsilon_i) \neq 0$ if and only if $\alpha \neq 0$. We cannot regress ε_i on w_i because we do not observe ε_i. But we can try to replace ε_i with $\hat{\varepsilon}_i$, (the residuals from the 2SLS esimtate of β NOTE this is not the same as the second stage residuals).

Sargan proposed to use nR^2 from this regression as the test stat for H_0 vs H_1:

$$S = nR^2 = n\frac{SSE}{SST} = n\frac{\hat{\varepsilon}'W(W'W)^{-1}W'\hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}}$$

Asymptotic Distribution of S: Note S is invariant wrt transformations $W \to W \times A$ where A is any invertible matrix. Therefore wlog we assume W rotated and scaled so that $W(w_i w_i') = I_l$ As $n \to \infty$:

$$\frac{1}{\sqrt{n}}W'\varepsilon = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i\varepsilon_i \xrightarrow{d} N(0, Var(w_i\varepsilon_i)) = N(0, \sigma^2 I_l) = \sigma N(0, I_l)$$

$$\frac{1}{n}W'W \xrightarrow{p} E(w_i w_i')^{-1} = I_l$$

and $\frac{1}{n}W'X \xrightarrow{p} E(w_i x_i') = Q$ where Q is some full column rank matrix. On the other hand:

$$\frac{1}{\sqrt{n}}W'\hat{\varepsilon} = \frac{1}{\sqrt{n}}W'(Y - X\hat{\beta}_{2SLS}) = \frac{1}{\sqrt{n}}W'(Y - X(X'P_W X)^{-1}X'P_W Y)$$

$$= \frac{1}{\sqrt{n}} W'(\varepsilon + X(X'P_W X)^{-1} X'P_W \varepsilon)$$

$$= (I - W'X(X'P_W X)^{-1} X'P_W) \frac{1}{\sqrt{n}} W'\varepsilon$$

$$\xrightarrow{d} (I - Q(Q'Q)^{-1}Q')\sigma N(0, I_l)$$

Therefore,

$$\hat{\varepsilon}'W(W'W)^{-1}W'\hat{\varepsilon} = \frac{1}{\sqrt{n}}\hat{\varepsilon}'W(\frac{1}{n}W'W)^{-1}\frac{1}{\sqrt{n}}W'\hat{\varepsilon}$$

$$\xrightarrow{d} \sigma^2 N'(I - Q(Q'Q)^{-1}Q')N$$

Lemma 14.4.1. $N'(I - Q(Q'Q)^{-1}Q')N \sim \chi^2(l-k)$

Proof. We have $Q'Q = I_k$ and $Q : l \times k$ where $l > k$ We define Q_c as the $l \times (l-k)$ orthonormal complement matrix such that $[Q\, Q_c]$ together form an $l \times l$ complete orthogonal matrix. Thus $[Q\, Q_c][Q\, Q_c]' = I_l$

$$\Rightarrow QQ' + Q_c Q_c' = I_l$$

$$\Rightarrow Q_c Q_c' = I_l - QQ'$$

Thus

$$N'(I - Q(Q'Q)^{-1}Q')N = N'Q_c Q_c' N$$

$$= (Q_c'N)'(Q_c'N)$$

But $Q_c'N \sim N(0, Q_c'I_l Q_c) = N(0, I_{l-k})$ Thus

$$(Q_c'N)'(Q_c'N) = \sum_{i=1}^{l-k}(z_i)^2 \sim \chi^2(l-k)$$

□

Thus:

$$\hat{\varepsilon}'W(W'W)^{-1}W'\hat{\varepsilon} \xrightarrow{d} \sigma^2\chi^2(l-k)$$

Finally, $\frac{\hat{\varepsilon}'\hat{\varepsilon}}{n} \xrightarrow{p} \sigma^2$ (sim to lec 8 proof) Therefore:

$$S = n\frac{\hat{\varepsilon}'W(W'W)^{-1}W'\hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}} \xrightarrow{d} \chi^2(l-k)$$

We reject the null of the instrument exogeneity when s is larger than a critical value of $\chi^2(l-k)$

Note:-

The test cannot be performed in the just-identified situation ($l = k$). Then $W'X$ has full rank and so is thus invertible.

$$\frac{1}{\sqrt{n}}W'\hat{\varepsilon} = (I - W'X(X'P_W X)^{-1}X'W(W'W)^{-1})\frac{1}{\sqrt{n}}W'\varepsilon$$

$$= (I - W'X(X'W(W'W)^{-1}W'X)^{-1}X'W(W'W)^{-1})\frac{1}{\sqrt{n}}W'\varepsilon$$

$$= (I - W'X(W'X)^{-1}W'W(X'W)^{-1}X'W(W'W)^{-1})\frac{1}{\sqrt{n}}W'\varepsilon$$

$$(I - I)\frac{1}{\sqrt{n}}W'\varepsilon = 0$$

14.5 Appendix

14.5.1 Chi-squared asymptotic result

Lemma 14.5.1. For $\vec{z} \sim N(0, V)$ We have

$$z'V^{-1}z \xrightarrow{d} \chi^2(p)$$

where p is the number of elements in z.

Proof. As V symmetric we can write its spectral decomposition:

$$V = Q\Lambda Q' = Q\Lambda^{1/2}\Lambda^{1/2}Q'$$

where Q orthogonal and Λ diagonal with eigenvalues $\lambda_1, ..., \lambda_p$.

$$\therefore z'V^{-1}z = z'(Q\Lambda^{1/2}\Lambda^{1/2}Q')^{-1}z$$

$$= ((\Lambda^{1/2}Q)^{-1}z)'((\Lambda^{1/2}Q)^{-1}z)$$

But

$$(\Lambda^{1/2}Q)^{-1}z \sim N(0, (\Lambda^{1/2}Q)^{-1}V(Q'\Lambda^{1/2})^{-1})$$

$$= N(0, (\Lambda^{1/2}Q)^{-1}Q\Lambda^{1/2}\Lambda^{1/2}Q'(Q'\Lambda^{1/2})^{-1})$$

$$= N(0, I_p)$$

Therefore $(\Lambda^{1/2}Q)^{-1}z$ is a vector of p independent standard normals.
Therefore $((\Lambda^{1/2}Q)^{-1}z)'((\Lambda^{1/2}Q)^{-1}z)$ is the sum of p independent standard normals squared, which is $\chi^2(p)$. □

14.5.2 Limited Info Maximum Likelihood

- no finite sample moments the same as 2SLS (so will have outliers)
- but better than 2sls with weak instruments

Recall the same linear regression model:

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

$$x_i' = w_i'\pi + u_i'$$

$$\Rightarrow y_i = w_i'\pi\beta + u_i'\beta + \varepsilon_i$$

Let $(y_i, x_i) = Y_i'$

$$\Rightarrow Y_i' = w_i'(\pi\beta, \pi) + (u_i'\beta + \varepsilon_i, u_i')$$

Transposing

$$Y_i = \underbrace{\begin{pmatrix} \beta'\pi' \\ \pi' \end{pmatrix}}_{\Gamma(\beta,\pi)} w_i + \underbrace{\begin{pmatrix} \beta'u_i + \varepsilon_i \\ u_i \end{pmatrix}}_{e_i}$$

$$\Rightarrow Y_i = \Gamma(\beta, \pi)w_i + e_i$$

Assume:

$$e_i | w_i \sim N(0, \Omega)$$

We can then write likelihood function, and maximise wrt parameters to find $\hat{\beta}_{ML} = \hat{\beta}_{LIML}, \hat{\pi}_{ML}$ and $\hat{\Omega}_{ML}$

