Ordinary Least Squares and Variance

$$Y = X\beta + \varepsilon \qquad E[\varepsilon | X] = 0$$

Estimate β by $\hat{\beta} = (X'X)^{-1}X'Y$

$$E[\hat{\beta}] = \beta$$

$$Var(\hat{\beta} | X) = \sigma^2 (X'X)^{-1} \quad \text{if } Var(\varepsilon | X) = \sigma^2 I_n$$

Homoskedasticity

Suppose $\Sigma := Var(\varepsilon | X)$ is something else

- Heteroskedasticity: ε_i's are independent of each other but dependent on X_i.

  by $Var(\varepsilon_i | X_i) = \sigma^2(X_i)$ for some nonnegative function

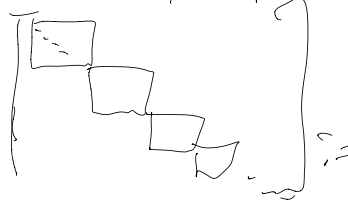
- Clustering: ε's are separately groups with shared characteristics and so nonzero covariances

Shows up with "panel data" forms

two indices $i = (j, t)$ $j = 1 \dots J$ $t = 1 \dots T$

each unit j comes in groups of size T

ε may have form



$\sigma_{jt,s} := Cov(\varepsilon_{j,t}, \varepsilon_{j,s}) \neq 0$

- Sequential structure:

- Time series or spatial data have built-in correlation

Units arrange in ordered set $1, \ldots, T$

$\sigma_{|s-t|} = \text{cov}(\varepsilon_s, \varepsilon_t)$ may depend on $|s-t|$

$$\Sigma = \begin{bmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \cdots \\ \cdots & \ddots & \ddots & \ddots \\ \cdots & \ddots & \ddots & \sigma_2 \\ \cdots & \cdots & \sigma_1 & \sigma^2 \end{bmatrix}$$

- What happens to OLS when $\Sigma \neq \sigma^2 I$

Gauss Markov no longer true. Can improve on OLS

Σ is positive definite and symmetric

Let's further suppose Σ^{-1} exists, will also be positive definite and symmetric

We can define square root of $\underset{N\times N}{\Sigma^{-1}} = \underset{N\times N}{C'} \underset{N\times N}{C}$

not unique, example: Cholesky decomposition

C is upper triangular $C_{ij} = 0$ if $i > j$

C' is lower triangular

- Use the spectral decomposition

If M is positive definite and symmetric on $\mathbb{R}^{N\times N}$

$M = V' D V$ where D is diagonal, $D_{ij} = 0$ for $i \neq j$

V is orthogonal $V'V = I_N$

$VV' = I_N$

D is nonnegative and entries are the eigenvalues of M

columns of V are eigenvectors

ie $Mv_i - d_{ii} I_N = 0 \qquad v_i \neq 0$,

Square root: $C = D^{\frac{1}{2}} V \qquad D^{\frac{1}{2}}$ takes square root of Diagonal entries

$d_1 = \max_{\|v\|=1, v \in \mathbb{R}^N} v' M v$

Restrict since set $\{v \in \mathbb{R}^N : \|v\| = \sqrt{\sum v_i^2} = 1\}$ is compact and $v'Mv$ is continuous

$f(v) = v' M v - \lambda (v'v - 1)$

$\frac{\partial f}{\partial v} = 0$ at optimum $\quad 2v'M - 2\lambda v' = 0$ FOC

$Mv = \lambda v \quad$ at $v = v_1$

So λ is some eigenvalue for eigenvector v_1

So $v_1' M v_1 = v_1' \lambda v_1 = \lambda v_1' v_1 = \lambda$

So optimum is $\lambda = d_{11}$

Define Q_1 as $[v_1, Z]$ where Z is

$e_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{pmatrix}$ i^{th} orthogonal to v_1 (defined by projecting basis vectors of $\mathbb{R}^N$ onto orthogonal complement of space of V_1

$$MQ_1 = [d_{11} v_1, MZ]$$

So $Q_1' M Q_1 = \begin{bmatrix} d_{11} & * \\ 0 & Z'MZ \end{bmatrix}$ 0's by orthogonality

