Background: Previously, we learned properties of estimators $\hat{\theta}$ when we observe iid data $\{w_i\}$ drawn from some distribution $P(\theta)$ (the estimators are functions of the data w_i. Now, we will observe data $\{w_i\} = \{y_i, x_i\}$ from some distribution $P(\theta)$ where we think of y as the "outcome/dependent" variable and x as the "covariate/independent" variables. We will be interested in the conditional mean $E_P(Y_{1x1} \mid X_{d-1x1}) \equiv \mu_P(X_{d-1x1})$. Though the functional form of $\mu_P(X)$ can be quite general, we study here the case where the specific linear functional form where $\mu_P(X) = \beta_0(P) + \beta_1(P)X = X'_{1xd}\ \beta_{dx1}$ (Here we add a constant to X to include the intercept term)

→ We characterize true $\beta = (\beta_0(P), \beta_1(P))$ as the minimizer of MSE (i.e. **least squares**): $\beta = \arg\min_{b \in R^e} E_P\left[\left(Y_{1x1} - X'_{1xd}\ b_{dx1}\right)^2\right]$

→ (Assuming the random vector $X_{dx1}$is such that $E(XX')^{-1}$ and Y is a random scalar variable with first 2 moments finite)

Then: True $\beta = (\beta_0(P), \beta_1(P))$ is **uniquely** given by $\beta_{dx1} = \left[E_P(X_{dx1}X'_{1xd})\right]^{-1} E_P(X_{dx1}Y_{1x1})$

→ By Analogy principle: $\hat{b}_{dx1} = \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right] = \left[\frac{1}{n}X'X\right]^{-1}\left[\frac{1}{n}X'Y\right]$ *where* X_{nxd} Y_{nx1} *are observed iid data*

→ This is the Ordinary Least Squares estimator of the "true" $\beta = (\beta_0(P), \beta_1(P))$

OLS Estimator: Suppose we observe $\{W_i\}_{i=1}^n = \{Y_i, X_i\}_{i=1}^n$ iid from some distribution P.

- 2 Expressions: $\hat{b} = (X'X)^{-1}X'y = (X'X/n)^{-1}X'y/n = S_{XX}^{-1}s_{XY}$
- 2 Methods of Derivation:
  1. Linear Algebra: Find the "best" approximation of ***y*** in the image of X by finding $b\ s.t. \|y - Xb\|$ *smallest*

     → Find orthogonal projection of y onto the space spanned by the columns of X

     $\|y - Xb\| \le \|y - X\beta\|\ \forall\ \beta \in \Re^n \Leftrightarrow Xb = \Pr oj_{\mathrm{Im}(X)}(y) = y^{\|}$ *by def or orth. proj*

     $\Rightarrow y^{\perp} = y - y^{\|} = y - Xb \in \left[\mathrm{Im}(X)\right]^{\perp} = Ker(X')$

     $\Rightarrow X'(y - Xb) = 0$

     $\Rightarrow X'y = X'Xb$

     If $X'X$ *ivertible, then* $b_{OLS} = (X'X)^{-1}X'y$ *and the "best" approx is* $\hat{y} = Xb = X(X'X)^{-1}X'y = P_{im(X)}\ y$

  2. Calculus: Minimize SSR/ sum of prediction error

     $SSR(b) = (y - Xb)'(y - Xb) = y'y - y'Xb - b'X'y + b'X'Xb = y'y - 2y'Xb + b'X'Xb$

     (since y'Xb a scalar and therefore equals its transpose)

     $FOC: \frac{\partial SSR(b)}{\partial b} = \frac{\partial (y'y - 2y'Xb + b'X'Xb)}{\partial b} = -2y'X + 2X'Xb = 0$

     $\Rightarrow X'Xb = y'X = X'y$ *bc* $y'X$ *scalar*

     If $X'X$ *invertible, then* $b_{OLS} = (X'X)^{-1}X'y$ *and the "best" approx is* $\hat{y} = Xb = X(X'X)^{-1}X'y = P_{im(X)}y$

- **Projection and Annihilator Matrices (P and M)**[1]

$$P_{nxn} \equiv X(X'X)^{-1}X',\ M_{nxn} \equiv I - P, \text{ and } A_{kxn} \equiv (X'X)^{-1}X'$$

Properties:

a) I = P + M
b) P and M are symmetric and idenpotent[2]
c) PX = X (Since projecting X onto the column space of X gives you the same thing)
d) MX = 0 (Since vectos in im(X) is orthogonal to the M space)
e) $y = Py + My = y^{\|} + y^{\perp}$ (We can always right a vector in Rn as the projection onto 2 orthogonal subspaces. Follows from a.)
f) $\hat{y} = Py$ (Fitted y is just the orthogonal projection of y onto the column space of x)
g) A matrix returns the linear combination of X that is the projection of a vector onto column space of X: Ay = β, XAy=X β
h) PM = 0, MA' = 0, PA'=A [3] (This is intuitive bc for any y, Ay lives in column space of X)
i) Trace(M) = n-k
j) E(SSR) = E($\varepsilon'M\varepsilon$)= σ^2 Trace(M)

[1] P is the projection onto the column space of X and M is the projection onto the space that is orthogonal to Im(X).
[2] P idempotent → $P^2 = P$
[3] $PA' = X(X'X)^{-1}X'X(X'X)^{-1} = X(X'X)^{-1} = A',\ PM = P(I - P) = P - P^2 = P - P = 0,\ MA' = (I - P)A' = A' - PA' = A' - A' = 0$

- **SSR, SER, and Sampling Error: Relationship to ε (True error term)**

$$SSR = Sum\ of\ \ squared\ \ OLS\ \ residuals = (y - X\hat{b})'(y - X\hat{b}) = e'e = \varepsilon' M\varepsilon$$ [4]

$$s^2 = OLS\ \ estimate\ \ of\ \ \sigma^2 (\text{var}\, of\ error\ term) \equiv \frac{SSR}{n-K} = \frac{e'e}{n-K}$$ [5]

$$SER = Std\ Err\ of\ the\ regression = Est\ of\ \ std\ dev\ of\ \ error\ term = \sqrt{s^2}$$

$$Sampling\ Error \equiv \hat{b} - \beta = (X'X)^{-1}X'y - \beta = (X'X)^{-1}X'(X\beta + \varepsilon) - \beta = (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon - \beta = (X'X)^{-1}X'\varepsilon = A\varepsilon$$

(sampling error is the projection of the error matrix onto the column space of X → why??)

How do we think about the difference between the sampling error and the true error?
What does it mean to project a random vector onto a space?

Models

A Model is a set of restrictions on the joing distribution of the dependent and independent variables- i.e. a model is a set of joint distributions satisfying a set of assumptions.

We will see 3 models, each of which makes a set of assumptions about the joint distribution of (y,x)

M1: Classical Regression (Assumptions 1~5) (with Gaussian Errors: Assumption 6)
M2: Generalized Least Squares - Relax Conditional Homoskdasticity and No Serial Correlation (Relax Assumption 4a and 4b)
M3: Relax Everything

[4] BC $y = X\beta + \varepsilon \Rightarrow My = MX\beta + M\varepsilon \Rightarrow My = M\varepsilon$ by d) $\Rightarrow e = M\varepsilon \Rightarrow e'e = \varepsilon' M'M\varepsilon = \varepsilon' MM\varepsilon = \varepsilon' M\varepsilon$ (by symmetry and idempotency)

[5] n-K is the degrees of freedom. The idea is that K parameters need to be estimated (i.e. β_{Kx1}) before obtaining the residual vector e used to calculate s^2. More specifically, e has to satisfy the k normal equations: X'Xb = X'y → X'(y-e) = X'y → X'e = 0

5 Key Assumptions to Classical Regression Model (M1):

1. (Linearity)[6]: Relationship between dependent var and regressors is linear.

$$\boxed{Y_{nx1} = X_{nxk}\beta_{kx1} + \varepsilon_{nx1} \quad or \quad y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} + \varepsilon_i \ (i = 1,2,...,n)}$$

2. (Strict Exogeneity): Regressors are strictly exogenous s.t. error terms are random with (conditional and unconditional) mean 0

$$\boxed{\begin{aligned} &E(\varepsilon_i \mid X_{nxk}) = E(\varepsilon_i \mid x_{1_{kx1}}, x_{2_{kx1}}, ..., x_{n_{kx1}}) = 0 \ \ (i = 1,2,...,n) \Leftrightarrow E(\varepsilon \mid X) = 0 \\ &\Rightarrow E(Y - X\beta \mid X) = 0 \Rightarrow E(Y \mid X) = E(X\beta \mid X) = X\beta \\ &\Rightarrow E(Y \mid X) = X\beta = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} \end{aligned}}$$

Intuition: On average our model makes the "right" guess of y! Here the mean is conditional on regressors for **all** observations. In other words, if we take the joint distribution of the nK + 1 random variables, $f(\varepsilon_i, x_1, ..., x_n)$, and consider the conditional distribution $f(\varepsilon_i \mid x_1, ..., x_n)$. In general the conditional mean $E(\varepsilon_i \mid X) = E(\varepsilon_i \mid x_1, x_2, ..., x_n)$ is a nonlinear function of $(x_1, \ldots, x_n)$. The strict exogeneity assumption says that this function is a constant of value 0.