15 Irrelevant and Weak Instruments

15.1 Irrelevant Instruments

Occurs when $E(w_i x_i')$ is not of full column rank, violating IV1. In this case, parameter β is not identified. Consider:

$$y_i = x_i\beta + \varepsilon_i$$
$$x_i - w_i\gamma + e_i$$

with one-dimernsional endogenous variable x_i, and instrument w_i. Satisfying $E(w_i\varepsilon_i) = 0$, but fails the relevance assumption, so that $\gamma = 0$ and hence $E(w_i x_i) = 0$.
The system of equations (moment conditions):

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

is equivalent to

$$\begin{cases} E(w_i(y_i - x_i\beta)) = 0 \\ E(w_i x_i) = 0 \end{cases} \Leftrightarrow \begin{cases} E(w_i y_i) \\ E(w_i x_i) \end{cases}$$

which tells us nothing about β and so it is not identified.
We can still compute IV estimator as unlikely $W'X$ exactly zero in a finite sample:

Proposition 15.1.1. Under non-identifiability:

- $\hat{\beta}_{IV}$ does not converge in probability to a limit. Instead it converges in distribution to a RV. In particular, $\hat{\beta}_{IV}$ is not consistent.
- The limiting distribution of $\hat{\beta}_{IV}$ is not centered at β but instead has its median at $\beta + \rho$, like β_{OLS}.
- $\hat{\beta}_{IV}$ will have very wild fluctuations in finite samples, since the ratio ξ_0/ξ_2 is Cauchy distributed and has no moments, due its fat tails.

Proof. For simplicity we assume homoskedasticity and suppose:

$$E(e_i|w_i) = E(\varepsilon_i|w_i) = 0$$
$$Var(e_i|w_i) = Var(\varepsilon_i|w_i) = 1, Cov(e_i, \varepsilon_i|w_i) = \rho \neq 0$$
$$Ew_i = 0, Ew_i^2 = 1$$

By CLT:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\begin{pmatrix} w_i\varepsilon_i \\ w_i e_i \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \xi_1 \\ \xi_2 \end{pmatrix} \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$$

Note: $\xi_0 = \xi_1 - \rho\xi_2$ is a normal random variable, independent from ξ_2

$$E(\xi_0|\xi_2) = E(\xi_1 - \rho\xi_2|\xi_2) = E(\xi_1|\xi_2) - \rho\xi_2 = 0$$

Then (using $\gamma = 0 \Rightarrow x_i = e_i$)

$$\hat{\beta}_{OLS} - \beta = \frac{\frac{1}{n}\sum x_i\varepsilon_i}{\frac{1}{n}\sum x_i^2} = \frac{\frac{1}{n}\sum e_i\varepsilon_i}{\frac{1}{n}\sum e_i^2} \xrightarrow{p} \rho \neq 0$$

$$\hat{\beta}_{IV} - \beta = \frac{\frac{1}{\sqrt{n}}\sum w_i\varepsilon_i}{\frac{1}{\sqrt{n}}\sum w_i x_i} = \frac{\frac{1}{\sqrt{n}}\sum w_i\varepsilon_i}{\frac{1}{\sqrt{n}}\sum w_i e_i} \xrightarrow{d} \frac{\xi_1}{\xi_2} = \rho + \frac{\xi_0}{\xi_2}$$

□

Proposition 15.1.2. Under non-identifiability, the β t-statistics will diverge, such that we may conclude statistically significant estimates when they are in fact useless. (we prove this explicitly below in the case when $\rho \to 1$, i.e. lots of endogeneity)

$$|t| \xrightarrow{p} \infty$$

Proof.

$$\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - x_i\hat{\beta}_{IV})^2 = \frac{1}{n}\sum(\varepsilon_i - x_i(\hat{\beta}_{IV} - \beta))^2$$

$$= \frac{1}{n}\sum(\varepsilon_i - e_i(\hat{\beta}_{IV} - \beta))^2$$

$$\frac{1}{n}\sum\varepsilon_i^2 - 2(\hat{\beta}_{IV} - \beta)\frac{1}{n}\sum\varepsilon_i e_i + (\hat{\beta}_{IV} - \beta)^2\frac{1}{n}\sum e_i^2$$

$$\xrightarrow{d} 1 - 2\rho\left(\rho + \frac{\xi_0}{\xi_2}\right) + \left(\rho + \frac{\xi_0}{\xi_2}\right)^2$$

$$= 1 - \rho^2 + \left(\frac{\xi_0}{\xi_2}\right)^2$$

Therefore, the t-statistic for $H_0 : \beta = \beta_0, H1 : \beta \neq \beta_0$ has the asymptotic distribution:

$$t = \frac{\hat{\beta}_{IV} - \beta_0}{\sqrt{\hat{Var}(\hat{\beta}_{IV})}}; \qquad Var(\hat{\beta}_{IV}) = \sigma^2(D'C^{-1}D)^{-1}$$

Replacing with sample analogues:

$$\hat{Var}(\hat{\beta}_{IV}) = \hat{\sigma}^2(\hat{D}'\hat{C}^{-1}\hat{D})^{-1}$$

In the 1D case:

$$\hat{D} = \frac{1}{\sqrt{n}}\sum w_i x_i$$

$$\hat{C} = \frac{1}{n}\sum w_i^2$$

$$\therefore t = \frac{\hat{\beta}_{IV} - \beta}{\sqrt{\hat{\sigma}^2\frac{1}{n}\sum w_i^2}/\frac{1}{\sqrt{n}}|\sum w_i x_i|} = \frac{\hat{\beta}_{IV} - \beta}{\sqrt{\hat{\sigma}^2\frac{1}{n}\sum w_i^2}/\frac{1}{\sqrt{n}}|\sum w_i e_i|}$$

$$\xrightarrow{d} \frac{\rho + \frac{\xi_0}{\xi_2}}{\sqrt{1 - \rho^2 + \left(\frac{\xi_0}{\xi_2}\right)^2}/|\xi_2|}$$

This distribution is non-normal. Note when $\rho \to 1$

$$Var(\xi_0) = Var(\xi_1 - \rho\xi_2) = 1 - 2\rho^2 + \rho^2 \xrightarrow{0}$$

so that

$$\xi_0 \xrightarrow{p} 0 \text{ and } 1 - \rho^2 + (\frac{xi_0}{\xi_2})^2 \xrightarrow{p} 0$$

This implies that

$$|t| \xrightarrow{p} \infty$$

□

15.2 Weak Instruments

When $E(w_i x_i')$ is of full column rank but $E(w_i w_i')^{-1} E(w_i x_i')$ (the coefficient matrix in the first stage regression) is small, the instruments, although relevant, are **weak**.

Proposition 15.2.1. Under weak instruments, $\hat{\beta}_{IV}$ will still not be consistent for β and will again have a non normal distribution.