Can apply same proof to $Z'MZ$ which is in $\mathbb{R}^{N-1 \times N-1}$

and symmetric and positive definite and independent of β

for set sequence $d_{11}, d_{22}, \ldots d_{N,N}$ of eigenvalues

and $Q_1, Q_2, \ldots$ eventually reach V

We can define estimators after from OLS

Apply $E[C(Y - X\beta)|X] = CE[\varepsilon|X] = 0$

so $CE[Y|X] = E[CX|X]\beta$

can premultiply by $X'C'$ to get

$$E[X'C'CY|X] = E[X'C'CX|X]\beta$$

so estimate by

$$\hat{\beta}^{GLS} = (X'C'CX)^{-1}X'C'CY$$

$$= (X'\Sigma^{-1}X)^{-1}X'\Sigma^{-1}Y$$

Generalized Least squares

Properties: Unbiased:

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

Conditional Variance $E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'|X] =$

$$E\left[(X'\Sigma^{-1}X)^{-1}(X'\Sigma^{-1}\varepsilon\varepsilon'\Sigma^{-1}X)(X'\Sigma^{-1}X)^{-1}|X\right]$$

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

Can apply same argument as in Gauss Markov theorem to show this is "best" linear unbiased estimator

Problem: don't know Σ usually

Replace Σ by an estimate to get "feasible GLS"

Depending on properties of $\hat{\Sigma}$ estimator, might be good or bad in finite samples.

Large sample properties of OLS

Use summation representation:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{k1} & \cdots & x_{K1} \\ \vdots & & \vdots & & \vdots \\ x_{1i} & \cdots & x_{ki} & \cdots & x_{Ki} \\ \vdots & & \vdots & & \vdots \\ x_{1N} & \cdots & x_{kN} & \cdots & x_{KN} \end{bmatrix} = \begin{bmatrix} x_1' \\ \vdots \\ x_N' \end{bmatrix}$$

$$x_i' = (x_{1i}, \ldots x_{ki}, \ldots x_{Ki})$$

K variables for observation i

So $X'X = \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} x_1' \\ \vdots \\ x_N' \end{bmatrix} = \sum_{i=1}^{N} x_i x_i'$

similarly $X'y = \sum_{i=1}^{N} x_i y_i$ since y_i is scalar

So OLS has

$$\hat{\beta} = (X'X)^{-1}X'y = \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1}\left(\sum_{i=1}^{N} x_i y_i\right)$$

Use to write as sample averages

$$= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right)$$

Assume w.w (y_i, x_i') are i.i.d and $E[\varepsilon_i | x_i] = 0$ for all i

Consistency

$$\hat{\beta} = \beta + \left[\frac{1}{N}\sum x_i x_i'\right]^{-1}\left[\frac{1}{N}\sum x_i \varepsilon_i\right]$$

Take $N \to \infty$

$$\frac{1}{N}\sum x_i x_i' \xrightarrow{p} E[x_i x_i'] \equiv \Sigma_{xx}$$

by Law of Large numbers if this exists

Now, if Σ_{xx}^{-1} exists, by continuity

$$\text{plim}\left[\frac{1}{N}\sum x_i x_i'\right]^{-1} = \left[\text{plim}\ \frac{1}{N}\sum x_i x_i'\right)^{-1}$$

$$= \Sigma_{xx}^{-1}$$

Next: by CLN

$$\frac{1}{N}\sum x_i \varepsilon_i \xrightarrow{p} E[x_i \varepsilon_i] = E\left[x_i E(\varepsilon_i | x_i)\right]$$

$$= E[x_i \cdot 0] = 0$$

So $\hat{\beta} \xrightarrow{p} \beta$: consistency

In fact, we didn't need $E[\varepsilon_i | x_i] = 0$ for this Argument just used $E[x_i \varepsilon_i] = 0$ - this is entirely general: doesn't restrict distributional form

N.A. $E[x_i y_i] = E[x_i x_i' \beta] + \underbrace{E[x_i \varepsilon_i]}_{0 \text{ by assumption}}$