Note that this assumption is also not too restrictive if we allow for a constant term in the regressor.

Implications of strict exogeneity:

a) Unconditional Mean = 0:

$$\boxed{E(\varepsilon_i) = E(E(\varepsilon_i \mid X)) = E(0) = 0}$$

b) Each *Regressor is orthogonal to the error term for **all** observations:*

$$\boxed{E(x_{jk}\varepsilon_i) = E(E(x_{jk}\varepsilon_i \mid x_{jk})) = E(x_{jk}E(\varepsilon_i \mid x_{jk})) = 0 \ \ (i, j = 1,...,n; k = 1,..,K)}$$

$$since \quad E(\varepsilon_i \mid x_{jk}) = E(E(\varepsilon_i \mid X) \mid x_{jk}) = 0$$

c) Each *regressor and each error term are uncorrelated: (This follows from a and b, since the assumption says that each of the regressors is linearly uncorrelated with the error term)*

$$\boxed{Cov(x_{jk}, \varepsilon_i) = E(x_{jk}\varepsilon_i) - E(x_{jk})E(\varepsilon_i) = 0}$$

3. (No Multicollinearity): Regressors are linearly independent from each other

Rank(X) = K with probability 1

Note: This allows us to say that (X'X) is invertible!

4a. *(Homoskedasticity): Conditional second moment is a constant*

$$Var(\varepsilon_i \mid X) = E(\varepsilon_i^2 \mid X) - E(\varepsilon_i \mid X)^2 = E(\varepsilon_i^2 \mid X) = \sigma^2 > 0 \quad for \ i = 1,2,...n \ ^{7} \ \boxed{\Leftrightarrow Var(\vec{\varepsilon} \mid X) = E(\varepsilon\varepsilon' \mid X) = \sigma^2 I_n}$$

4b. (No Serial Correlation between Observations):

$$Cov(\varepsilon_i, \varepsilon_j \mid X) = E(\varepsilon_i \varepsilon_j \mid X) = 0 \quad for \ i, j = 1,2,...n; i \neq j$$

(This follows from 4., since the conditional var-cov matrix is diagonal, the off-diag. elements are 0)

Note: 4 and 5 → the nxn matrix of conditional second moments $E(\varepsilon\varepsilon' \mid X) = Var(\varepsilon \mid X)$ is spherical, i.e. proportional to the identity matrix.

For Gaussian Model: (For purposes of hypothesis testing, we often add the following distributional assumption)

5. (Normality of the Error Term):

$$\varepsilon \mid X \sim N(0, \sigma^2 I_n) \Leftrightarrow Y \mid X \sim N(X'\beta, \sigma^2 I_n)$$

Note: These assumptions are about the ENTIRE sample. We impose no iid assumption.

If (**y**,**X**) is a random sample, i.e. $\{y_i, x_i\}$ iid across observations, then our assumptions can be restated for single observations…

1. $Y_{nx1} = X_{nxk}\beta_{kx1} + \varepsilon_{nx1} \quad or \quad y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_K x_{iK} + \varepsilon_i \ (i = 1,2,...,n)$ (same as before)
2. $E(\varepsilon_i \mid x_i) = 0 \ (i = 1,2,...,n) (since \ (x_i, \varepsilon_i) \perp (x_j, \varepsilon_j) \ for \ i \neq j \ so \ E(\varepsilon_i \mid X_{nxk}) = E(\varepsilon_i \mid x_{1_{kx1}}, x_{2_{kx1}}, ..., x_{n_{kx1}}) = E(\varepsilon_i \mid x_i) = 0)$
3. *Rank(X) = K with probability 1*(Same)
4. $E(\varepsilon_i^2 \mid x_i) = \sigma^2 > 0 \ \forall i \ \ and \ \ Cov(\varepsilon_i \varepsilon_j \mid x_i, x_j) = 0$

The implication of iid sample is that the joint distribution of (x_i, ε_i) **does not depend on *i*.** So the unconditional second moment $E(\varepsilon_i^2)$ is constant across *i* – **we get unconditional homoskedasticity with iid, and the <u>functional form</u> of the conditional second moment** $E(\varepsilon_i^2 \mid x_i)$ is same across *i*! **However**, the value of the conditional second moment is same across *i* does not follow, so assumption 4 (conditional homoskedasticity) remains restrictive for random samples.

[6] This is not as restrictive an assumption as it seems, because many non-linear relationships can be linearized (e.g. by "logging")

[7] Homoskedasticity says that the conditional variance of the error term is a constant → conditional second moment is constant by strict exogeneity.

Classical Regression (assumptions 1 ~5): Properties of OLS Estimator

1. Finite Sample Properties of M1 OLS estimator[8]

(a) **Unbiasedness**: Under 1 ~3, $E(b \mid X) = \beta$

Note: Strict exogeneity is CRITICAL in proving this, anything short of it will not suffice. For example, it is not enough to assume $E(\varepsilon_i \mid \bar{x}_i) = 0 \ \forall i$ *or* $E(\bar{x}_i \cdot \varepsilon_i) = 0 \ \forall i$ (*orthogonality*). Since most time series models do not satisfy strict exogeneity and at most the orthogonality condition, the OLS estimator is not unbiased.

(b) **Variance of $\mathbf{b_{OLS}}$**: Under 1~5, $Var(\hat{b} \mid X) = \sigma^2 (X'X)^{-1}$

(c) **Gauss-Markov Th**: Under 1~5, the OLS estimator is **efficient** in the class of linear and unbiased estimator. That is, for any unbiased $\hat{\beta}$ that is linear in y,

$Var(\hat{\beta} \mid X) \geq Var(\hat{b}_{OLS} \mid X)$ in the matrix sense (i.e. $Var(\hat{\beta} \mid X) - Var(b_{OLS} \mid X)$ is a p.s.d. matrix)

Meaning: For any regression coeff, the variance of the OLS est is no larger than that of any other linear unbiased estimator.

$Var(\hat{\beta} \mid X)_{kxk} \geq Var(b_{OLS} \mid X)_{kxk}$ *p.s.d* $\Rightarrow a'\left[Var(\hat{\beta} \mid X) - Var(b_{OLS} \mid X)\right]a \geq 0$ *for any* $kx1$ *vector* a

$\therefore$ *for* $a = (0,0,...,1,0,0,...)$ *vector with* $i-th$ *element* 1 *and* 0 *o.w., above also true*

$\Rightarrow Var(\hat{\beta}_i \mid X) \geq Var(\hat{b}_i \mid X)$ *for* $i = 1,...,k$

BLUE (Best Linear Unbiased Estimator): Gauss-Markov says that since OLS is linear and unbiased; then, by above proof, among all other linear and unbiased estimators of β, the OLS estimator is efficient in the sense that its conditional variance matrix $Var(b_{OLS} \mid X)_{kxk}$ is the smallest among the linear unbiased estimators → That's why OLS is called BLUE.

Note: **This does not mean that OLS estimator is the most efficient among all linear estimators! (why?)**

(d) **Covariance of OLS Estimator and errors[9]**: Under 1~4, $Co\,\text{var}(b_{OLS}, e \mid X) = 0$ *where* $e \equiv y - X\,b$

More precisely, this means that every random RV in $\hat{b}$ is uncorrelated with every RV in e: i.e.

$Cov(\hat{b}_i, e_j) = E((\hat{b}_i - E(\hat{b}_i))(e_j - E(e_j))) = 0$ *for* $i = 1,...,k$ *and* $j = 1,...,n$

Or, in "outer product" form: $Cov\,(\hat{b}_{kx1}, e_{nx1}) = E((\hat{b} - E(\hat{b} \mid X))_{kx1}(e - E(e \mid X))'_{1xn} \mid X) = 0$

---

[8] *a.* $E(\hat{b} \mid X) = E\left[(X'X)^{-1}X'y \mid X\right] = (X'X)^{-1}X'E(y \mid X) = (X'X)^{-1}X'X\beta$ *by strict exogeneity*
$= \beta$

b. $Var(\hat{b} \mid X) = E\left[\hat{b}\hat{b}' \mid X\right] - E\left[\hat{b} \mid X\right]E\left[\hat{b} \mid X\right]' = E\left[\hat{b}\hat{b}' \mid X\right] - \beta\beta'$ *by* (a)

$= E\left[(X'X)^{-1}X'yy'X(X'X)^{-1} \mid X\right] - \beta\beta' = (X'X)^{-1}X'E\left[yy' \mid X\right]X(X'X)^{-1} - \beta\beta'$

$= (X'X)^{-1}X'(Var(y \mid X) + E(y \mid X)E(y \mid X)')X(X'X)^{-1} - \beta\beta'$