Proof. Consider the same 1D model as previous, except instead of $E(w_i x_i) = 0$ We assume it is 'small', modelled by the 'local-to-zero' assumption:

$$E(w_i x_i) = \gamma = \frac{1}{\sqrt{n}}\mu$$

where μ is fixed, which will yield useful asymptotic approximations for $\hat{\beta}_{IV}$. Large μ corresponds to relatively strong instruments, whereas small μ corresponds to almost irrelevant instruments.
For $\hat{\beta}_{OLS}$ as before:

$$\hat{\beta}_{OLS} - \beta = \frac{\frac{1}{n}\sum x_i \varepsilon_i}{\frac{1}{n}\sum x_i^2} = \frac{\frac{1}{n}\sum(\frac{\mu}{\sqrt{n}} w_i + e_i)\varepsilon_i}{\frac{1}{n}\sum(e_i^2 + \frac{2\mu}{\sqrt{n}} w_i e_i + \frac{\mu^2}{n} w_i^2)} \xrightarrow{p} \rho \neq 0$$

For $\hat{\beta}_{IV}$:

$$\hat{\beta}_{IV} - \beta = \frac{\sum w_i \varepsilon_i}{\sum w_i x_i} = \frac{\frac{1}{\sqrt{n}}\sum w_i \varepsilon_i}{\frac{1}{\sqrt{n}}\sum w_i(\frac{\mu}{\sqrt{n}} w_i + e_i)}$$

$$= \frac{\frac{1}{\sqrt{n}}\sum w_i \varepsilon_i}{\mu \frac{1}{n}\sum w_i^2 + \frac{1}{\sqrt{n}}\sum w_i e_i} \xrightarrow{d} \frac{\xi_1}{\mu + \xi_2} = \frac{\xi_0}{\mu + \xi_2} + \rho\frac{\xi_2}{\mu + \xi_2}$$

Thus as in the case of irrelevant instruments, $\hat{\beta}_{IV}$ is not consistent for β and has a non-normal asymptotic distribution. When $\mu \to \infty$, so the instruments become strong, the consistency is restored because $\frac{\xi_1}{\mu + \xi_2} \xrightarrow{p} 0$ as $\mu \to \infty$. □

Proposition 15.2.2. The t-statistic for β will inflate as instruments become weaker and the distribution will become closer to $\chi^2(1)/\|\mu|$.

Proof. First consider $\hat{\sigma}^2$:

$$\hat{\sigma}^2 = \frac{1}{n}\sum(y_i - x_i\hat{\beta}_{IV})^2 = \frac{1}{n}\sum(\varepsilon_i - x_i(\hat{\beta}_{IV} - \beta))^2$$

$$\frac{1}{n}\sum\varepsilon_i^2 - 2(\hat{\beta}_{IV} - \beta)\frac{1}{n}\sum\varepsilon_i x_i + (\hat{\beta}_{IV} - \beta)^2\frac{1}{n}\sum x_i^2$$

But

$$\frac{1}{n}\sum\varepsilon_i x_i = \frac{1}{n}\sum\varepsilon_i\left(\frac{\mu}{\sqrt{n}}w_i + e_i\right) \xrightarrow{p} \rho \text{ and}$$

$$\frac{1}{n}\sum x_i^2 \xrightarrow{p} 1$$

Hence,

$$\hat{\sigma}^2 \xrightarrow{d} 1 - 2\rho\frac{\xi_1}{\mu + \xi_2} + \left(\frac{\xi_1}{\mu + \xi_2}\right)^2$$

$$= 1 - \rho^2 + \left(\rho - \frac{\xi_1}{\mu + \xi_2}\right)^2$$

$$= 1 - \rho^2 + \left(\frac{\rho\mu - \xi_0}{\mu + \xi_2}\right)^2$$

Thus for the t-statistic:

$$t = \frac{\hat{\beta}_{IV} - \beta}{\sqrt{\hat{\sigma}^2\frac{1}{n}\sum w_i^2}/\frac{1}{\sqrt{n}}|\sum w_i x_i|} \xrightarrow{d} \frac{\xi_1/(\mu + \xi_2)}{\sqrt{1 - \rho^2 + (\frac{\rho\mu - \xi_0}{\mu + \xi_2})^2}/|\mu + \xi_2|}$$

$$= \frac{\xi_1}{sgn(\mu + \xi_2)\sqrt{(1 - \rho^2) + (\frac{\rho\mu - \xi_0}{\mu + \xi_2})^2}}$$

This has a non-normal distribution.
Again we consider the extreme case where $\rho = 1$ (lots of endogeneity in x_i):
$\rho = 1 \Rightarrow \xi_1 = \xi_2 = \xi \sim N(0, 1)$, $\xi_0 = \xi - \rho\xi = 0$

$$t \xrightarrow{d} \frac{\xi(\mu + \xi)}{|\mu|}$$

With $|\mu|$ very large (strong instruments), t is almost normal, but when $|\mu|$ is small, t is almost $\chi^2(1)/|\mu|$.
As $\mu \to 0$, $t \xrightarrow{p} \infty$. Thus there is a multiplicity of possibilities when μ varies. □

15.2.1 Classifying Weak Instruments

Stock and Yogo (2005) classify strength of instruments based on the size distortion of the nominal 5% significance asymptotic t-test, which rejects the null hypothesis iff $|t| > 1.96$.

No distortion implies:

$$Pr(|t| > 1.96) \to 0.05$$

But with $\mu < \infty$, such convergence will not take place:

$$Pr(|t| > 1.96) = Pr\left(|\frac{\xi(\mu + \xi)}{|\mu|}| > 1.96\right) \nrightarrow 0.05$$

Thus the actual (as opposed to nominal(i.e. intended)) size of the test will not be 5%.

Stock and Yogo suggested that a 'tolerable' actual size should be perhaps not larger than 15%.

Let τ^2 be such that, whenever $\mu^2 \geq \tau^2$ the actual size of the t-test is below 15%. τ^2 is found through simulating the $\frac{\xi(\mu+\xi)}{|\mu|}$ distribution.

Then proposed to use the F-stat from first stage regression (normally used to test hypothesis $\mu = 0$) to test hypothesis that $\mu^2 \leq \tau^2$.

By simulation they found the appropriate critical value is **approx 10**, though this is dependent on choosing 15% as the benchmark distortion and number of instruments used.
But for not too many instruemnts critical value remains around 10 so this is used in empirical literature often to determine if instruments weak.

16 Generalised Method of Moments.

16.1 GMM setup

The basic principle of the method of moments is to choose the parameter estimate so that the corresponding sample moments are equal to the corresponding population moments. Moment equation models take the following form. Let $\mathbf{g}_i(\boldsymbol{\beta})$ be a known $\ell \times 1$ function of the i^{th} observation, and a $k \times 1$ parameter $\boldsymbol{\beta}$. Denote the true parameter value in population as $\boldsymbol{\beta}_0$. A moment equation model is summarised by the moment equations

$$\mathbb{E}[\mathbf{g}_i(\boldsymbol{\beta}_0)] = 0.$$

Example. **Mean:** $g_i(\mu) = y_i - \mu$.
Instrumental Variables: $\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{w}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$. (Exogeneity of w_i)
Linear Regression: $\mathbf{g}_i(\boldsymbol{\beta}) = \mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})$. (uncorrelatedness of x_i and u_i)

When $\ell = k$, the moment equations are just identified (k equations for k unknowns), and we solve the moment equations for $\hat{\boldsymbol{\beta}}$

$$\bar{g_n} = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}) = 0.$$

When $\ell < k$ the system is under-identified we typically have no solution to the system, and cannot hope to solve for $\hat{\boldsymbol{\beta}}$.
When $\ell > k$ the system is over-identified there are an infinite number of solutions to the system (more variables than unknowns), each one giving us a different MOM estimate of $\boldsymbol{\beta}$. 2SLS solved this problem by using the k dimensional vector $\hat{\pi} w_i$ obtained from the first stage regression. Alternatively, GMM aims to minimise the distance between the sample and population moments. That is,

Definition 16.1.1: GMM estimator

$$\hat{\beta}_{GMM} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta})\right)' W_n \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\boldsymbol{\beta})\right) = \arg\min_{\boldsymbol{\beta}} \bar{g_n}' W_n \bar{g_n}.$$

where W_n is a positive-definite $l \times l$ weighting matrix.

