Boston University Questrom School of Business

MF793 – Fall 2021

Eric Jacquier

ORDINARY LEAST SQUARES

Main differences with the previous note:

We assume the model is linear and …. we put our hats on!

A Regression OLS: simple regression

B Multiple regression: general model and OLS properties

C Statistical properties of OLS

Recommended Readings: Hansen**:** Ch 2, 3 if desired

Greene: Ch 2, 3 (ignore or skim 3.5.1, 3.6).

Only what corresponds to what we do

A. Regression and OLS, basic models

A1 Basic model

Results in previous note – on the *Conditional Expectation Function* – were theoretical.

- We first discussed the conditional expectation from a true but unknown model - as in the "true unknown mean".

- Then we discussed using a linear model to approximate the true unknown CEF.

  We proved what β should be so the linear model approximates the desired true model properties as best as possible.

- Result was theoretical: β a function of (true but unknown!) expectation of the data.

Now we have data, what do we do ?

- .. We talk about estimation! – equivalent of the sample mean for the conditional expectation

- Data: y_i, x_i. Model for the data is: $y_i = x_i \beta + \varepsilon_i$, ε_i: true unknown **noise**

- Given a candidate value b as an estimate of the unknown β, we have: $y_i = x_i b + e_i$

  e_i: residual for observation i, an *estimate* of the noise ε_i.

  e: the vector of residuals

- Criterion to find the best β:

  Sum of errors? silly

  Sum of absolute errors? No analytics but can be done

  A Sample estimate of the MSE: Residual Sum of Squares aka Sum of Squared Errors:

  $\sum_i e_i^2 = e'e = \sum_i (y_i - bx_i)^2$

  Minimize: Set derivative with respect to b equal to 0

  $\frac{\partial e'e}{\partial b} = 2b\sum_i x_i^2 - 2\sum_i x_i y_i = 0$

  Known as the **OLS normal equation:**

$$\boldsymbol{b\sum_i x_i^2 = \sum_i x_i y_i}$$

$$\boldsymbol{b = \frac{\sum_i x_i y_i}{\sum_i x_i^2}}$$

$\beta = \frac{E(xy)}{E(x^2)}$

  Is it a minimum? Check the second derivative

- We do not have the **true** mean squared error $E(y-bx)^2$.
  We computed its sample mean, the **sample mean** of squared errors, and minimized it along b.

- What is b if $x_i = 1$, $i = 1, \ldots, T$? **The sample mean is the simplest of all OLS estimates.**

- This is a toy model: no intercept is very unrealistic …. Let's get (a bit) more serious.

A2 First (somewhat) serious model: regression with intercept

- Linear model: $y_i = \alpha + x_i \beta + \varepsilon_i$, $E(\varepsilon_i \mid x) = 0$

- $E(\varepsilon_i \mid x) = 0$ .... could be wrong. Will need to check this.
  - Maybe the true model is not linear ! The true unknown CEF m_x is unknown.
  - Here we write a feasible linear model. It is known but likely incorrect.
  - Serious econometricians consider this notion of the existence of a true model, naïve at best. All models are wrong, some maybe useful.

- Find (a, b): $y_i = a + b x_i + e_i$, that minimize the sum of squared errors **SSE**

$$\min_{a,b} SSE \equiv \min_{a,b} \sum_i (y_i - a - bx_i)^2 \qquad [1]$$

$$\frac{\partial SSE}{\partial a} = 2\sum_i (y_i - a - bx_i)(-1) = 0$$

$\sum y_i = \overline{\sum a} - b \sum x_i$

$$\boldsymbol{\bar{y} = a + b\bar{x}}$$

Regression line goes through the sample mean point $(\bar{x}, \bar{y})$

- Now find b: Substitute a into the SSE in [1]

$$SSE = \sum_i (y_i - a - bx_i)^2 = \sum_i (y_i - (\bar{y} - b\bar{x}) - bx_i)^2$$

$$SSE = \sum_i (y_i - \bar{y})^2 + b^2 (x_i - \bar{x})^2 - 2b(y_i - \bar{y})(x_i - \bar{x})$$

optimize with respect to b:

$$\frac{\partial SSE}{\partial b} = 0 = 2b \sum (x_i - \bar{x})^2 - 2 \sum (y_i - \bar{y})(x_i - \bar{x})$$

$$\boldsymbol{b} = \frac{\widehat{\boldsymbol{Cov}}(\boldsymbol{x}, \boldsymbol{y})}{\widehat{\boldsymbol{Var}}(\boldsymbol{x})}$$