$= (X'X)^{-1}X'(Var(y \mid X) + E(y \mid X)E(y \mid X)')X(X'X)^{-1} - \beta\beta'$

$= (X'X)^{-1}X'Var(y \mid X)X(X'X)^{-1} + (X'X)^{-1}X'X\beta\beta'X'X(X'X)^{-1} - \beta\beta'$

$= (X'X)^{-1}X'Var(y \mid X)X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2 X(X'X)^{-1}$ *by* 5

$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$ *bc* σ^2 *a cons*

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

c. $\hat{\beta}$ *linear in y, so we write as* $\hat{\beta} = Cy$ *for some matrix C. Let* $D \equiv C - A \Rightarrow C = D + A$ *where* $A \equiv (X'X)^{-1}X'$

$\hat{\beta} = Cy = (D + A)y = Dy + Ay = D(X\beta + \varepsilon) + \hat{b} = DX\beta + D\varepsilon + \hat{b}$

$\Rightarrow E(\hat{\beta} \mid X) = E(DX\beta + D\varepsilon + \hat{b} \mid X) = DX\beta + E(D\varepsilon \mid X) + E(\hat{b} \mid X) = DX\beta + DE(\varepsilon \mid X) + \beta$ *since* $\hat{b}$ *unbiased*

$\Rightarrow \beta = DX\beta + DE(\varepsilon \mid X) + \beta$ *since* $\hat{\beta}$ *unbiased*

$\Rightarrow DX\beta = 0$ *since* $E(\varepsilon \mid X) = 0$ *by strict exogeneity* (*this also means that* $DX = 0$ *for any non – zero* β)

$\Rightarrow \hat{\beta} = D\varepsilon + \hat{b}$

$\Rightarrow \hat{\beta} - \beta = D\varepsilon + \hat{b} - \beta = D\varepsilon + A\varepsilon = (D + A)\varepsilon$ (*by property of OLS – Sampling Error*)

$\therefore Var(\hat{\beta} \mid X) = Var(\hat{\beta} - \beta \mid X) = Var((D + A)\varepsilon \mid X) = (D + A)Var(\varepsilon \mid X)(D + A)'$ *since* D, A *functions of* X

$= \sigma^2(D + A)(D + A)'$ *since* $Var(\varepsilon \mid X) = \sigma^2 I_n$ *by cond hom.*

$= \sigma^2(DD' + AD' + DA' + AA')$ *and we know* $DA' = DX(X'X)^{-1} = 0$ *bc* $DX = 0$, *and* $AA' = (X'X)^{-1}X'X(X'X)^{-1} = (X'X)^{-1}$

$= \sigma^2(DD' + (X'X)^{-1}) = \sigma^2 DD' + \sigma^2(X'X)^{-1} = \sigma^2 DD' + Var(\hat{b} \mid X)$

$\geq Var(\hat{b} \mid X)$ (*since* DD' *p.s.d*)

[9] *d.* $\hat{b} - E(\hat{b} \mid X) = \hat{b} - \beta = A\varepsilon$, *and* $e - E(e \mid X) = M\varepsilon - ME(\varepsilon \mid X) = M\varepsilon$

$\Rightarrow Cov(\hat{b}, e) = E(A\varepsilon(M\varepsilon)' \mid X) = E(A\varepsilon\varepsilon'M' \mid X) = AE(\varepsilon\varepsilon' \mid X)M' = \sigma^2 AM = \sigma^2 AM' = 0$ *since* $MA' = 0$ (A *lives in* $im(X)$ *proj onto* $M = 0$)

2. Finite Sample Properties of s^2 [10]

Recall, $s^2 \equiv \frac{SSR}{n-K} = \frac{e'e}{n-K} = \frac{(y - X\,b_{OLS})'(y - X\,b_{OLS})}{n-K}$

(a) **Unbiasedness of s^2**: Under 1~4, $E(s^2 \mid X) = \sigma^2$, provided n > K (so that s^2 is well-defined)

Implication: By law of iterated expectation, $E(s^2) = E(E(s^2 \mid X)) = \sigma^2$

(b) **Estimate of Var(b|X)** [11]: Since s^2 is the estimate of σ^2, a natural estimate of $Var(\hat{b} \mid X) = \sigma^2 (X'X)^{-1}$ is

$$\hat{Var}(\hat{b} \mid X) = s^2 (X'X)^{-1} = e'e(X'X)^{-1}/(n-K)$$

Using the above results on the variance of b, and the following (0.) distribution of b, we can now test hypothesis about OLS regression coefficients, assuming normality in the error terms

<u>(Finite Sample) Hypothesis Testing in Gaussian Regression (i.e. under Normality assumption):</u>

0. Preliminary: Distribution of b → $\boxed{\hat{b} \mid X \sim N(\beta, \sigma^2 (X'X)^{-1}) \text{ and sampling error } (\hat{b} - \beta) \mid X \sim N(0, \sigma^2 (X'X)^{-1})}$

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

Since $\varepsilon \mid X \sim N(0, \sigma^2 I_n)$ *by 2 and 6,* $\Rightarrow b \mid X$ *also normal*

We know from finite sample property of OLS estimator that $E(\hat{b} \mid X) = \beta$ and $Var(\hat{b} \mid X) = \sigma^2 (X'X)^{-1}$

So, $\boxed{\hat{b} \mid X \sim N(\beta, \sigma^2 (X'X)^{-1}) \text{ and sampling error } (\hat{b} - \beta) \mid X \sim N(0, \sigma^2 (X'X)^{-1})}$

1. Testing Hypotheses about Individual OLS Regression Coefficients

Suppose H_0: $\beta_i = \overline{\beta}_i$, *K:* $\beta_i \neq \overline{\beta}_i$

$$\boxed{\text{Under Null}, (\hat{b}_i - \overline{\beta}_i) \mid X \sim N(0, \sigma^2 \left((X'X)^{-1}\right)_{ii}) \Rightarrow z_k \equiv \frac{\hat{b}_i - \overline{\beta}_i}{\sqrt{\sigma^2 \left((X'X)^{-1}\right)_{ii}}} \sim N(0,1)}$$

If we do not know the true population variance σ^2, then we can conduct hypothesis testing using sample S.E. and the t-distribution.

Under the null $\beta_i = \overline{\beta}_i$,

$$\boxed{t_k \equiv \frac{\beta_i - \overline{\beta}_i}{SE(\beta_i)} = \frac{\beta_i - \overline{\beta}_i}{\sqrt{s^2 \left((X'X)^{-1}\right)_{ii}}} \sim t(n-K)}$$ [12]

---

[10] Since s^2= e'e/n-K, the proof amounts to showing that E(e'e|X) = $(n-K)\sigma^2$:

From previous, $e'e = \varepsilon' M \varepsilon$. *Show* $1.\ E(\varepsilon' M \varepsilon \mid X) = \sigma^2 \cdot Trace(M)$ *and* $2.\ Trace(M) = n - K$

1. $E(\varepsilon' M \varepsilon \mid X) = E\left(\sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij}\varepsilon_i\varepsilon_j \mid X\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} m_{ij} E(\varepsilon_i\varepsilon_j \mid X) = \sigma^2 \sum_{i=1}^{n} m_{ii}$ since $E(\varepsilon_i\varepsilon_j \mid X) = 0$ *for* $i \neq i$ *by no serial corr* $= \sigma^2 \cdot Trace(M)$

2. $Trace(M) = Trace(I - P) = n - Trace(P) = n - K$

and $Trace(P) = Trace(X(X'X)^{-1}X') = Trace(X'X(X'X)^{-1}) = Trace(I_K) = K$ since $Trace(AB) = Trace(BA)$

[11] This is the standard error of the OLS estimate that gets spit out by STATA

[12] Recall… **Def**: *Let Z have a normal distribution with mean 0 and variance 1. Let V have a chi-square distribution with v degrees of freedom. Suppose that Z and V are independent. Then,* $T = \frac{Z}{\sqrt{V/v}} \sim t_v$

Here, $\frac{\hat{b}_i - \overline{\beta}_i}{\sqrt{s^2 \left((X'X)^{-1}\right)_{ii}}} = \frac{\hat{b}_i - \overline{\beta}_i}{\sqrt{\sigma^2 \left((X'X)^{-1}\right)_{ii}} \sqrt{\frac{s^2}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{e'e/(n-K)}{\sigma^2}}} = \frac{Z}{\sqrt{\frac{e'e/\sigma^2}{n-K}}}$, $\frac{\hat{b}_i - \overline{\beta}_i}{\sqrt{\sigma^2 \left((X'X)^{-1}\right)_{ii}}} \sim N(0,1) = Z$ and $\frac{e'e}{\sigma^2} \sim \chi^2(n-K)$

Note: $\frac{e'e}{\sigma^2} = \frac{(M\varepsilon)'M\varepsilon}{\sigma^2} = \frac{\varepsilon'M'M\varepsilon}{\sigma^2} = \frac{\varepsilon'M\varepsilon}{\sigma^2} = \frac{\varepsilon'}{\sigma} M \frac{\varepsilon}{\sigma}$ since $\varepsilon \sim N(0, \sigma^2 I_n) \Rightarrow \frac{\varepsilon}{\sigma} \sim N(0,1)$

Fact: if $x \sim N(0, I_n)$, *and A a symmetric projection matrix (i.e. idempotent), then,* $x'Ax \sim \chi^2(rank(A))$. Here, rank(M) = trace(M) = n-K

Fact: if A idempotent, then rank(A) = trace(A)

2. Testing Linear (Joint) Hypotheses about OLS Regression Coefficients (Wald Test)

We can generalize the above to test null hypothesis that impose a restriction on not only a single individual coefficient but a linear combination of them, written as a system of linear equation.