The object we are minimising is called the GMM criterion function, further denoted as $J_n(\beta)$. If the criterion function is unweighted: $W_n = I_\ell$, then $J_n(\beta) = \bar{g_n(\beta)}' \bar{g_n(\beta)} = ||\bar{g_n(\beta)}||^2$, the square of the Euclidean length. Since we restrict attention to positive definite weight matrices W, the criterion J_n is always non-negative.

Note:-

In the just identified case, the GMM estimator is the same as the MOM estimator, regardless of the choice of W_n. $\hat{\beta}_{MOM}$ solves $\bar{g}_n(\beta_{MOM}) = 0$, hence $J_n(\beta_{MOM}) = 0$. Since $J_n(\beta) \geq 0$, $\hat{\beta}_{MOM}$ is also the GMM estimator.

GMM is known as an extremum estimator, since it is obtained by minimising a criterion function. Another example of an extremum estimator is MLE.

Example (Overidentified IV). Given the moment conditions $\mathbb{E}[z_i(y_i - x_i'\beta)] = 0$ where z_i is $\ell \times 1$ and x_i is $k \times 1$, we can construct the GMM criterion function as follows:

$$J_n(\beta) = (Z'y - Z'X\beta)'W_n(Z'y - Z'X\beta)$$

The GMM estimator minimises this criterion function, the FOCs are:

$$\begin{aligned} 0 &= \frac{\partial J_n(\beta)}{\partial \beta} \\ &= \frac{\partial}{\partial \beta}\left(y'ZW_nZ'y - y'ZW_nZ'X\beta - \beta'X'ZW_nZ'y + \beta'X'ZW_nZ'X\beta\right) \\ &= -2X'ZW_nZ'y + 2X'ZW_nZ'X\beta \\ \Rightarrow \hat{\beta}_{GMM} &= (X'ZW_nZ'X)^{-1}X'ZW_nZ'y \end{aligned}$$

While the estimator depends on W_n, the dependence is only up to scale. For example, if we multiply W_n by a constant $c > 0$, then $\hat{\beta}_{GMM}$ is unchanged. When W is fixed by the user we call $\hat{\beta}_{GMM}$ the **one-step GMM estimator**. The GMM estimator is similar to the 2SLS estimator, in fact they are equal when $W_n = (Z'Z)^{-1}$.

Theorem 16.1.1. If $W_n = (Z'Z)^{-1}$, then $\hat{\beta}_{GMM} = \hat{\beta}_{2SLS}$. Furthermore if $k = \ell$, then $\hat{\beta}_{GMM} = \hat{\beta}_{IV}$.

Example (CAPM). Consider the case where a representative agent makes decisions about consumption and investment to maximise lifetime discounted expected utility conditional on their information set. Suppose the agent has CRRA utility

$$\mathbb{E}\left(\sum_{i=1}^{\infty} \delta_0^i U(c_{t+i}|\mathcal{I}_t)\right), \quad U(c_t) := \frac{c_t^{\gamma_0} - 1}{\gamma_0}$$

The budget constraint is

$$c_t + p_t q_t = r_t q_{t-1} + w_t$$

where p_t is the price of a financial asset, q_t is the amount of asset owned at time t, r_t is the gross return to the asset, and w_t is the wage. The FOC (euler equation) for this problem is:

$$p_t c_t^{\gamma_0 - 1} = \mathbb{E}_t\left(\delta_0 r_{t+1} c_{t+1}^{\gamma_0 - 1} | \mathcal{I}_t\right)$$

We thus have the conditional moment condition:

$$\mathbb{E}_t\left(\delta_0 \frac{r_{t+1}}{p_t} \frac{c_{t+1}^{\gamma_0 - 1}}{c_t^{\gamma_0 - 1}} | \mathcal{I}_t\right) = 0$$

This implies an infinite number of unconditional moment conditions, we select the following (arbitrarily):

$$\mathbb{E}_t\left(z_t \left[\delta_0 \frac{r_{t+1}}{p_t} \frac{c_{t+1}^{\gamma_0 - 1}}{c_t^{\gamma_0 - 1}}\right]\right) = 0$$

Where z is any vector of variables known by time t (EG consumption growth, returns in recent periods etc.) This is $dim(z_t)$ equations for 2 unknowns γ_0 and δ_0. We can estimate this model using GMM.

Note:-

A conditional moment condition is much stronger than an unconditional condition, the conditional condition implies an infinite set of unconditional moments:

$$\begin{aligned}\mathbb{E}[f(z_i)g(x_i,\beta_0)] &= \mathbb{E}[\mathbb{E}[f(z_i)g(x_i,\beta_0)|z_i]] \\ &= \mathbb{E}[f(z_i)\mathbb{E}[g(x_i,\beta_0)|z_i]] \\ &= 0\end{aligned}$$

above we arbitrarily selected $f(z_i) = z_i$, but any function of any subset of variables in z_i will do. See here for a brief discussion of the optimal choice, however this is beyond the scope of this course.

16.2 Asymptotic properties of GMM

(GMM0) $\frac{1}{n}\sum_{i=1}^n g_i(\beta) \xrightarrow{p} \mathbb{E}g_i(\beta)$ and $\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta'}g_i(\beta) \xrightarrow{p} \mathbb{E}\frac{\partial}{\partial\beta'}g_i(\beta)$, uniformly over β from a compact subset B of $\mathbb{R}^k$.

(GMM1) $\mathbb{E}\frac{\partial}{\partial\beta'}g_i(\beta)$ has rank k (i.e. full column rank) in a neighbourhood of β_0. Furthermore $W_n \xrightarrow{p} W$, a positive definite matrix.

(GMM2) $\mathbb{E}g_i(\beta_0) = 0$ and β_0 is the only root of equation $\mathbb{E}g_i(\beta) = 0$ for $\beta \in B$.

(GMM3) $\frac{1}{n}\sum_{i=1}^n g_i(\beta)g_i(\beta)' \xrightarrow{p} V(\beta)$ uniformly $\beta \in B$ and $\frac{1}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^n g_i(\beta_0) \xrightarrow{d} N(0, V(\beta_0))$ with $V(\beta_0) > 0$.

(GMM4) $\mathbb{E}g_i(\beta)$ and $g_i(\beta)$ are continuously differentiable, and $\frac{\partial}{\partial\beta'}\mathbb{E}g_i(\beta) = \mathbb{E}\frac{\partial}{\partial\beta'}g_i(\beta)$.

GMM0 does not require iid-ness of the data, it just says that an LLN should hold. As discussed with serial correlation, there exist LLNs for dependent data.

Theorem 16.2.1. Under (GMM0)-(GMM4)

1. $\hat{\beta}_{GMM} \xrightarrow{p} \beta_0$
2. $\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \xrightarrow{d} N(0, (D_0'WD_0)^{-1}D_0'WV_0WD_0(D_0'WD_0)^{-1})$ where $D_0 = \mathbb{E}\frac{\partial}{\partial\beta'}g_i(\beta_0)$ and $V_0 = V(\beta_0)$ is the variance of the asymptotic distribution of .

Proof. 1. By GMM0 and GMM1 $W_n \to W > 0$ and $\frac{1}{n}\sum_{i=1}^n g_i(\beta) \to \mathbb{E}g_i(\beta)$ in probability, then

$$\begin{aligned}J_n(\beta) &= \left(\frac{1}{n}\sum_{i=1}^n g_i(\boldsymbol{\beta})\right)' W_n \left(\sum_{i=1}^n g_i(\boldsymbol{\beta})\right) \\ &\xrightarrow{p} \mathbb{E}g_i(\beta)'W\mathbb{E}g_i(\beta) \geq 0\end{aligned}$$

Using GMM2, we have that β_0 is the only root of $\mathbb{E}g_i(\beta) = 0$, hence β_0 is the minimiser of $J_n(\beta)$, and $\hat{\beta}_{GMM} \xrightarrow{p} \beta_0$.