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

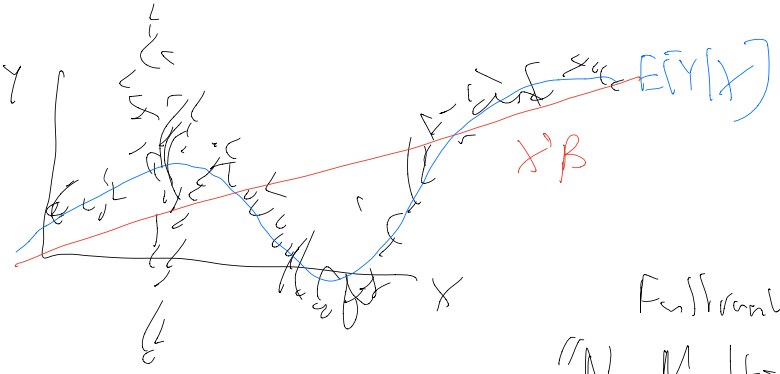
and we can define β as the object from any distribution with $E x_i x_i'$ full rank

with β defined this way,

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

satisfy $E[x_i \varepsilon_i] = 0$

Interpretation: OLS estimates line of best fit of y given x, even if conditional expectation not linear.



Full rank $E x_i x_i'$ assumption "No Multicollinearity"

Gives a unique line of best fit

If it fails, can choose one of many lines of best fit arbitrarily.

Use a pseudo-inverse instead of inverse.

Distribution Theory

- Need new results about convergence in distribution
- Relationship of convergence in probability

- Suppose $\hat{\theta}_n \xrightarrow{p} \theta$ a constant

  Then $\hat{\theta}_n \xrightarrow{d} \theta$

Proof:

$$Pr(\hat{\theta} < z) = Pr(\hat{\theta} - \theta < z - \theta)$$

$$\geq Pr(|\hat{\theta} - \theta| < |z - \theta|)$$

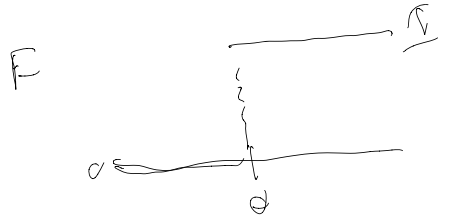
$$\to 1 \quad \text{as } n \to \infty \text{ if } z > \theta$$

by convergence in probability

Can also show by symmetric argument

$$Pr(\hat{\theta} < z) \to 0 \text{ if } z < \theta$$

So CDF converges to CDF of a constant θ at all points of continuity



Slutsky's Theorem:

Suppose $A_n \xrightarrow{d} Z$, $B_n \xrightarrow{p} b$ a constant, Then

$$(A_n, B_n) \xrightarrow{d} (Z, b)$$

Statistics can be replaced by their plim in asymptotic distributions.

Note: statement that $A_n \xrightarrow{d} Z$, $B_n \xrightarrow{d} W$ does not imply $(A_n, B_n) \xrightarrow{d} (Z, W)$

joint distribution is not defined for Z, W

Asymptotic Distribution of OLS

$$\sqrt{N}(\hat{\beta} - \beta) = \underbrace{\left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1}}_{\xrightarrow{p} \Sigma_{xx}^{-1}} \underbrace{\left(\frac{1}{\sqrt{N}}\sum x_i \varepsilon_i\right)}_{\text{apply CLT}}$$

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i \varepsilon_i \xrightarrow{d} Z \sim N\left(0, E\left[x_i \varepsilon_i \varepsilon_i x_i'\right]\right)$$

So by Slutsky and continuous mapping

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} \Sigma_{xx}^{-1} Z$$

$$\sim N\left(0, \Sigma_{xx}^{-1}\left(E x_i \varepsilon_i \varepsilon_i x_i'\right)\Sigma_{xx}^{-1}\right)$$

$$Var(Av) = A\,Var(v)\,A'$$

If $E[\varepsilon_i | x_i] = \sigma^2$ then variance simplifies to

$\sigma^2 \Sigma_{xx}^{-1}$ which can be estimated by $\hat{\sigma}^2 (X'X)^{-1}$ as before.