- **Fitted value**: $\hat{\boldsymbol{y}}_{\boldsymbol{i}} = \boldsymbol{a} + \boldsymbol{x}_{\boldsymbol{i}}\, \boldsymbol{b}$

- **Residual**: $\boldsymbol{e}_{\boldsymbol{i}} = y_i - a - x_i b = (y_i - \bar{y}) - b(x_i - \bar{x})$ **[2]**

"*deviation form*"

Recall $\bar{y} = a + b\bar{x}$

- Is the model well specified? Must check if it matches properties of the CEF

A3 Properties of OLS residuals

Recall LN 9, we found some obvious properties of the **correct** unknown model (CEF) and error. We hope that our … *likely wrong* … approximate linear model also has these properties

- Theory says $E(\varepsilon_i) = 0$ From [2]: $\mathbf{\Sigma_I e_i} = \sum(y_i - a - bx_i) = T\bar{y} - Ta - Tb\bar{x}$
  - Because of the intercept, the residuals have zero sample mean by construction. $= 0$
  - So … the regression with no intercept has a potential problem. Always use an intercept.

- Theory says $\mathrm{Cov}(\varepsilon , x) = 0$

Is the model well specified?

Noise must truly be unrelated to the predictor X ... or it's not a noise !

$$\sum(x_i - \bar{x})e_i = \sum(x_i - \bar{x})((y_i - \bar{y}) - b(x_i - \bar{x}))$$ **Prove it**

$$= \sum(x_i - \bar{x})(y_i - \bar{y}) - \sum b(x_i - \bar{x})^2 \quad = 0$$

- By construction, OLS residuals are orthogonal to the predictor x.
  .... even if the true noise is related to x !

This is a problem: We need other ways to check for model specification: graphical analysis.

Cov(ε,x) = 0 is an assumption of the model, $\widehat{\mathrm{Cov}}$(e,x) = 0 is always true for the OLS.

B (Now the serious) Multiple Regression Model: OLS Properties.

Know the difference between **multiple** and **multivariate** regression
Multiple: Several X variables to do a good job forecasting one Y variable
Multivariate: Forecasting several Y variables with the same several X variables

B1 OLS model in Matrix form

- $y_i = \beta_0 \mathbf{1} + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$, $i = 1, 2, \ldots, T$ **1**: a vector of ones
  $y_i = x_i' \beta + \varepsilon_i$, $i = 1, 2, \ldots, T$

  Intercept: $\mathbf{x_{0i} = 1}$, for $i = 1, 2, \ldots, T$

  Stack the observations:

$$\begin{aligned} y_1 &= x_1' \beta + \varepsilon_1 \\ y_2 &= x_2' \beta + \varepsilon_2 \\ &\ldots \\ y_T &= x_T' \beta + \varepsilon_T \end{aligned}$$

- Linear model in Matrix notation: $\underset{T\times 1}{Y} = \underset{T\times K}{X} \; \underset{K\times 1}{\beta} + \underset{T\times 1}{\varepsilon}$

  Intercept: is the first of the k+1 betas, corresponds to the first column of X, a column of ones.
  X: T x K data matrix
  ε: unobservable noise vector

- **Residual vector**: $\mathbf{e = Y - X b} \neq \varepsilon$! For a choice b for the unknown vector β
- **Sum of squares**: now written as the inner product $\mathbf{e'e} = (y - Xb)'(y - Xb)$

B2 OLS properties in matrix form, the HAT matrix, the M matrix

- Min e'e ⇔ Min (Y - Xβ)' (Y - Xβ)