2. Given that $\hat{\beta}_{GMM} \xrightarrow{p} \beta_0$, and that $\frac{\partial}{\partial\beta}g_i$ is differentiable, we can compute the taylor expansion as follows:

$$\frac{\partial}{\partial\beta}J_n(\hat{\beta}_{GMM}) \approx \frac{\partial}{\partial\beta}J_n(\beta_0) + \frac{\partial^2}{\partial\beta\partial\beta'}J_n(\hat{\beta}_{GMM})(\hat{\beta}_{GMM} - \beta_0)$$

Since the LHS $= 0$ by definition of $\hat{\beta}_{GMM}$, we have that

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \approx -\left(\frac{\partial^2}{\partial\beta\partial\beta'}J_n(\hat{\beta}_{GMM})\right)^{-1}\left(\sqrt{n}\frac{\partial}{\partial\beta}J_n(\beta_0)\right)$$

Consider the first term in the above expression, we can write it as follows:

$$\begin{aligned}
\frac{\partial^2}{\partial\beta\partial\beta'}J_n(\hat{\beta}_{GMM}) &= \frac{\partial^2}{\partial\beta\partial\beta'}\left(\frac{1}{n}\sum_{i=1}^{n}g_i(\hat{\beta}_{GMM})\right)' W_n \left(\frac{1}{n}\sum_{i=1}^{n}g_i(\hat{\beta}_{GMM})\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\beta\partial\beta'}g_i(\hat{\beta}_{GMM})'W_n g_i(\hat{\beta}_{GMM}) + \frac{1}{n}\sum_{i=1}^{n}g_i(\hat{\beta}_{GMM})'W_n\frac{\partial^2}{\partial\beta\partial\beta'}g_i(\hat{\beta}_{GMM}) \\
&+ 2\sum_{j,r=1}^{l}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2}{\partial\beta\partial\beta'}g_{ij}(\beta_0)\right)W_{n,jr}\left(\frac{1}{n}\sum_{i=1}^{n}g_{ir}(\beta_0)\right) \\
&= 2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta'}g_i(\hat{\beta}_{GMM})\right)' W_n \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta'}g_i(\hat{\beta}_{GMM})\right) + o_p(1) \\
&\xrightarrow{d} 2\left(\mathbb{E}\frac{\partial}{\partial\beta'}g_i(\beta_0)\right)' W \left(\mathbb{E}\frac{\partial}{\partial\beta'}g_i(\beta_0)\right) \\
&:= 2D_0'WD_0
\end{aligned}$$

g_{ij} and g_{ir} are the j^{th} and r^{th} elements of g_i respectively, and $W_{n,jr}$ is the element in the j-th row and r-th column of W_n. The cross term is $o_p(1)$ since the final term converges to $\mathbb{E}g_{ir}(\beta_0) = 0$ by GMM2.

Consider the final term in the above expression, we can write it as follows:

$$\begin{aligned}
\sqrt{n}\frac{\partial}{\partial\beta}J_n(\beta_0) &= \sqrt{n}\frac{\partial}{\partial\beta}\left(\frac{1}{n}\sum_{i=1}^{n}g_i(\beta_0)\right)' W_n \left(\frac{1}{n}\sum_{i=1}^{n}g_i(\beta_0)\right) \\
&= 2\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\beta'}g_i(\beta_0)\right)' W_n \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}g_i(\beta_0)\right) \\
&\xrightarrow{d} 2\left(\mathbb{E}\frac{\partial}{\partial\beta'}g_i(\beta_0)\right)' W \left(N(0, \mathbb{E}[g_i(\beta_0)g_i(\beta_0)'])\right) \\
&:= 2D_0'WN(0, V_0)
\end{aligned}$$

Combining these two results, we have that

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) &\xrightarrow{d} (2D_0'WD_0)^{-1}2D_0'WN(0, V_0) \\
&\sim N(0, (D_0'WD_0)^{-1}D_0'WV_0WD_0(D_0'WD_0)^{-1})
\end{aligned}$$

□

16.3 Efficient GMM

The question remains as to how we pick W, and thus W_n. A simple choice is the identity matrix, i.e. $W_n = I_\ell$, or another fixed positive definite matrix. However these choices result in a loss of efficiency.

Theorem 16.3.1. The efficient choice of weighting matrix (that minimises asymptotic vari-

ance) is

$$W = V_0^{-1}$$

where V_0 is the asymptotic variance of $\sqrt{n}\bar{g_n}(\beta_0)$.

Intuitively $\hat{\beta}_{GMM}$ is trying to set each component of the ℓ dimensional vector $\frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}_{GMM})$ as close to zero as possible. However, if some components have particularly large variance (and are thus less informative), then it makes sense to place less weight on them.

Proof. If $W = V_0^{-1}$, then we can simplify the asymptotic variance as follows:

$$\begin{aligned}(D_0'WD_0)^{-1}D_0'WV_0WD_0(D_0'WD_0)^{-1} &= (D_0'V_0^{-1}D_0)^{-1}D_0'V_0^{-1}V_0V_0^{-1}D_0(D_0'V_0^{-1}D_0)^{-1} \\ &= (D_0'V_0^{-1}D_0)^{-1}D_0'V_0^{-1}D_0(D_0'V_0^{-1}D_0)^{-1} \\ &= (D_0'V_0^{-1}D_0)^{-1}\end{aligned}$$

It remains to show that the difference between the inefficient and efficient weighting matrices $(D_0'WD_0)^{-1}D_0'WV_0WD_0(D_0'WD_0)^{-1} - (D_0'V_0^{-1}D_0)^{-1}$ is positive semi-definite. For my first trick, I will pull the following matrix out of my ass: $A = I_\ell - V_0^{-1/2}D_0(D_0'V_0^{-1}D_0)^{-1}D_0'V_0^{-1/2}$. Observe:

$$\begin{aligned}&(D_0'WD_0)^{-1}D_0'WV_0^{1/2}AV_0^{1/2}WD_0(D_0'WD_0)^{-1} \\ &= (D_0'WD_0)^{-1}D_0'WV_0^{1/2}\left(I_\ell - V_0^{-1/2}D_0(D_0'V_0^{-1}D_0)^{-1}D_0'V_0^{-1/2}\right)V_0^{1/2}WD_0(D_0'WD_0)^{-1} \\ &= (D_0'WD_0)^{-1}D_0'WV_0^{1/2}V_0^{1/2}WD_0(D_0'WD_0)^{-1} \\ &\quad - (D_0'WD_0)^{-1}D_0'WV_0^{1/2}V_0^{-1/2}D_0(D_0'V_0^{-1}D_0)^{-1}D_0'V_0^{-1/2}V_0^{1/2}WD_0(D_0'WD_0)^{-1} \\ &= (D_0'WD_0)^{-1}D_0'WV_0WD_0(D_0'WD_0)^{-1} \\ &\quad - \cancel{(D_0'WD_0)^{-1}D_0'WD_0}(D_0'V_0^{-1}D_0)^{-1}\cancel{D_0'WD_0(D_0'WD_0)^{-1}} \\ &= (D_0'WD_0)^{-1}D_0'WV_0WD_0(D_0'WD_0)^{-1} - (D_0'V_0^{-1}D_0)^{-1}\end{aligned}$$

Note that A is the residual maker matrix onto the space spanned by $D_0V^{-1/2}$, is symmetric and idempotent, and thus positive semi-definite. The difference therefore has the form $B'AB = B'A'AB = (AB)'(AB)$ and is clearly positive s.d. . □

Estimation of V_0

We can estimate V_0 and then estimate GMM in a two step procedure. The first step uses a consistent (but not necessarily efficient) estimate of β to construct the optimal weighting matrix W_n. The second step uses this weighting matrix $W_n = \hat{V}_0^{-1}$ to estimate β efficiently. This is known as the two-step GMM estimator.