Suppose H_0: $R_{\#rxK}\beta_{Kx1} = r_{\#rx1}$ *where* ***R*** *and* ***r*** *are known and specified by the hypothesis,*

#r *is the dimension or* ***r*** *or # of equations (#r ≤ K)*

And Rank(R) = #r *(we impose R to have full row rank to make sure that there are no redundant equations and that equations are consistent with each other)*

We reject/don't reject null based on the following finite sample distribution of Wald Statistic[13]

Under the null (If null were true), the Wald statistic is F(#r,n-K) distributed. (Note: This is also the Lagrange Mult Stat)

$$F \equiv \frac{(R\hat{b}-r)'\left[R(X'X)^{-1}R'\right]^{-1}(R\hat{b}-r)/\#r}{s^2} = (R\hat{b}-r)'\left[R\hat{Var}(\hat{b}\mid X)R'\right]^{-1}(R\hat{b}-r)/\#r \sim F_{\#r,n-K}$$

Decision Rule: Reject for large values of W!

Alternatively, we can use the following (more convenient) expression of F, using the likelihood-ratio principle.[14]

$$F = \frac{(SSR_R - SSR_U)/\#r}{SSR_U/(n-K)}$$

3. Note on t vs. F Statistics:

A. F and t tests are equivalent when testing hypotheses about individual coefficients:

Since hypotheses about individual coefficients are linear hypotheses, the t-test of H_0: $\beta_i = \overline{\beta}_i$ can be written as

$R\beta = r$ *where* $R = vector\ w.kth\ element = 1$ *and* $r = \overline{\beta}_i \Rightarrow F = (\beta_i - \overline{\beta}_i)\left[s^2\left[(X'X)^{-1}\right]_{ii}\right](\beta_i - \overline{\beta}_i) = t^2$ (T^2~F in general)

B. If null is that a set of individual regression coefficient equal certain values, then F is preferred to T.
   1. If the size (sig level) in each of the 2 t-tests is alpha, then overall sizeis not alpha
   2. F test is a likelihood ratio test and LR tests have certain desirable properties. (See next)

Finally, we have to show that $\hat{b}_i - \overline{\beta}_i$ and $e'e/\sigma^2$ are independently distributed.

Recall: For (X,Y) Jointly normal, if Cov(X,Y) = 0 → X, Y independent. Since, $\hat{b} = \beta + A\varepsilon$ *and* $e = M\varepsilon$ → b and e are jointly normal conditional on X. Also they are uncorrelated, so b and e are independently distributed conditional on X. Since Z and $\frac{e'e}{\sigma^2}$ are functions of b and e, therefore they are also independent.

[13]**Recall the F statistic. Def:** *Let U ~ ChiSq(u) RV, V ~ ChiSq(v), U and V independent RV. Then,* $F = \frac{U/u}{V/v} \sim F_{u,v}$

We can write $W = \frac{w/\#r}{q/n-K}$ $where\ w = (R\hat{b}-r)'(\sigma^2R(X'X)^{-1}R')^{-1}(R\hat{b}-r)/\#r$ *and* $q = \frac{e'e}{\sigma^2}$

1. *w~chi-sq(#r) under the null: Under null,* $R\hat{b} - r = R\hat{b} - R\beta = R(\hat{b}-\beta)$, $\hat{b} \sim N(\beta, \sigma^2(X'X)^{-1}) \Rightarrow R\hat{b} \sim N_{\#r}(R\hat{b}, \sigma^2R(X'X)^{-1}R')$

   Recall, if ***x is an m-dimensional random vector with*** $x \sim N_m(\mu, \Sigma_{mxm})$ ***, then*** $(x-\mu)\Sigma^{-1}(x-\mu)' \sim \chi^2(m)$

   So, w ~Chi-Sq(#r)
2. *q is Chi-Sq(n-K) as shown above*
3. *w and q independent by same reasoning above: w is a function of b and q a function of e, b and e independent → w and q independent.*

(Note: This derivation of the F-Ratio is based on the Wald-Principle, and therefore often called the Wald Statistic, because it is based on ONLY the unrestricted estimator, which is not constrained to satisfy the restrictions of the null hypothesis.)

[14]See 270 PS3 Q. 7

Relation to Maximum Likelihood (with normally distributed errors)

1. **ML estimator of (β, σ^2) [15]: Suppose Assumptions 1~5 hold. Then the MLE of β is the OLS estimator $\hat{b}$**

   and MLE of $\sigma^2 = \frac{1}{n} e'e = \frac{SSR}{n} = \frac{n-K}{n} s^2$

2. **OLS/ML estimator of β is UMVUE (since it reaches CRLB and is unbiased)[16]: Under assumptions 1 ~6, the OLS estimator is UMVUE in that any other unbiased (but not necessarily linear) estimator has larger conditional variance in the matrix sense.**

   Note: OLS estimator s^2 of σ^2 does not attain the CRLB (since $\text{Var}(s^2 \mid X) = 2\sigma^4 / n-K$)[17] and is not unbiased, therefore not UMVUE.

---

[15] Under assumptions 1~6,

$$y \mid X \sim N(X\beta, \sigma^2 I) \Rightarrow p(y \mid X; \beta, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2}(y-X\beta)'(\sigma^2 I)^{-1}(y-X\beta)\right\} = (2\pi\sigma^2)^{-n/2} \exp\left\{-\frac{1}{2\sigma^2}(y-X\beta)'(y-X\beta)\right\}$$

$$\Rightarrow \log p(y \mid X; b, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y-Xb)'(y-Xb)$$

$$Max_{b,\sigma^2} -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y-Xb)'(y-Xb)$$

1. $FOC\ wrt\, b\ (Hold\ \sigma^2\ fixed): \frac{1}{2\sigma^2} X'(2I_n)(y-Xb) = 0 \Rightarrow X'y = X'Xb$

$$\Rightarrow \hat{b} = (X'X)^{-1}X'y$$

2. $FOC\ wrt\, \sigma^2: -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y-Xb)'(y-Xb) = 0 \Rightarrow \sigma^2 = (y-Xb)'(y-Xb)/n$

$$\Rightarrow (by\ ML\ \text{Principle, plug in}\ \hat{b}),\ \hat{\sigma}^2 = (y-Py)'(y-Py)/n = e'e/n = \frac{1}{n} SSR(\hat{b}) = \frac{n-K}{n} s^2$$

[16] Recall the CR Inequality: *Suppose that T(**X**) is a real valued statistic and* $E_\theta[T(X)] = \psi(\theta)_{1x1}$ *and let* $\frac{\partial \psi(\theta)}{\partial \theta}$ *denote the dx1 vector of partial derivatives and suppose that* $I(\theta)^{-1}$ *is nonsingular and the conditions of the above theorem hold. Then, for all* $\theta \in \Theta$ *,*

$$Var(T(X)) \geq \frac{\partial \psi(\theta)}{\partial \theta}' I(\theta)^{-1} \frac{\partial \psi(\theta)}{\partial \theta}' = \frac{\partial \psi(\theta)}{\partial \theta}' \frac{I_1(\theta)^{-1}}{n} \frac{\partial \psi(\theta)}{\partial \theta}' \ \text{if}\ X_i\ i.i.d.\ p(\cdot, \theta)\ \text{where}\ I(\theta) = -E_\theta\left(\frac{\partial^2}{\partial\theta\partial\theta'} \log p(x \mid \theta)\right)$$