  $\min_{\beta} \; (Y'Y \; - \; 2\,\beta' X' Y \; + \; \beta' X' X \beta)$

- Normal equation: $-2\,X'Y + 2\,X'X\,\beta = 0$ <- k x 1 vector of partial derivatives $\partial e'e/\partial\beta$

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

  Notation: we use **b** or $\hat{\beta}$ to denote the **estimator** of β

- Regression plane goes through the Data sample means: $\bar{Y} = \bar{x}'\hat{\beta}$ Prove it

  Look at the 1st row of Normal equation

- What is X'e at the OLS estimate? **X'e =** $X'(Y - X\hat{\beta})$ = 0 Prove it.

  $= X'Y - X'X(X'X)^{-1}X'Y = 0$

- X'X Sample Cross-product matrix of the **K** regressors (including the intercept)

  <u>Need to invert X'X</u> => X is kxk, it needs to be of full **rank K**

  None of the K variables can be written as a linear function of the others

  How many observations do we need? **T** $\geq$ **K**

- Is the solution in [1] a minimum? $\frac{\partial^2 e'e}{\partial\beta\partial\beta'} = X'X$ must be positive definite matrix (like a covariance matrix, which it is!)

- Fitted value is a projection of Y on the space of X: the **P matrix**: (aka the **HAT** matrix)

  $\hat{Y} = X\hat{\beta} = X(X'X)^{-1}X'Y = \mathbf{P}\,\mathbf{Y}$ dim(P) = TxT

  $P = X(X'X)^{-1}X'$ P matrix projects The T dimensions of Y onto the space of X.

  $= X$

  PX = X ... we are already in X !

  Symmetric: **P = P'**

  Projecting twice: **PP' = P'P = PP = P** Idempotent, already in X !