In the case of i.i.d. $g_i(\beta)$, we can estimate $\hat{\beta}$ consistently with a sub optimal weighting matrix and compute

$$\hat{V}_0 = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta})g_i(\hat{\beta})' \quad \text{or}$$

$$\hat{V}_0 = \frac{1}{n}\sum_{i=1}^{n}\left(g_i(\hat{\beta}) - \bar{g_n}(\hat{\beta})\right)\left(g_i(\hat{\beta}) - \bar{g_n}(\hat{\beta})\right)'$$

If the data are serially correlated, a version of the Newey-West estimator can be used to estimate V_0.

$$\hat{\Gamma}_j = \frac{1}{n}\sum_{i=j+1}^{n} g_i(\hat{\beta})g_{i-j}(\hat{\beta})' \quad \text{and obtain}$$

$$\hat{V}_0 = \frac{1}{n}\sum_{j=0}^{G}\frac{G+1-j}{G+1}\left(\hat{\Gamma}_j + \hat{\Gamma}_j'\right)$$

Example (2SLS under homoskedasticity). Consider linear regression $y_i = x_i'\beta_0 + \varepsilon_i$ with k endogenous variables and $\ell > k$ instruments. The moment conditions $g_i(\beta) = z_i(y_i - x_i'\beta)$ are distributed as:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i\varepsilon_i \to N(0, \sigma^2\mathbb{E}[z_i z_i'])$$

under IV3 (homoskedasticity). Hence $V_0 = \sigma^2\mathbb{E}[z_i z_i'] = \sigma^2 Z'Z$. The efficient weighting matrix is therefore $W_n = (Z'Z)^{-1}$, since the weighting matrix is invariant to scale (so we can drop the σ^2). From the previous example we know that the GMM estimator is given by $\hat{\beta}_{GMM} = (X'ZW_nZ'X)^{-1}X'ZW_nZ'y$, substituting in the efficient weighting matrix we obtain:

$$\hat{\beta}_{GMM} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y = (X'P_ZX)^{-1}X'P_Zy = \hat{\beta}_{2SLS}$$

and thus, 2SLS can be viewed as efficient GMM under homoskedasticity.
However, under heteroskedasticity, 2SLS is inefficient relative to GMM which would instead be based on :

$$W_n = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i^2 z_i z_i'\right)^{-1} \quad \text{or its centred version}$$

$$W_n = \left(\frac{1}{n}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i z_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i z_i\right)\left(\hat{\varepsilon}_i z_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\varepsilon}_i z_i\right)'\right)^{-1}$$

$\hat{\varepsilon}$ can be obtained using any consistent (but inefficient) method, like 2SLS or unweighted GMM. If the model is correctly specified such that $\mathbb{E}z_i\varepsilon_i = 0$, then the centred and un-centred versions of W_n are asymptotically equivalent.

Overidentification test
Previously we studied Sargan's test of over-identifying restrictions in the context of 2SLS under homo. This can be generalised to GMM (allowing for serial correlation and heteroskedasticity). In the overidentified case we have ℓ equations

$$\mathbb{E}(g_i(\beta_0)) = 0$$

that have to be satisfied by $k < \ell$ parameters. Since $\frac{1}{n}\sum_{i=1}^{n} g_i(\beta_0) \xrightarrow{p} \mathbb{E}g_i(\beta_0)$ and $\hat{\beta}_{GMM} \xrightarrow{p} \beta_0$ we can use $\frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}_{GMM})$ to test the hypothesis:

$$H_0 : \mathbb{E}(g_i(\beta_0)) = 0$$

Assuming the efficient weight matrix is being used, the criterion function is

$$J(\hat{\beta}_{GMM}) = n\bar{g}_n'\hat{V}_0^{-1}\bar{g}_n = \left(\frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}_{GMM})\right)'\hat{V}_0^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\beta}_{GMM})\right)$$

Theorem 16.3.2. Under assumptions GMM0-4, as $n \to \infty$,

$$J(\hat{\beta}_{GMM}) \xrightarrow{d} \chi^2(\ell - k)$$

Proof. This was not presented in the course, WORK IN PROGRESS.

For any positive semi-definite matrix A we can decompose it as $A = U\Lambda U' = (U\Lambda^{1/2})(U\Lambda^{1/2})' = BB'$. The covariance matrix V_0 is positive definite, thus Λ is a diagonal matrix with positive entries. The inverse of V_0 is thus also positive definite and can be written as $V_0^{-1} = CC'$ (and $V_0 = C'^{-1}C^{-1}$).
Subbing this into J:

$$\begin{aligned} J(\hat{\beta}_{GMM}) &= n(\bar{g}_n' \hat{V}_0^{-1} \bar{g}_n) \\ &= n(\bar{g}_n' C C' \bar{g}_n) \\ &= n(C'\bar{g}_n(\hat{\beta}))' C'\bar{g}_n(\hat{\beta}) \\ &= n(C'\bar{g}_n(\hat{\beta}))'(C'\hat{V}_0 C)^{-1} C'\bar{g}_n(\hat{\beta}) \end{aligned}$$

Where the last step follows from the fact that $C'\hat{V}_0 C = I_\ell$.
We can write the average of the moment conditions as

$$\bar{g}_n(\beta) = \frac{1}{n} Z'e$$

For my next trick I will pull the following matrix out of my ass:

$$D_n = I_\ell - C' \left(\frac{1}{n} Z'X\right) \left(\left(\frac{1}{n} X'Z\right) \hat{V}_0^{-1} \left(\frac{1}{n} Z'X\right)\right)^{-1} \left(\frac{1}{n} X'Z\right) \hat{V}_0^{-1} C'^{-1}$$

and observe that

$$\begin{aligned} D_n C'\bar{g}_n(\beta) &= I_\ell C'\bar{g}_n - C' \left(\frac{1}{n} Z'X\right) \left(\left(\frac{1}{n} X'Z\right) \hat{V}_0^{-1} \left(\frac{1}{n} Z'X\right)\right)^{-1} \left(\frac{1}{n} X'Z\right) V_0^{-1} C'^{-1} C'\bar{g}_n \\ &= \dots \\ &= C'\bar{g}_n(\hat{\beta}) \end{aligned}$$

D_n is a residual maker matrix onto the space spanned by $C'Z'X$, it converges to the following $D_n \xrightarrow{p} I_\ell - R(R'R)^{-1}R'$ where $R = C'\mathbb{E}(z_i x_i')$.
Further note the asymptotic distribution of $C'\bar{g}_n(\beta)$

$$\begin{aligned} n^{1/2} C'\bar{g}_n(\beta) &= C' \frac{1}{sqrtn} z_i e_i \\ &\xrightarrow{d} C' N(0, V_0) \\ &\sim N(0, C'V_0 C) \\ &\sim N(0, I_\ell) \end{aligned}$$