Here, $\theta = (\beta, \sigma^2)'_{k+1x1} \Rightarrow \frac{\partial^2}{\partial\theta\partial\theta'} \log p(y \mid X; \theta) = \begin{bmatrix} \frac{\partial^2 \log p(y \mid X;\theta)}{\partial\beta\partial\beta'}_{(KxK)} & \frac{\partial \log p(y \mid X;\theta)}{\partial\beta\partial\sigma^{2,}}_{(Kx1)} \\ \frac{\partial \log p(y \mid X;\theta)}{\partial\sigma^2\partial\beta'}_{(1XK)} & \frac{\partial \log p(y \mid X;\theta)}{\partial\sigma^2\partial\sigma^{2,}}_{(1x1)} \end{bmatrix}_{(K+1xK+1)}$

$$\frac{\partial^2 \log p(y \mid X;\theta)}{\partial\beta\partial\beta'} = \frac{\partial}{\partial\beta'}\left(\frac{1}{\sigma^2} X'(y-X\beta)\right) = -\frac{1}{\sigma^2} X'X \Rightarrow -E\left(-\frac{1}{\sigma^2} X'X\right) = \frac{1}{\sigma^2} X'X$$

$$\frac{\partial^2 \log p(y \mid X;\theta)}{\partial\beta\partial\sigma^2} = \frac{\partial}{\partial\sigma^2}\left(\frac{1}{\sigma^2} X'(y-X\beta)\right) = -\frac{1}{\sigma^4} X'(y-X\beta) \Rightarrow -E\left(-\frac{1}{\sigma^4} X'(y-X\beta)\right) = \frac{1}{\sigma^4} X'E(y-X\beta) = 0$$

$$\frac{\partial^2 \log p(y \mid X;\theta)}{\partial\sigma^2\partial\sigma^{2,}} = \frac{\partial}{\partial\sigma^2}\left(-\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(y-X\beta)'(y-X\beta)\right) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y-X\beta)'(y-X\beta) \Rightarrow -E\left(\frac{n}{2\sigma^4} - \frac{1}{\sigma^6}(y-X\beta)'(y-X\beta)\right) = \frac{n}{2\sigma^4}$$

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

[17] $s^2 = \frac{e'e}{n-K} = \frac{\varepsilon' M \varepsilon}{n-K}, \varepsilon \sim N(0, \sigma^2 I_n) \Rightarrow \frac{\varepsilon'}{\sigma} M \frac{\varepsilon}{\sigma} \sim \chi^2(rank(M)) = \chi^2(n-K) \Rightarrow \frac{s^2(n-K)}{\sigma^2} \sim \chi^2(n-K) \Rightarrow Var\left(\frac{s^2(n-K)}{\sigma^2}\right) = 2(n-K)$

$$\Rightarrow Var(s^2)\frac{(n-K)^2}{\sigma^4} = 2(n-K) \Rightarrow Var(s^2) = \frac{2\sigma^4}{(n-K)}$$

Trinity and Hypothesis Testing under Gaussian Errors:

M2: Generalized Least Squares (GLS) – Relaxing assumptions 4a and 4b (Conditional Homoskedasticity and No Serial Correlation)

(we observe (y, X, ε), which we assume to satisfy linearity, strict exogeneity, and no multicollineaity, but $E(\varepsilon\varepsilon' | X) \neq \sigma^2 I_n$)

0. **What it Means:** If error is not (conditionally) homoskedastic, the values of the diagonal elements of $E(\varepsilon\varepsilon' | X)$ are not the same.
   If there is correlation in errors between observations, then off-diagonal elements of $E(\varepsilon\varepsilon' | X)$ are non-0.
1. **Consequences of Relaxing Assumptions 4a** (Conditional Homoskedasticity) **and 4b** (No Serial Correlation):
   A. Gauss Markov Theorem no longer holds for the OLS estimator $\hat{b} = (X'X)^{-1}X'y$: **BLUE is some other estimator**
   B. T-test (for linear restriction on a single coefficient) and F-test (joint linear restriction) is no longer valid: **t-ratio and Wald statistic/f-ratio no longer t and f distributed.**
   C. OLS still unbiased, since unbiasedness does not require assumptions 4a and 4b, only assumptions 1 ~ 3
2. **GLS Model (M2) Assumptions:**

   Assumptions 1 ~ 3 from M1 (Linearity, Strict Exogeneity, and No Multicollinearity)

   Assumption 4c: $\boxed{E(\varepsilon\varepsilon' | X) = \sigma^2 V(X)}$ for some V(X) known symmetric, positive definite (non-singular/invertible) matrix.
   (We factor out σ^2 from every element for convenience sake – not necessary)
3. **GLS Estimator and BLUE: If the matrix function $V(X)=V^{-1}$ is known, then by a clever transformation of the data, we can obtain a model that satisfies Gaussian Model assumptions and the OLS estimator for this model (GLS) will be BLUE**

   For any symmetric positive definite matrix V, there exists nonsingular matrix C s.t. $V^{-1} = C'C$[18].

   $$\boxed{Let\ \ \tilde{y} \equiv Cy,\ \ \tilde{X} \equiv CX,\ \ \tilde{\varepsilon} \equiv C\varepsilon}$$

   - $(\tilde{y}, \tilde{X}, \tilde{\varepsilon})$ satisfies assumptions 1 ~ 4:
     1. Linearity: From assumption 1 for (y, X, ε), our model $y = X\beta + \varepsilon$ can be re-written as: $\tilde{y} = \tilde{X}\beta + \tilde{\varepsilon}$
     2. Strict Exogeneity: $E(\tilde{\varepsilon} \mid \tilde{X}) = E(\tilde{\varepsilon} \mid X)$ $(X \text{ and } \tilde{X} \text{ contain same inf}) = E(C\varepsilon \mid X) = CE(\varepsilon \mid X) = 0$ *by Assumption*2
     3. No Multicollinearity: C nonsingular → $Rank(\tilde{X}) = Rank(X) = K$ with probability 1
     4. Conditional Homoskedasticity: $E(\tilde{\varepsilon}\tilde{\varepsilon}' | X) = E(C\varepsilon\varepsilon' C' | X) = CE(\varepsilon\varepsilon' | X)C' = CV(X)C' = C\left(\sigma^2 V\right)C' = \sigma^2 CVC' = \sigma^2 I_n$
   - ***GLS Estimator is just the OLS Estimator applied to the transformed model!***

     $$\boxed{\hat{\beta}_{GLS} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{y} = \left[(CX)'(CX)\right]^{-1}(CX)'Cy = \left[X'C'CX\right]^{-1}X'C'Cy = (X'V^{-1}X)^{-1}X'V^{-1}y}$$
4. **Finite Sample Properties of GLS Estimator**[19]
   (a) Unbiasedness: Under GLS assumptions 1 ~ 3, $E(\hat{\beta}_{GLS} \mid X) = \beta$
   (b) Expression for the Variance: Under assumptions 1~4, $Var(\hat{\beta}_{GLS} \mid X) = \sigma^2(\tilde{X}'\tilde{X})^{-1} = \sigma^2(X'V(X)^{-1}\tilde{X})^{-1}$
   (c) Efficiency of GLS (Gauss-Markov Applied): Under assumptions 1 ~4, the GLS is efficient in that the conditional variance of any unbiased estimator that is linear in **y** is greater than or equal to $Var(\hat{\beta}_{GLS} \mid X)$ in the matrix sense.

Note on the limiting nature of GLS:

The "nice" finite sample properties of GLS rest on the strict exogeneity assumption as well as the knowledge of V(X) (i.e. what is V as a function of the data X). As we'll see in time-series contexts (where errors are often serially correlated), strict exogeneity is too strong and limits the usefulness of the GLS procedure. So neither OLS nor GLS have the nice finite-sample properties such as unbiasedness. However, given that the regressors are pre-determined (weaker assumption than strict exogeneity), then the OLS estimator, which ignores the serial correlation in the error, will have some good large sample properties such as consistency and asymptotic normality. GLS will not have these properties under the weaker assumption.

V *symmetric* $\Rightarrow V$ *orthogonally diagonalizable by spectral theorem*
$\Rightarrow V = K\Lambda K', , K$ *a matrix of orthonormal eigenvectors* (*orthogonal*), *and* Λ *matrix of eigenvalues*

[18] K *orthogonal* $\Rightarrow K'K = I \Rightarrow K' = K^{-1}$
$\therefore V^{-1} = K\Lambda^{-1}K'$ (*bc* $V^{-1}V = I \Rightarrow V^{-1}K\Lambda K' = I \Rightarrow V^{-1} = K\Lambda^{-1}K'$)
$= K\Lambda^{-1/2}\Lambda^{-1/2}K' = C'C$ *where* $C = \Lambda^{-1/2}K'$
Note: This also implies that $V^{-1} = C'C \Rightarrow (C')^{-1}V^{-1}(C)^{-1} = I \Rightarrow CV^{-1}C' = I \Rightarrow CVC' = I$

[19] Proofs are same as before but applied to the transformed data $(\tilde{y}, \tilde{X}, \tilde{\varepsilon})$

5. 3 Special Cases:

A. Weighted Least Squares (Heteroskedasticity Correction): Conditional heteroscedasticity (Var of y varies with x), no corr errors.

e.g. We observe $\{y_i, x_i\}_{i=1}^n$ *iid* with $y_i = x_i'\beta + \varepsilon_i$ where $E(\varepsilon_i \mid X) = 0$ and no serial correlation (assumptions 1-3, 4b hold).

Conditional heteroskedasticity: $E(\varepsilon_i^2 \mid X) = [\sigma(x_i)]^2$ *for some known function* $\sigma(x_i) \Rightarrow Var(y_i \mid X) = \sigma^2(x_i)$

Variance of y varies with x!