Otherwise, if $E(\varepsilon_i^2 | X_i)$ is not always constant, limit still holds, we can estimate the variance

$$\Sigma_{xx}^{-1} E X_i \varepsilon_i \varepsilon_i X_i' \Sigma_{xx}^{-1}$$

Called "Robust" or "Heteroskedasticity Consistent" Variance estimation, or Eicker-Huber-White

Estimate by sample analogues: replace expectation by average, replace ε_i by $\hat{\varepsilon}_i = Y_i - X_i \hat{\beta}$

$\left(\frac{1}{N} \sum X_i X_i'\right)^{-1} \xrightarrow{p} \Sigma_{xx}^{-1}$ so that's handled

Check $\frac{1}{N} \sum X_i X_i' \hat{\varepsilon}_i^2$

$$= \underbrace{\frac{1}{N} \sum X_i X_i' \varepsilon_i^2}_{\xrightarrow{p} E[X_i X_i' \varepsilon_i^2]} + \underbrace{\frac{1}{N} \sum X_i X_i' \left(\hat{\varepsilon}_i^2 - \varepsilon_i^2\right)}$$

Need to show is that

$$\frac{1}{N} \sum X_i X_i' \left(\hat{\varepsilon}_i^2 - \varepsilon_i^2\right) \xrightarrow{p} 0$$

$$\hat{\varepsilon}_i = Y_i - X_i \hat{\beta} = X_i \beta + \varepsilon_i - X_i \hat{\beta} = -X_i (\hat{\beta} - \beta) + \varepsilon_i$$

$$\text{so } \hat{\varepsilon}_i^2 = (\beta - \hat{\beta}) X_i X_i' (\beta - \hat{\beta}) + \varepsilon_i^2 - 2 \varepsilon_i X_i' (\hat{\beta} - \beta)$$

$$\text{So } \frac{1}{N} \sum X_i X_i' \left(\hat{\varepsilon}_i^2 - \varepsilon_i^2\right) =$$

$$\frac{1}{N}\sum x_i x_i' (\beta-\hat\beta)' x_i x_i' (\beta-\hat\beta) - \frac{2}{N}\sum_{i=1}^{N} x_i x_i' \varepsilon_i x_i' (\beta-\hat\beta)$$

$$= \sum_{j=1}^{K}\sum_{k=1}^{K} \underbrace{(\beta_j-\hat\beta_j)}_{\xrightarrow{p}0}\underbrace{(\beta_k-\hat\beta_k)}_{\xrightarrow{p}0} \underbrace{\frac{1}{N}\sum x_i x_i' x_{ji} x_{ki}}_{\xrightarrow{p} E[x_i x_i' x_{ji} x_{ki}] \text{ if finite}} - \sum_{k=1}^{K} \underbrace{(\beta_k-\hat\beta_k)}_{\xrightarrow{p}0} \underbrace{\frac{2}{N}\sum_{i=1}^{N} x_i x_i' \varepsilon_i x_{ki}}_{\xrightarrow{p} E[x_i x_i' \varepsilon_i x_{ki}]}$$

$$\xrightarrow{p} 0$$

$$\text{So } \hat\Sigma^{\text{robust}} = \left(\frac{1}{N}\sum x_i x_i'\right)^{-1}\left(\frac{1}{N}\sum x_i x_i' \hat\varepsilon_i^2\right)\left(\frac{1}{N}\sum x_i x_i'\right)^{-1}$$

$$\xrightarrow{p} \Sigma_b \text{ asymptotic variance of } \hat\beta$$

If data is iid, you can almost always use this formula; if homoskedastic, this consistent, but not, as we will see.

reg y x, robust

Application of these formulas

Tests (and later, confidence intervals)

Goal is to distinguish between "hypotheses"

H_0 (null hypothesis)

H_1 (alternative hypothesis)

Two disjoint sets of probability measures over data
usually corresponding to disjoint parameter sets.
We will use shorthand:

$H_0: \beta = \beta_0$, $H_1: \beta \neq \beta_0$

is shorthand for $H_0 = \{$ probability measures
under some set of assumptions where population
regression coefficient $= \beta_0 \}$

and $H_1 = \{$ set of measures satisfying same assumptions except
β in population $\neq \beta_0 \}$

Test: a function of data $\phi: \{Z_i\}_{i=1}^N \to \{0, 1\}$

that picks H_0 or H_1

(can have a "randomized test" that is also function
of additional random variable(s))