Thus (with some abuse of notation):

$$\begin{aligned} J(\hat{\beta}_{GMM}) &= n(C'\bar{g}_n(\hat{\beta}))'(C'\hat{V}_0C)^{-1}C'\bar{g}_n(\hat{\beta}) \\ &= (D_n\sqrt{n}C'\bar{g}_n(\beta))'(C'\hat{V}_0C)^{-1}D_n\sqrt{n}C'\bar{g}_n(\beta) \\ &= (D_n\sqrt{n}C'\bar{g}_n(\beta))'(C'\hat{V}_0C)^{-1}D_n\sqrt{n}C'\bar{g}_n(\beta) \quad \xrightarrow{d} (DN(0,I_\ell))'\cancel{(C'V_0C)^{-1}}DN(0,I_\ell) \\ &= (DN(0,I_\ell))'DN(0,I_\ell) \end{aligned}$$

D is a residual maker matrix with rank equal to the number of non zero eigenvalues (i.e. $\ell - k$), thus $DN(0, I_\ell)$ is a $\ell - k$ dimensional vector of independent standard normal random variables (same logic as deriving distribution of error variance in OLS). Thus the asymptotic distribution of $J(\hat{\beta}_{GMM})$ is $\chi^2(\ell - k)$. □

17 Panel data. Fixed effects.

17.1 Time invariant heterogeneity

A panel is a set of observations on individuals, collected over time. An observation is the pair (y_{it}, x_{it}), where i denotes the individual and t denotes time. The standard panel data specification is that there is an individual-specific effect δ_i that is constant over time:

$$y_{it} = x_{it}\beta + \delta_i + \varepsilon_{it}$$

such that $\mathbb{E}[\varepsilon_{it}|x_{i1}, \ldots, x_{iT}, \delta_i] = 0$ (strict exogeneity). If δ_i is observed then we can estimate β by OLS of $y_{it} - \delta_i$ on x_{it}. If δ_i is unobserved then we require either the lack of correlation between δ_i and x_{it} or the availability of an instrument w_i which is correlated with x_{it} but not $\delta_i + \varepsilon_{it}$.

First differences

If neither of the above are available we can still consistently estimate β by considering the regression in first differences:

$$y_{it} - y_{i,t-1} = (x_{it} - x_{i,t-1})'\beta + (\varepsilon_{it} - \varepsilon_{i,t-1}) \tag{17.1}$$

which is often written as the regression of Δy_{it} on Δx_{it}, where $\Delta y_{it} = y_{it} - y_{i,t-1}$ and $\Delta x_{it} = x_{it} - x_{i,t-1}$. Running OLS will not be optimal, since there is serial correlation in the error term. Thus the standard errors will be wrong and the estimator inefficient. To solve this we can use GLS.

17.2 Within-Group (or simply within) estimator

Let

$$y_{it} = x_{it}'\beta + \delta_i + \varepsilon_{it} \quad \text{where}$$

$$\underset{T\times 1}{y_i} = \begin{pmatrix} y_{i1} \\ \vdots \\ y_{iT} \end{pmatrix}, \quad \underset{T\times k}{x_i} = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iT} \end{pmatrix}, \quad \underset{T\times 1}{\varepsilon_i} = \begin{pmatrix} \varepsilon_{i1} \\ \vdots \\ \varepsilon_{iT} \end{pmatrix}$$

$$\underset{nT\times 1}{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1T} \\ \vdots \\ y_{n1} \\ \vdots \\ y_{nT} \end{pmatrix}, \quad \underset{nT\times k}{X} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \underset{nT\times 1}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix} = \begin{pmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{1T} \\ \vdots \\ \varepsilon_{n1} \\ \vdots \\ \varepsilon_{nT} \end{pmatrix}, \quad \underset{n\times 1}{\delta} = \begin{pmatrix} \delta_1 \\ \vdots \\ \delta_n \end{pmatrix}$$

We assume strict exogeneity $\mathbb{E}[\varepsilon_i|x_i, \delta_i] = 0$ and $Var(\varepsilon_i|x_i, \delta_i) = \sigma^2 I_T$. We can rewrite equation (1) in terms of the $(T-1) \times T$ matrix D with $D_{tt} = -1, D_{t,t+1} = 1$, and all other entries zero:

$$Dy_i = Dx_i\beta + D\delta_i + D\varepsilon_i = Dx_i\beta + D\varepsilon_i$$

$$\underset{(T-1)\times T}{D} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \Rightarrow \underset{(T-1)\times 1}{Dy_i} = \begin{pmatrix} y_{i2} - y_{i1} \\ \vdots \\ y_{iT} - y_{i,T-1} \end{pmatrix} = \Delta y_i$$

$$\Rightarrow \underset{(T-1)\times k}{Dx_i} = \begin{pmatrix} x_{i2} - x_{i1} \\ \vdots \\ x_{iT} - x_{i,T-1} \end{pmatrix} = \Delta x_i$$

This setup can also be written in terms of the matrices $Y, X, \varepsilon, \delta$, by considering the block diagonal matrix of D. Our assumptions become $\mathbb{E}[\varepsilon|X,\delta] = 0$ and $Var(\varepsilon|X,\delta) = \sigma^2 I_{nT}$. Let ℓ be a column vector of ones of length T. Then we can write the model as:

$$Y = X\beta + C\delta + \varepsilon \quad \text{where} \quad C = \begin{pmatrix} \ell & 0 & \dots & 0 \\ 0 & \ell & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \ell \end{pmatrix}$$

$$\begin{pmatrix} D & 0 & \dots & 0 \\ 0 & D & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D \end{pmatrix} Y = \begin{pmatrix} D & 0 & \dots & 0 \\ 0 & D & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D \end{pmatrix} X\beta + \begin{pmatrix} D & 0 & \dots & 0 \\ 0 & D & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D \end{pmatrix} \varepsilon$$

Note:-

Definition: The Kronecker product of two matrices A and B, denoted as $A \otimes B$, is a matrix formed by taking each element of A and multiplying it by the entire matrix B.

Example: Let $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$ and $B = \begin{pmatrix} 5 & 6 \\ 7 & 8 \end{pmatrix}$. The Kronecker product $A \otimes B$ is given by:

$$A \otimes B = \begin{pmatrix} 1 \cdot B & 2 \cdot B \\ 3 \cdot B & 4 \cdot B \end{pmatrix} = \begin{pmatrix} 5 & 6 & 10 & 12 \\ 7 & 8 & 14 & 16 \\ 15 & 18 & 20 & 24 \\ 21 & 24 & 28 & 32 \end{pmatrix}$$

Properties

- $(A \otimes B)' = A' \otimes B'$
- $A \otimes (C + D) = A \otimes C + A \otimes D$

There are many other useful properties, but these are all we need for this lecture.

We can now write the first difference estimator more succinctly as:

$$(I_n \otimes D)Y = (I_n \otimes D)X\beta + (I_n \otimes D)\varepsilon$$

GM Assumptions

GM1 $\text{rank}(I_n \otimes D)\text{X} = \text{k}$

This is equivalent to $\text{rank}(D)\text{X} = \text{k}$, which is equivalent to $\text{rank}(X) = \text{k}$

GM2 $\mathbb{E}[\Delta\varepsilon|\Delta X] = 0$

$\mathbb{E}[(I_n \otimes D)\varepsilon|\Delta X] = (I_n \otimes D)\mathbb{E}[\varepsilon|\Delta X] = 0$

GM3 Homoskedasticity and no serial correlation in $\Delta\varepsilon$

Even if we assume that ε_{it} is homoskedastic and serially uncorrelated, $\Delta\varepsilon_{it}$ will not be. $Cov(\varepsilon_{it} - \varepsilon_{i,t-1}, \varepsilon_{i,t-1} - \varepsilon_{i,t-2}) = -Var(\varepsilon_{it}|x_i) = -\sigma^2 \neq 0$.