Then, V and C matrices are given by:

$$V(X) = \begin{bmatrix} \sigma^2(x_1) & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & \sigma^2(x_n) \end{bmatrix} \Rightarrow V(X)^{-1} = \begin{bmatrix} 1/\sigma^2(x_1) & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & 1/\sigma^2(x_n) \end{bmatrix} \Rightarrow C = \begin{bmatrix} 1/\sigma(x_1) & 0 & 0 \\ 0 & . & 0 \\ 0 & 0 & 1/\sigma(x_n) \end{bmatrix} (so\, C'C = V^{-1})$$

Then we obtain the transformed system (divide everything by $\sigma(x_1)$)

$$\tilde{y}_i = \tilde{x}_i'\beta + \tilde{\varepsilon}_i \Leftrightarrow \frac{y_i}{\sigma(x_i)} = \frac{x_i}{\sigma(x_i)}'\beta + \frac{\varepsilon_i}{\sigma(x_i)}$$

If we know the function $\sigma(x_i)$ then we can implement the above scheme and perform OLS on the transformed model. (In practice we don't know such function, so we either have to assume a functional form or estimate this unknown function)

Note: In this model, observations with higher conditional variance get a lower weight and vice versa.

B. Prais-Whinston Correction (Serially Correlated Errors)

e.g. Suppose now we have $y_i = x_i'\beta + \varepsilon_i$ with $E(\varepsilon_i \mid X) = 0$ and $E(\varepsilon_i^2 \mid X) = \sigma^2$ (i.e. assumptions 1 – 4a hold),

but $Cov(\varepsilon_i, \varepsilon_j) = \sigma^2 \rho^{|i-j|}$ $|\rho| < 1$ (this form of correlation is called AR(1)).

Then, V and C matrices are given by: $V(X) = \sigma^2 \begin{bmatrix} 1 & \rho & ... & \rho^{n-1} \\ \rho & 1 & \rho & ... \ \rho^{n-2} \\ \rho^2 & \rho & 1 & . \\ & & & 1 \end{bmatrix} \Rightarrow C = \begin{bmatrix} 1 & 0 & ... & 0 \\ -\rho & 1 & 0 & \\ 0 & -\rho & 1 & \\ 0 & 0 & -\rho & 1 \end{bmatrix}$

C. Random Effects Model (Can also be interpreted as GLS) – See later.

Group Regressions vs. Pooled Regressions

Running OLS regression by group is same as pooling regression using dummy variables (the difference is what assumptions you make about the conditional variance – if you assume that the conditional variance varies by group, then run separate regressions. If you assume that the conditional variance is the same across groups – i.e. conditional homoskedasticity holds – then run pooled regression. Either way, the OLS coefficient point estimates are the same).

(SEE HW1 of 271)

Large Sample Theory of OLS Estimator[20]: In previous section, assumptions of M1 and M2 are fairly strict. We show here that relaxing many of the assumptions, OLS (while do not have the finite sample properties) will have nice asymptotic properties.

1. Basic Time Series Concepts:

Stochastic process : a sequence of random variables $\{z_i\}$

Realization/Sample Path of a stochastic process: a realization of $\{z_i\}$ → sequence of real numbers

Time Series: If the sequence of rv's is indexed by time, the stochastic process is called a time series. (we often will use "time series" to refer to the realization and the stochastic process)

Ensemble Mean: $\{\bar{z}_i\}$ ("true" mean of each of the rv's in the sequence)

Need for Ergodic Stationarity: The fundamental problem in time-series analysis is that we observe the realization of the process only once. (i.e. we get only 1 sample and 1 observation of realized $\{z_i\}$!) Ideally, we would like to observe history many times over to obtain more samples, but clearly this is not feasible. But if each of the z_i's come from the same distribution (**stationarity**) , then we can view each realization of $\{z_i\}$ as n realizations from the same distribution. Furthermore, if the process is not too persistent (**ergodicity**), then each element of $\{z_i\}$ will contain some information not available from the other elements. In this case, the time average over the elements of $\{z_i\}$ will be **consistent for the ensemble mean**!

Defining Stationary and Erogodic Processes

Strictly Stationary Processes: A stochastic process $\{z_i\}$ $(i = 1,2...)$ is (strictly) stationary if the joint distribution of $(z_i, z_{i_1}, z_{i_2}, ..., z_{i_r})$ depends only on (for any given finite integer r and any set of subscripts $i_1,\ldots,i_r$) i_1-i, i_2-i, …,i_r-i but not i. (e.g. joint distribution of $(\mathbf{z}_1,\mathbf{z}_5)$ is same as $(\mathbf{z}_{12},\mathbf{z}_{16})$ → what matters is the relative position in the sequence not the absolute position!)

Weakly Stationary Process: A stochastic process $\{z_i\}$ $(i = 1,2...)$ is weakly (or covariance) stationary if:

i. $E(\mathbf{z}_i)$ does not depend on i and
ii. $Cov(\mathbf{z}_i,\mathbf{z}_{i-j})$ exists, is finite, and depends only on j but not on i (e.g. $Cov(\mathbf{z}_1,\mathbf{z}_5) = Cov(\mathbf{z}_{12},\mathbf{z}_{16})$)

Ergodic Process: A stationary process $\{z_i\}$ is said to be ergodic if, for any two bounded functions: $f : \Re^K \to \Re,\ g : \Re^L \to \Re$

$$\lim_{n\to\infty} \left|E\left(f(z_i,...,z_{i+k})g(z_{i+n},...,z_{i+n+l})\right| = \left|E\left(f(z_i,...,z_{i+k})\right)\right|\left|E\left(g(z_{i+n},...,z_{i+n+l})\right)\right|$$

Heuristically, **a stationary process is ergodic (i.e. ergodic stationarity) if it is asymptotically independent** → i.e. any 2 rv's or random vectors positioned far apart in the sequence are almost independently distributed. **Ergodic stationarity is important in developing large sample theory because of the ergodic theorem**.

Ergodic Theorem: Let $\{z_i\}_{i=1}^n$ be a stationary and ergodic process with $E(z_i) = \mu$. Then, $\boxed{\bar{z}_n \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow[a.s.]{} \mu}$

Idea: This generalizes Kolmogorov's LLN[21], becau… se **ergodic theorem allows for serial dependence** (whereas Kol's rules it out by iid assumption), **provided that the serial dependence disappears in the "long-run" (by stationary ergodic), i.e. asymptotically in large samples.**

Implication: Any moment of a stationary and ergodic process (if exists and finite) is consistently estimated by the sample moment. [22]

Vector Process, Martingales, Martingle Differences

Martingales Vector Process: A vector process $\{g_i\}$ is called a martingale if $E(g_i \mid g_{i-1},...,g_1) = g_{i-1}$ *for* $i \geq 2$

Note: The conditioning set $g_{i-1},...,g_1$ is often called the information set, and $\{g_i\}$ is called a martingale since its information set is its own past values.

Martingale Difference Sequence: A vector process $\{g_i\}$ with $E(g_i) = 0$ is called a margingale difference sequence (m.d.s.) or martingale difference if the expectation conditional on its past values is also 0: $\boxed{E(g_i \mid g_{i-1},...,g_1) = 0 \quad for \quad i \geq 2}$

Important Property: A martingale difference sequence has no serial correlation $Cov(g_i,g_j) = 0$

Ergodic Stationary Martingale Differences CLT: Let $\{g_i\}$ be a vector martingale difference sequence that is stationary and ergodic with $E(g_i g_i') = \Sigma$ [23], and let $\bar{g} \equiv \frac{1}{n}\sum_{i=1}^{n} g_i$. Then, $\boxed{\sqrt{n}\bar{g} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_i \xrightarrow{D} N(0,\Sigma)}$

---

[20] OLS procedure is important in econometrics because it has good asymptotic properties for a class of models (different from M1) that are useful in economics. We present a model here with the widest range of economic applications, it replies on no distributional assumption (normality in error terms not necessary) and the strict exogeneity assumption is replace by a much weaker assumption that they are predetermined.

[21] Kolmogorov's Second Strong Law of Large Numbers: Let $\{z_i\}_{i=1}^n$ be iid with $E(z_i) = \mu$ .Then, $\bar{z}_n \equiv \frac{1}{n}\sum_{i=1}^{n} z_i \xrightarrow[a.s.]{} \mu$

[22] Since if $\{z_i\}$ ergodic stationary, then $\{f(z_i)\}$ also ergodic stationary for any measurable function *f(.)*.

2. **Assumptions**[24]: To study large sample properties of OLS, we relax assumption of M1 further and impose some add'l assumptions
The idea is that under the following set of (less strict) assumptions, though we can't say nice finite sample properties, we can claim nice asymptotic properties.