- Residual e = Y - X$\hat{\beta}$ and the **M matrix**

  e = Y - X$\hat{\beta}$ = Y – PY = **(I – P) Y** = **M Y**

  M = I - P M matrix projects Y onto the space of residuals

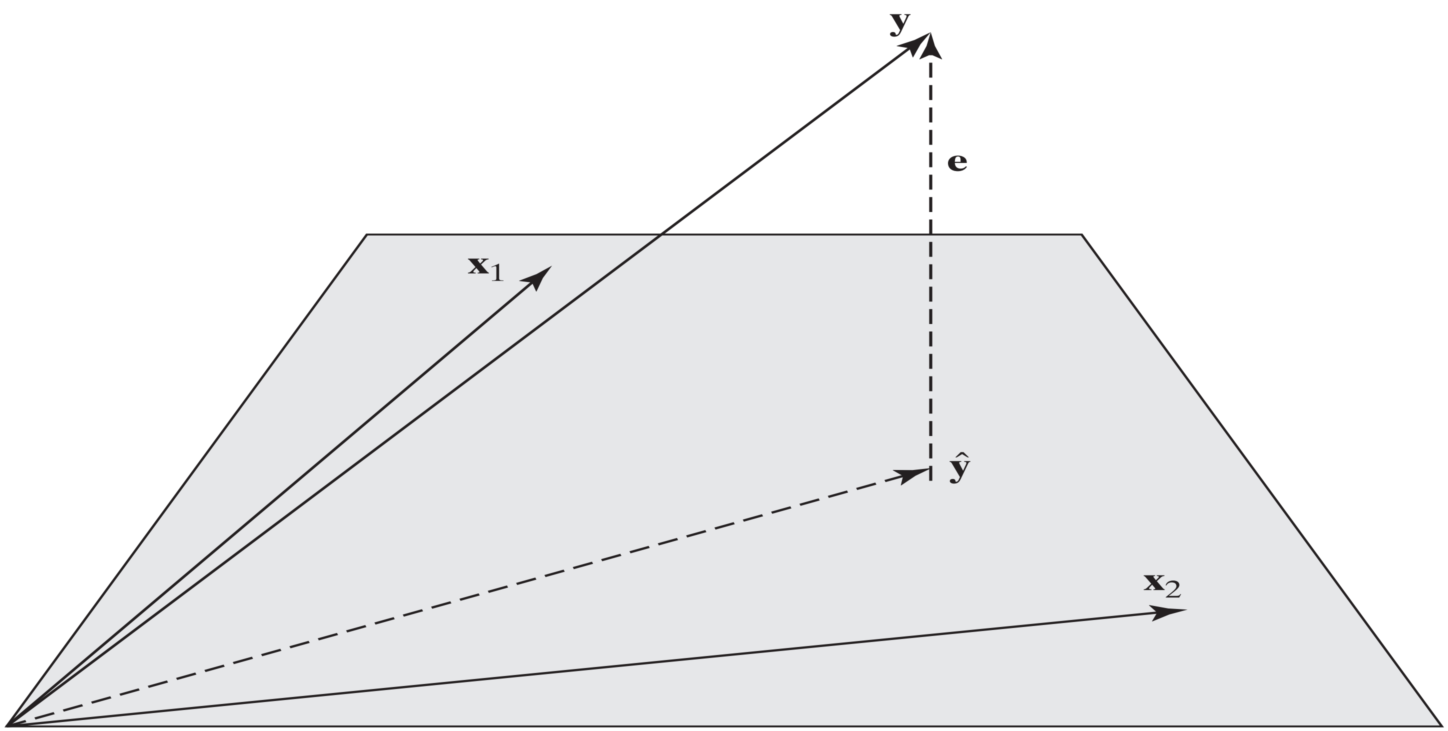
  M is symmetric, idempotent

- Recall residuals are orthogonal to X, consistent with:

  MX = (I – P) X = X – X = **0**

  MP = (I – P) P = 0

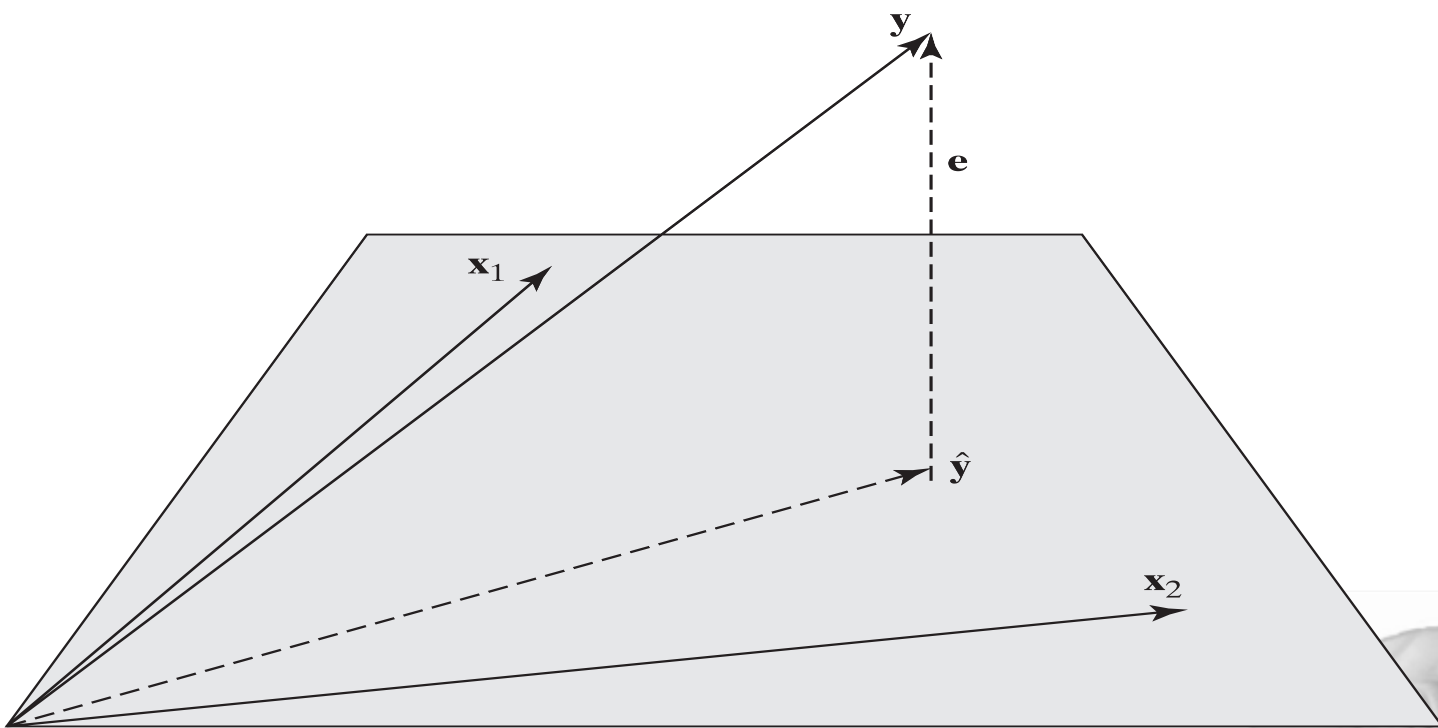
- $Y = \hat{Y} + e = PY + MY$



- ... The Variance decomposition …. It is just Pythagoras:

$$\mathbf{Y'Y = (\hat{Y} + e)'(\hat{Y} + e) = \hat{Y}'\hat{Y} + e'e} + 2\,Y'M'PY = \mathbf{\hat{Y}'\hat{Y} + e'e}$$

- $Y = \hat{Y} + e = PY + MY$



- ... The variance decomposition …. It is just Pythagoras:

$$Y'Y = (\hat{Y} + e)'(\hat{Y} + e) = \hat{Y}'\hat{Y} + e'e + 2\,Y'M'PY = \hat{Y}'\hat{Y} + e'e$$

Ωηατ δο ψου μεαν **ϑΥΣΤ** Πψτηαγορασ, ψου δωεεβ ?!

$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'Y$

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

B3 What is the purpose of the multiple regression ?

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

- **When X_1 and X_2 are not correlated, The estimates of β_1 , β_2 are equal to those from two separate regression Y on X_1 and Y on X_2** Prove it

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

- Heard on the street:

"If X_1 and X_2 are correlated it's a problem for the estimation of β_1 , β_2 because of multicollinearity"

- The purpose of the multiple regression **is exactly** to properly estimate β_1 and β_2 when X_1, X_2 are correlated, by taking into account their correlation.

- Result: $Y = X\beta + \varepsilon$, where $X = (X_1 \mid X_\varepsilon)$ $\hat{\beta}_\varepsilon$

  When X_1 and X_2 are correlated, the estimate of β_2 is equal to the OLS estimate of the regression of not (Y on X_2), but of (the residual of Y on X1) on (the residual of X2 on X1)

  No Proof

  $Y \perp X_1 \longrightarrow X_2 \perp X_1$

- We only have a problem when X_1 and X_2 are extremely highly correlated.
  As their correlation increases, the variance of b_1 and b_2 increases.
  If X_1 and X_2 are perfectly correlated, β_1 and β_2 are not separately identifiable.

  It's a silly situation easily avoided: Don't use quasi-perfectly correlated regressors!

- Warning: This "silly situation" is more likely to occur in situation with very large data sets and little subject matter knowledge, as *mechanical* non parametric models may use many Xs.

$y'y = \hat{y}'\hat{y} + e'e$ $\quad X\hat{\beta} = \hat{y}$

B4. Variance decomposition, R-square and Adjusted R-square

$T\times 3 \quad 3\times 1 \quad T\times 1$

- The mean-centered variance decomposition (similar but different from P. 11, no proof)

$$(Y - \mathbf{1}\bar{Y})'(Y - \mathbf{1}\bar{Y}) = (\hat{Y} - \mathbf{1}\bar{Y})'(\hat{Y} - \mathbf{1}\bar{Y}) + e'e \rightarrow \text{Small}$$

SST — **Total Sum Squares** | **SSR** — **Regression SS** | **SSE** — **SS Errors**

- **$R^2 = 1 - SSE / SST$** $\quad R^2$ is a measure of *goodness of fit.*

  Problem: R^2 increases automatically as X variables are added $\quad$ Why ?