Since GM1 and GM2 hold, the FD below is unbiased and consistent.

$$\hat{\beta}_{FD} = \left(\sum_{i=1}^{n}(Dx_i)'(Dx_i)\right)^{-1}\left(\sum_{i=1}^{n}(Dx_i)'(Dy_i)\right)$$

However GM3 is violated so the OLS estimate is not BLUE. To get the efficient estimator we need to use GLS.

Definition 17.2.1: GLS

If instead of GM3 we have that $Var(\varepsilon|X) = \Omega$ where Ω is known, then we can transform the data by premultiplying by $\Omega^{-1/2}$ to get: $Y^* = \Omega^{-1/2}Y, X^* = \Omega^{-1/2}X, \varepsilon^* = \Omega^{-1/2}\varepsilon$. Then we can run OLS on $Y^* = X^*\beta + \varepsilon^*$ to get the efficient estimator $\hat{\beta}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}Y$. Note that Ω can be scaled without any effect on the estimator.

We can see this results in efficient estimates by considering $Var(\varepsilon^*|X^*) = \Omega^{-1/2}Var(\varepsilon|X)\Omega^{-1/2} = \Omega^{-1/2}\Omega\Omega^{-1/2} = I_{nT}$. We now derive the GLS estimator for the FD model. We assume that the original regression satisfies $\mathbb{E}[\varepsilon|X] = 0$ and $Var(\varepsilon|X) = \sigma^2 I_{nT}$. Then,

$$\begin{aligned}
Var((I_n \otimes D)\varepsilon|X) &= (In \otimes D)Var(\varepsilon|X)(In \otimes D)' \\
&= \sigma^2 \begin{pmatrix} D & 0 & \dots & 0 \\ 0 & D & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D \end{pmatrix}\begin{pmatrix} D' & 0 & \dots & 0 \\ 0 & D' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & D' \end{pmatrix} \\
&= \sigma^2 \begin{pmatrix} DD' & 0 & \dots & 0 \\ 0 & DD' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & DD' \end{pmatrix}
\end{aligned}$$

We thus have our covariance matrix to use for scaling,

$$\Omega = \sigma^2 \begin{pmatrix} DD' & 0 & \dots & 0 \\ 0 & DD' & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & DD' \end{pmatrix} \Rightarrow \Omega^{-1/2} = \frac{1}{\sigma}\begin{pmatrix} (DD')^{-1/2} & 0 & \dots & 0 \\ 0 & (DD')^{-1/2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & (DD')^{-1/2} \end{pmatrix}$$

Thus,

$$\begin{aligned}
\hat{\beta}_{GLS-FD} &= \left(((I_n \otimes D)X)'\Omega^{-1}(I_n \otimes D)X\right)^{-1}((I_n \otimes D)X)'\Omega^{-1}(I_n \otimes D)Y \\
&= \left(X'(I_n \otimes D)'(I_n \otimes (DD')^{-1})(I_n \otimes D)X\right)^{-1} X'(I_n \otimes D)'(I_n \otimes (DD')^{-1})(I_n \otimes D)Y \\
&= \left(X'(I_n \otimes D')(I_n \otimes (DD')^{-1})(I_n \otimes D)X\right)^{-1} X'(I_n \otimes D')(I_n \otimes (DD')^{-1})(I_n \otimes D)Y \\
&= \left(X'(I_n \otimes D'(DD')^{-1}D)X\right)^{-1} X'(I_n \otimes D'(DD')^{-1}D)Y
\end{aligned}$$

Note that $D'(DD')^{-1}D$ is a projection matrix onto the space spanned by columns of D' (and thus is symmetric and idempotent). We now consider the column vector of 1s ℓ. Since the FD of ℓ is zero (i.e. always $1-1$), we have that $D'\ell = 0$. Thus a projection onto the space spanned by D' is equivalent to projection onto the space orthogonal to ℓ.

$$\Rightarrow D'(DD')^{-1}D = I_T - \ell(\ell'\ell)^{-1}\ell' = \frac{1}{T}\ell\ell'$$

Then,

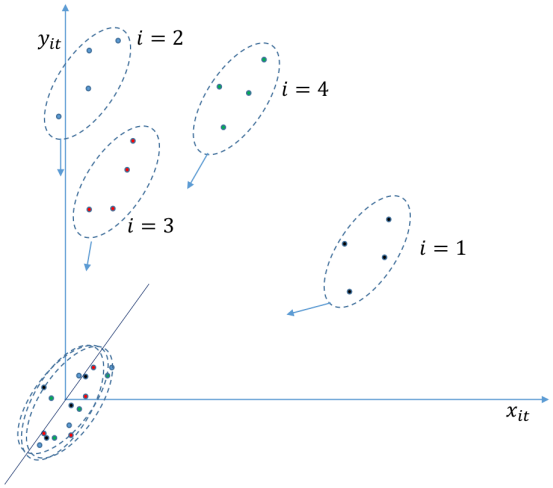
$$\begin{aligned}
\hat{\beta}_{GLS-FD} &= \left(X'(I_n \otimes (I_T - \frac{1}{T}\ell\ell'))X\right)^{-1} X'(I_n \otimes (I_T - \frac{1}{T}\ell\ell'))Y \\
&= \left(\sum_{i=1}^{n} x_i'(I_T - \frac{1}{T}\ell\ell')x_i\right)^{-1}\left(\sum_{i=1}^{n} x_i'(I_T - \frac{1}{T}\ell\ell')y_i\right) \\
&= \left(\sum_{i=1}^{n} x_i'(x_i - \ell\frac{\ell'x_i}{T})\right)^{-1}\left(\sum_{i=1}^{n} x_i'(y_i - \ell\frac{\ell'y_i}{T})\right)
\end{aligned}$$

$$x_i - \ell\frac{\ell'x_i}{T} = \begin{pmatrix} x_{i1} \\ \vdots \\ x_{iT} \end{pmatrix} - \frac{1}{T}\begin{pmatrix} \ell'x_i \\ \vdots \\ \ell'x_i \end{pmatrix} = \begin{pmatrix} x_{i1} - \frac{1}{T}\sum_{t=1}^{T} x_{it} \\ \vdots \\ x_{iT} - \frac{1}{T}\sum_{t=1}^{T} x_{it} \end{pmatrix} = \underbrace{\begin{pmatrix} x_{i1} - \bar{x}_i \\ \vdots \\ x_{iT} - \bar{x}_i \end{pmatrix}}_{T \times k}$$

$$\begin{aligned}
\hat{\beta}_{GLS-FD} &= \left(\sum_{i=1}^{n} x_i'(x_i - \bar{x}_i)\right)^{-1}\left(\sum_{i=1}^{n} x_i'(y_i - \bar{y}_i)\right) \\
&= \left(\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'\right)^{-1}\left(\sum_{i=1}^{n}\sum_{t=1}^{T}(x_{it} - \bar{x}_i)(y_{it} - \bar{y}_i)\right)
\end{aligned}$$

This estimator is known as the within or within-group estimator because it exploits only the variation of x_{it} and y_{it} within the groups of observations corresponding to the same individual but different time periods. It ignores the variability between groups by subtracting individual specific time averages from the data. Note that the errors are not uncorrelated, indeed every error term appears in every other equation (via the demeaning).

A geometric interpretation of the within transformation is given here. The individual effects are the 'intercepts' of the individual specific scatters. Here they are negatively correlated with $x_i t$ so fitting that regression line to the untransformed data results in negative bias. The within transformation shifts the data so that the individual specific scatters have zero intercepts, and thus the regression line is unbiased.



END OF LECTURE