The DGP[25] (sequence of RV's that generated our (finite) sample/data) satisfy the following:

Linearity:

$y_{i\,(1x1)} = x_i'_{(1xk)}\,\beta_{(kx1)} + \varepsilon_i \quad for\ i = \overline{1,n}$ where $\boldsymbol{x_i}$ a vector of explanatory variables, ε_i unobserved error term

Ergodic Stationarity: (Elements in the sequence are asymptotically independent so serial dependence disappears)

The K+1 dimensional vector stochastic process $\{y_i, \boldsymbol{x}_i\}$ is jointly stationary and ergodic.

(This assumption, again, allows for serial dependence to disappear as sample gets large, and allows us to invoke ergodic theorem and ergodic stationary CLT)

Note: iid samples are trivially ergodic stationary.

Predetermined Regressors / (Contemporanous) Orthogonality Condition:

All the regressors are **predetermined** in the sense that they are orthogonal to the contemporaneous error term:

$E(x_{ik}\varepsilon_i) = 0\ \forall i, \forall k \ \Leftrightarrow\ \mathrm{E}[\ \boldsymbol{x}_i\ .\ (y_i - \boldsymbol{x}_i'\beta)\ ] = 0 \ \Leftrightarrow \mathrm{E}(\mathbf{g}_i) = \mathbf{0}$ where $\mathbf{g}_i = \boldsymbol{x}_i\ .\ (y_i - \boldsymbol{x}_i'\beta) = \boldsymbol{x}_i\ .\varepsilon_i$

Note: This is weaker than strict exogeneity

Rank Condition: No multicollinearity in the limit

KxK matrix $\mathrm{E}(\boldsymbol{x}_i\,\boldsymbol{x}_i')$ is nonsingular (and hence finite). Denote $\mathrm{E}(\boldsymbol{x}_i\,\boldsymbol{x}_i') = \boldsymbol{\Sigma_{XX}}$

Martingale Difference with Finite Second Moments: $\boldsymbol{g}_i$ is a martingale difference sequence with finite second moments

$\{\ \boldsymbol{g}_i\ \}$ is a martingale difference sequence (so $\mathrm{E}(\mathbf{g}_i) = \mathbf{0}$ with $\mathrm{E}(\mathbf{g}_i \mid \mathbf{g}_{i-1}, \mathbf{g}_{i-2}, \ldots, \mathbf{g}_1) = \mathbf{0}$ for $i \geq 2$) → no serial correlation in $\boldsymbol{g}_i$

The KxK matrix of cross moments, $\mathrm{E}(\mathbf{g}_i\,\mathbf{g}_i')$ is nonsingular.

→ so, $\bar{g}_{nx1} = \dfrac{1}{n}\sum_{i=1}^{n} g_i \xrightarrow{P} E(g_i g_i') = S \equiv A\,\mathrm{var}(\bar{g}_{nx1})$ by Ergodic Differences CLT

Note: This assumption is stronger than 2.3 (since $\boldsymbol{g}_i$ m.s.d → $\mathrm{E}(\boldsymbol{g}_i) = 0$). We need it to derive asymptotic normality of OLS estimator.

Note: This assumption hard to interpret. Often we interpret a sufficient condition: the error term is serially uncorrelated and also is uncorrelated with the current and past regressors.

Note: S is a matrix of fourth moments

Note: This assumption also implies that the error terms are not serially correlated.

Add'l :

Finite fourth moments for Regressors: $\mathrm{E}[(x_{ik}x_{ij})^2]$ exists and is finite for all k,j = 1,2, …, K (For consistent estimation of S)

Conditional Homoskedasticity: $E(\varepsilon_i^2 \mid x_i) = \sigma^2 > 0$

---

[23] Since $\{g_i\}$ stationary, the matrix of cross moments does not depend on *i*. Also, we implicitly assume that all the cross moments exist and are finite.

[24] Comments about assumptions:

- iid is trivially ergodic stationary:
  If$\{y_i, \boldsymbol{x}_i\}$ is iid., i.e. we have random sample, then trivially satisfies ergodic stationarity.
  (Though trivial, this is an important special case of ergodic stationarity!)
- Predetermined vs. Strictly exogenous regressors:
  Predetermined regressors are not required to be strictly exogenous. This is a weaker assumption than in M1. Strict exogeneity implies that all the regressors are orthogonal to all the error terms (current, past, and future); on the other hand, predetermined regressors restricts only the **contemporaneous** relationship between the error term and the regressors.
- Rank Condition as no Multicollinearity in the limit:
  Since $\mathrm{E}(\boldsymbol{x}_i\,\boldsymbol{x}_i')$ is finite by 2.4, $\lim_n S_{XX} = \Sigma_{XX}$ with probability 1 by the ergodic theorem (where $S_{XX} \equiv \dfrac{1}{n}\sum_{i=1}^{n} x_i x_i'$). So, for n sufficiently large, S_{XX} is nonsingular (and equivalently rank($\mathbf{X}$)= K for sufficiently large n).
- S is a matrix of 4th moments:
  Since $\mathbf{g}_i = \boldsymbol{x}_i\ .\ (y_i - \boldsymbol{x}_i'\beta) = \boldsymbol{x}_i\ .\varepsilon_i$, we can rewrite $S = \mathrm{E}(\varepsilon_i^2 \boldsymbol{x}_i\boldsymbol{x}_i')$ → the (i, j)th element is $\mathrm{E}(\varepsilon_i^2 x_{ij}x_{ij}')$ which involves 4th moments. So, for consistent estimation of S we need an additional assumption that the 4th moments exist.
- A sufficient condition for $\mathbf{g}_i$ to be an m.d.s.:
  A sufficient condition that is easier to interpret is: $\mathrm{E}(e_i \mid e_{i-1}, e_{i-2}, \ldots, e_1, \mathbf{x}_i, \mathbf{x}_{i-1}, \ldots, \mathbf{x}_1) = 0$ → error term is serially uncorrelated and uncorrelated with the current and past regressors.
- By 2.2, if $\{y_i, \boldsymbol{x}_i\}$ stationary, then $\{\varepsilon_i\}$ also stationary → ε_i unconditionally homoskedastic in that $E(\varepsilon_i^2)$ does not depend on *i*. But this does not imply errors are conditionally homoskedastic, therefore we need an additional assumption 2.7.
  (e.g. $\varepsilon_I = \eta_i f(\mathbf{x}_i)$ where η_I ind $\mathbf{x}_i$ and $\mathrm{E}(\eta_i) = 0$. Then, $\mathrm{E}(\varepsilon_I^{\,2}) = \mathrm{E}(\eta_I^{\,2} f(\mathbf{x}_i)^2) = \mathrm{E}(\eta_i^2)\mathrm{E}(f(\mathbf{x}_i)^2)$ by independence → (does not depend on i),
  but $\mathrm{E}(\varepsilon_I^{\,2} \mid \mathbf{x}_i) = \mathrm{E}(\eta_I^{\,2} f(\mathbf{x}_i)^2 \mid \mathbf{x}_i) = \mathrm{E}(\eta_I^{\,2} \mid \mathbf{x}_i)\ f(\mathbf{x}_i)^2 = \mathrm{E}(\eta_I^{\,2})\ f(\mathbf{x}_i)^2$ by independence → depends on *i* since $\mathbf{x}_i$ varies across *i* !

[25] Data Generating Process (DGP) is the stochastic process / sequence of random variables that generated the finite sample ($\mathbf{Y}_{nx1}$, $\mathbf{X}_{nxd}$). Therefore, if we specify the DGP, the joint distribution of the finite sample ($\mathbf{Y}_{nx1}$, $\mathbf{X}_{nxd}$) can be determined!

In finite-sample theory, where the sample size is fixed and finite, we defined a model as a set of joint distributions of ($\mathbf{Y}_{nx1}$, $\mathbf{X}_{nxd}$). In large sample theory, a model is stated as a set of DGPs that satisfy a set of assumptions.

3. **Properties of the OLS Estimator of β and σ^2** : [26]

   A. Consistency of $\hat{b}$ *for* β : Under assumptions 2.1 – 2.4, $p\lim_{n\to\infty} \hat{b}_{kx1} = \beta_{kx1}$

   B. Asymptotic normality of $\hat{b}$: Under 2.1 – 2.5, $\sqrt{n}\left(\hat{b}-\beta\right)\xrightarrow{D} N(0, \Sigma_{XX}^{-1}\ S\ \Sigma_{XX}^{-1})$ *where* $A\text{var}(\hat{b}) = \Sigma_{XX}^{-1} S \Sigma_{XX}^{-1}$
   Recall: $E(\boldsymbol{x}_i \boldsymbol{x}_i') = \boldsymbol{\Sigma_{XX}}$ and $E(g_i g_i') = S \equiv A\text{var}(\bar{g}_{nx1})$

   C. Consistent Estimate of Avar($\hat{b}$): Suppose there is available a consistent estimator $\hat{S}$ of S_{kxk}. Then, under assumption 2.2, Avar($\hat{b}$) is consistently estimated by $\hat{A}\text{var}(\hat{b}) = S_{XX}^{-1}\ \hat{S}\ S_{XX}^{-1}$ *where* $S_{XX} = \frac{1}{n}\sum_{i=1}^{n} x_i x_i' = \frac{1}{n} X'X$

   D. Consistency of s^2 (estimation of variance of "true" error is consistent): Under assumptions 2.1 – 2.4,
   $s^2 \equiv \frac{e'e}{n-K} \xrightarrow{P} E(\varepsilon_i^2)$, *provided* $E(\varepsilon_i^2)$ *exists and finite* ($e_i = y_i - \boldsymbol{x_i}'\boldsymbol{b}_{OLS}$ = OLS residual for observation i)

   Note on Consistent Estimation of S: How do we obtain consistent estimator $\hat{S}$ of S_{kxk} from the sample $(\boldsymbol{y},\boldsymbol{X})$?