- $0<R^2<1$ only if regression has an intercept. R^2 is not as useful if there is no intercept. No proof

- Very minor improvement: $\bar{R}^2 = 1 - \frac{SSE/_{T-K}}{SST/_{T-1}}$ **Adjusted R^2**

  K increases as one adds X variables: the adjusted R^2 does not necessarily increase.

  This penalty is very minor, one can show that: (no proof)

Adjusted $\bar{R}^2$ increases when adding a variable to the regression

if its squared t-statistic is higher than 1

- R^2 is a good measure of fit for one given model, but we need better measures for model comparison ! (such as Akaike or BIC, will see later)

C Statistical Properties of the OLS

- So far, we said nothing about the statistical properties of **ε**. .. We did not need to!

- Is $\hat{\beta}_{OLS}$ a good estimator?

  Unbiased: Do we get the true β *on average, in repeated samples* if we use OLS?

  Precise: How far is $\hat{\beta}_{OLS}$ from the true β ... *on average, in repeated samples*?

  Is $\hat{\beta}_{OLS}$ *best*?

  Distribution of $\hat{\beta}_{OLS}$?

  R**epeated sampling** => different ε each time.

- To answer these questions, we must make assumptions about the statistical properties of the noise ε

C1. Assumptions of the Linear Model

- Already have:

.. Linearity ! $y = X\beta + \varepsilon$, Y is linear in X

.. Full rank Data matrix X has full column rank K - None of the k variables is a linear combination of the others

.. Exogenous X variables: $E(\varepsilon \mid X) = 0$ then $Cov(X, \varepsilon) = 0$, $E(g(X)\,\varepsilon) = 0$. As in LN9.

- Add: Homoskedastic and non correlated errors:

  $E(\varepsilon\varepsilon' \mid X) = \sigma^2 I_T$ generalizes the i.i.d sample from the mean estimation problem

.. Often, one assumes X is non-stochastic. That is, the analysis is done *"given X"*.

.. $\varepsilon \sim N$? No need yet, let's bring it only when / if we need it.

- Rewrite the OLS estimator as a function of the true noise:

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

$$\hat{\beta}_{OLS} = \beta + (X'X)^{-1}\ X'\varepsilon \qquad [1]$$

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

C2 Unbiasedness

LN 9 Assumptions of the model: $E(\varepsilon|X) = 0 \quad \Rightarrow E(\, g(X)\, \varepsilon\,) = 0$ [2]

Then [2] => $\mathbf{E(\hat{\beta}_{OLS})} = \beta + E[\,(X'X)^{-1} X' \varepsilon\,]$

$= \beta + 0 \quad \mathbf{= \beta}$ [3]

- Questions:

  1. Did we need to know the *distribution* of ε ? No

  2. What is the only assumption we needed ?

  3. When would we not have : $E(\varepsilon) = 0$?

  4. When would we not have: $E(\, \varepsilon \mid X) = 0$?

  Omitted some X_2 which is correlated with X

C3 Variance of $\hat{\beta}_{OLS}$

- $V(\hat{\beta}_{OLS}) = E[\,(\hat{\beta}_{OLS} - E(\hat{\beta}_{OLS}))\,(\hat{\beta}_{OLS} - E(\hat{\beta}_{OLS}))'\,]$ <- definition

  $= E[\,(\hat{\beta}_{OLS} - \beta)\,(\hat{\beta}_{OLS} - \beta)'\,]$ <- unbiased $E(\hat{\beta}_{OLS}) = \beta$

  $= E[\,(X'X)^{-1} X' \varepsilon\,(\,(X'X)^{-1} X' \varepsilon\,)'\,]$ use assumption [2] again

  $= (X'X)^{-1} X'\ \mathbf{E(\varepsilon\varepsilon')}\ X\,(X'X)^{-1}$

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

  New critical assumption: Noise is iid: $\mathbf{E(\varepsilon\varepsilon') = \sigma^2 I_T}$ $= \begin{pmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{pmatrix}$

  $T \times T$

  $\mathbf{V(\hat{\beta}_{OLS})} = (X'X)^{-1} X'\ \boldsymbol{\sigma^2 I}\ X\,(X'X)^{-1} = \boldsymbol{\sigma^2 (X'X)^{-1}}$ [4]

  Intuition? **Do it** for one regressor

  Note: This proof was " | X)", it can be generalized by iterated expectation. (No proof)

- **Gauss-Markov result**: $\hat{\beta}_{OLS}$ **is the Best Linear Unbiased Estimator** (BLUE) of β

  Linear estimator? Here, linear refers to linear in the noises; i.e., of the form K ε

  No proof

C4. Distribution of $\hat{\beta}_{OLS}$

Recall: $\hat{\beta}_{OLS} = \beta + \underbrace{(X'X)^{-1} X'} \varepsilon = \beta + \underbrace{\begin{bmatrix} \cdots \\ \vdots \\ \cdots \end{bmatrix}}_{K \times T} \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_T \end{bmatrix}$

$\hat{\beta}_{OLS}$ is a linear combinations of ε, the weights are the Xs.

Two possibilities

1. ε normally distributed => $\hat{\beta}_{OLS}$ exactly normally distributed

2. Don't know ε's distribution:
   Large sample => $\hat{\beta}_{OLS}$ **approximately** normally distributed by CLT

C5. Estimating σ

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

M is symmetric

OLS has nothing to say about the estimation of $\sigma^2 = E(\varepsilon^2)$

$e'e = \varepsilon' M' M \varepsilon$
$= \varepsilon' M \varepsilon$

We now have the residual **e,** an estimate of ε: **e** = My = M(Xβ+ε) = **Mε**

Crazy question: if I know e = Mε, why can't I back out the true ε from e? Answer ?

E(e'e|X) = E(ε'Mε | X) = E(tr(ε'Mε) | X) = E(tr(M εε') | X) by properties of the trace

= tr(M E(εε' | X)) = tr(M $\sigma^2 I_T$) = $\boldsymbol{\sigma^2}$ tr(M)

tr(M) = tr(I_T – P) = T – tr(X(X'X)$^{-1}$X')
= T – tr((X'X)$^{-1}$ X' X) = T – tr(I_K)
= **T – K**

E(e'e|X) = (T-k) σ². An unbiased estimator of σ^2 is:

$$s^2 = \frac{e'e}{T-K}$$

s^2 is also called the MSE of the regression, if ε is normal we can show that $\mathbf{(T-k)s^2/\sigma^2 \sim \chi^2(T-k)}$

Intuition: e'e is a sum of T squared normals but only T-k of them are independent.
Proof later, in χ2 results, LN 11

C7 Remaining Issues

- Estimate of the variance of $\hat{\beta}_{OLS}$: **Var($\hat{\beta}_{OLS}$) = $s^2(X'X)^{-1}$**

- Large Sample behavior of $\hat{\beta}_{OLS}$ with random X

  Requires X'X and X'y to be *well behaved* as the sample gets large, i.e.,:

$$\lim_{T\to\infty}\left(\frac{1}{T}X'X\right) = V(X) \qquad \lim_{T\to\infty}\left(\frac{1}{T}X'y\right) = Cov(X,y)$$

  Then: $plim\hat{\beta} = plim\left(\frac{1}{T}X'X\right)^{-1} plim\left(\frac{1}{T}X'y\right) = V(X)^{-1}\ Cov(X,y)$

  → before $plim\,\hat{\beta}$ $plim\,(X'X)^{-1}X'y = plim\frac{(X'X)^{-1}}{T}\ plim(X'y)$

  $\hat{\beta}_{OLS}$ is a consistent estimator of the true β

- Even If the noise ε is not normal, $\hat{\beta}_{OLS}$ is asymptotically normal by CLT with large sample.

- s^2 is a consistent estimator of σ^2 under reasonable circumstances