   - Use OLS Residuals for Errors:
     Since we don't observe $\boldsymbol{g_i g_{i'}} = \boldsymbol{x}_i .\varepsilon_i\ (\boldsymbol{x}_{i'} .\varepsilon_{i'})' = \varepsilon_i^2 \boldsymbol{x_i x_i}'$ (we don't observe true errors ε_i), we use a consistent estimator
     $\hat{\varepsilon}_i \equiv y_i - x_i'\hat{\beta}$ *for some consistent estimator* $\hat{\beta}$ *of* β

     Th (Consistent Estimation of S):
     Suppose the coefficient estimate $\hat{\beta}$ used for calculating the residual $\hat{\varepsilon}_i$ for $\hat{S}$ is consistent for β, and suppose $\mathbf{S} = E(\boldsymbol{g_i g_{i'}})$ exists and is finite. Then, under assumptions 2.1, .2, and 2.6, $\boxed{\hat{S} = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i x_i x_i'}$ is consistent for **S**.

     (Proof is similar to that for D)

4. Hypothesis Testing (Asymptotic distributions)
   (2.4 and 2.6)

   Implications under Conditional Homoskedasticity (2.6)

---

[26] Proofs:

A. $\hat{b} = (X'X)^{-1}X'Y = (X'X)^{-1}X'(X\beta+\varepsilon) = \beta + (X'X)^{-1}X'\varepsilon \Rightarrow \hat{b}-\beta = (X'X)^{-1}X'\varepsilon = \left[\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right]^{-1}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i\right] = S_{XX}^{-1}\bar{g}_n \Rightarrow \sqrt{n}\left(\hat{b}-\beta\right) = S_{XX}^{-1}\sqrt{n}\bar{g}_n$

$\frac{1}{n}\sum_{i=1}^{n} x_i x_i' = S_{XX} \xrightarrow{P} E(x_i x_i') \equiv \Sigma_{XX}$ *by Ergodic Theorem* (since $x_i x_i'$ *ergodic stationary by* 2.2) (*convergence is a.s. which implies in prob*)

$\Rightarrow S_{XX}^{-1} \xrightarrow{P} \Sigma_{XX}^{-1}$ *by CMT* (*we know* Σ_{XX}^{-1} *exists by* 2.4)

$\frac{1}{n}\sum_{i=1}^{n} x_i\varepsilon_i = \frac{1}{n}\sum_{i=1}^{n} g_i \xrightarrow{P} E(g_i) = 0$ *by Ergodic Theorem* since g_i *Ergodic Stationary* (*again, convergence is actually a.s.*)

$\therefore \hat{b}-\beta \xrightarrow{P} 0$ *by Slutsky*

B. $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} x_i\varepsilon_i = \sqrt{n}\bar{g}_n \xrightarrow{D} N(0,S)$ *by Stationary Ergodic CLT* $\Rightarrow \sqrt{n}\left(\hat{b}-\beta\right)\xrightarrow{D} N(0,\Sigma_{XX}^{-1}\ S\ \Sigma_{XX}^{-1})$ *by Slutsky*

C. From above, then $S_{XX}^{-1}\ \hat{S}\ S_{XX}^{-1} \xrightarrow{P} \Sigma_{XX}^{-1}\ S\ \Sigma_{XX}^{-1}$ *by Slutsky*

D. $s^2 = \frac{e'e}{n-K} = \frac{1}{n-K}\sum_{i=1}^{n} e_i^2 = \frac{n}{n-K}\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right)$

It suffices to show that $\frac{1}{n}\sum_{i=1}^{n} e_i^2 \xrightarrow{P} E\left(\varepsilon_i^2\right)$ since $n/n-K \xrightarrow{n\to\infty} 1$, *so by Slutsky* $s^2 \xrightarrow{P} E\left(\varepsilon_i^2\right)$

$e_i = y_i - x_i'\hat{b} = y_i - x_i'\hat{b} + x_i'\beta - x_i'\beta = \varepsilon_i + x_i'(\hat{b}-\beta) \Rightarrow e_i^2 = \varepsilon_i^2 + 2\varepsilon_i x_i'(\hat{b}-\beta) + \left[x_i'(\hat{b}-\beta)\right]^2$

$\Rightarrow \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + 2\varepsilon_i x_i'(\hat{b}-\beta) + \left[x_i'(\hat{b}-\beta)\right]^2 = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + 2(\hat{b}-\beta)\left(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i x_i'\right) + (\hat{b}-\beta)'\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)(\hat{b}-\beta) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 + 2(\hat{b}-\beta)'\bar{g}_n + (\hat{b}-\beta)'S_{XX}(\hat{b}-\beta)$

$2(\hat{b}-\beta)'\bar{g}_n \xrightarrow{P} 0$ *and* $(\hat{b}-\beta)'S_{XX}(\hat{b}-\beta) \xrightarrow{P} 0$ since $\hat{b}-\beta \xrightarrow{P} 0$ *and* $\bar{g}_n \xrightarrow{P} E(g_i) = 0$

5. Large sample properties of WLS estimator

<u>Note on Best Linear Predictor</u>: How do we interpret OLS if all we have is ergodic stationarity?

Motivation: Our theory about finite and large sample properties rely on the linearity assumption (as well as others). What if in reality none of the assumptions are satisfied (except ergodic stationarity – so for example if we only have iid sample) and we go ahead and apply OLS to the sample. What is it that we estimate?

OLS Interp.: OLS gives us the **best linear predictor, or the least squares projection** – i.e. the best way to linearly combine the explanatory variables to predict the dependent variable.

Th: Suppose we observe (y, **x**) and know its joint distribution. E(y | **x**) is the best predictor of y in that it minimizes MSE.
$E\left[(y - E(y \mid x))^2\right] \le E\left[(y - f(x))^2\right]$ *for any function* $f(x)$ [27] (assuming E(Y) and E(f(X)) finite)

Since, E(y | **x**) can be highly nonlinear, if we restrict the predictor to being a linear function of **x**, then we can show that the best predictor is OLS!

Proof: This is the same as our first derivation of the OLS estimator.
Suppose we want to find the "best" (i.e. minimize MSE) linear predictor/approximation of y based on **x**.
Thus, we find β^* that satisfies the orthogonality condition (i.e. the linear combination of x such that the expected distance is smallest (i.e. orthogonal term).
$E(\mathbf{x} \cdot (y - \mathbf{x}'\beta^*)) = 0 \rightarrow E(\mathbf{x}\mathbf{x}')\beta^* = E(\mathbf{x} \cdot y) \rightarrow \beta^* = [E(\mathbf{x}\mathbf{x}')]^{-1} E(\mathbf{x} \cdot y)$

Th: The least squares projection L(y | **x**) is the best linear predictor of y in that it minimizes
Pf:

Th:

OLS Consistently Estimates the Projection Coefficients

Now, suppose we have a sample of size n drawn from an ergodic stationary stochastic process $\{y_i, \mathbf{x}_i\}$ with the joint distribution $(y_i, \mathbf{x}_i)$ (which does not depend on *i* because of stationarity) identical to that of (y, **x**) in the above example → $E(\mathbf{x}_i\mathbf{x}_i') = E(\mathbf{x}\mathbf{x}')$

Then, $\left(\frac{1}{n}\sum_{i=1}^{n} x_i x_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i\right) = (X'X)^{-1}X'y$ consistently estimates β^*.

---

[27]
$E\left[(E(y \mid x) - f(x))^2\right] \ge 0 \Rightarrow E\left[E(y \mid x)^2 - 2E(y \mid x)f(x) + f(x)^2\right] \ge 0 \Rightarrow -2E\left[E(y \mid x)f(x) + f(x)\right] \ge -E\left[E(y \mid x)^2\right]$
$\Rightarrow -2E\left[yf(x)\right] + E(f(x)) \ge -2yE(y \mid x) + E(y \mid x)^2 \Rightarrow E(y^2) - 2E\left[yf(x)\right] + E(f(x)) \ge -2yE(y \mid x) + E(y \mid x)^2$
$\Rightarrow E\left[(y - f(x))^2\right] \ge E\left[(y - E(y \mid x))^2\right]$