ECON207 Session 5

OLS using Matrix Algebra

Anthony Tay

This Version: 31 Jul 2024

Session 5

- A Bit More Matrix Algebra
- OLS estimation of the Simple Linear Regression Model
  - Geometric Intuition: Projections
  - Statistic Properties
    - Completion of Basic Theory
    - Some Applications

Session 5.1

Session 5.1 A Bit More Matrix Algebra

- Differentiation of Matrix Forms
- Application to Optimization

Differentiation of Matrix Forms

Let $f(x)$ be some function of the $n \times 1$ vector x

e.g., $f(x) = x^{\mathrm{T}}Ax$ where A is some $n \times n$ matrix of constants

This is just a multivariable function $f(x_1, \dots, x_n)$

We *define*

$$\frac{df}{dx} = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} \\ \dfrac{\partial f}{\partial x_2} \\ \vdots \\ \dfrac{\partial f}{\partial x_n} \end{bmatrix} \quad \text{or} \quad \frac{df}{dx^{\mathrm{T}}} = \begin{bmatrix} \dfrac{\partial f}{\partial x_1} & \dfrac{\partial f}{\partial x_2} & \cdots & \dfrac{\partial f}{\partial x_n} \end{bmatrix}$$

Differentiation of Matrix Forms

Example: If $y = x^{\mathrm{T}}Ax$ where A is $n \times n$ and x is $n \times 1$, then

$$\frac{dy}{dx} = \frac{d}{dx}(x^{\mathrm{T}}Ax) = (A + A^{\mathrm{T}})x$$

$n = 3$ example:

$$x^T Ax = x_1^2 a_{11} + x_2^2 a_{22} + x_3^2 a_{33} + x_1 x_2(a_{12} + a_{21}) + x_1 x_3(a_{13} + a_{31}) + x_2 x_3(a_{23} + a_{32})$$

$$\left.\begin{aligned} \frac{dx^T Ax}{dx_1} &= 2x_1 a_{11} + x_2(a_{12} + a_{21}) + x_3(a_{13} + a_{31}) \\ \frac{dx^T Ax}{dx_2} &= x_1(a_{12} + a_{21}) + 2x_2 a_{22} + x_3(a_{23} + a_{32}) \\ \frac{dx^T Ax}{dx_3} &= x_1(a_{13} + a_{31}) + x_2(a_{23} + a_{32}) + 2x_3 a_{33} \end{aligned}\right\} \Rightarrow \quad \frac{dx^{\mathrm{T}}Ax}{dx} = (A + A^{\mathrm{T}})x$$

If A is symmetric, then $\dfrac{dx^{\mathrm{T}}Ax}{dx} = (A + A^{\mathrm{T}})x = 2Ax$ (compare $f(x) = ax^2 \Rightarrow f'(x) = 2ax$)

Differentiation of Matrix Forms

Suppose x is $n \times 1$ and $y = f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}$ (i.e., y is a vector of m different functions)

Then we define

$$\frac{dy}{dx^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \quad \text{or} \quad \frac{dy^{\mathrm{T}}}{dx} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Differentiation of Matrix Forms

E.g.: if $y = f(x) = Ax$ where $A = (a_{ij})_{m\times n}$ and $x^{\mathrm{T}} = [x_1 \quad x_2 \quad \ldots \quad x_n]$, then

$$\frac{dy}{dx^{\mathrm{T}}} = \frac{d(Ax)}{dx^{\mathrm{T}}} = A \quad \text{or} \quad \frac{dy^{\mathrm{T}}}{dx} = \frac{d(x^{\mathrm{T}}A^{\mathrm{T}})}{dx} = A^{\mathrm{T}}$$

Matrix analogue of the univariate differentiation rule $f(x) = ax \Rightarrow f'(x) = a$

Proof:

- The product Ax is an $m \times 1$ vector whose $i\text{-}th$ element is $\sum_{k=1}^{n} a_{ik}x_k$
- Therefore the $(i,j)th$ element of dy/dx^{T} is $(\partial/\partial x_j)\sum_{k=1}^{n} a_{ik}x_k = a_{ij}$
- This says that $dy/dx^{\mathrm{T}} = A$

Differentiation of Matrix Forms

E.g.: If $y = f(x)$ is a scalar-valued function of an $n \times 1$ vector of variables x, then

$$\frac{d}{dx^{\mathrm{T}}}\left(\frac{dy}{dx}\right) = \frac{d}{dx^{\mathrm{T}}}\begin{bmatrix} \dfrac{\partial y}{\partial x_1} \\ \dfrac{\partial y}{\partial x_2} \\ \vdots \\ \dfrac{\partial y}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial^2 y}{\partial x_1^2} & \dfrac{\partial^2 y}{\partial x_1 \partial x_2} & \cdots & \dfrac{\partial^2 y}{\partial x_1 \partial x_n} \\ \dfrac{\partial^2 y}{\partial x_2 \partial x_1} & \dfrac{\partial^2 y}{\partial x_2^2} & \cdots & \dfrac{\partial^2 y}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \dfrac{\partial^2 y}{\partial x_n \partial x_1} & \dfrac{\partial^2 y}{\partial x_n \partial x_2} & \cdots & \dfrac{\partial^2 y}{\partial x_n^2} \end{bmatrix}$$

- This is the Hessian matrix of $y = f(x)$. We usually write $\dfrac{d}{dx^{\mathrm{T}}}\left(\dfrac{dy}{dx}\right)$ as $\dfrac{d^2y}{dxdx^{\mathrm{T}}}$

Matrix Differentiation (Optimization)

Recall when finding min/max of $f(x, y)$:

- Stationary point: (x^*, y^*) such that $f'_x(x^*, y^*) = 0$ and $f'_y(x^*, y^*) = 0$
- If $f(x, y)$ is concave; stationary point is maximum point
- If $v_1^2 \dfrac{\partial^2 f(x,y)}{\partial x^2} + 2v_1 v_2 \dfrac{\partial^2 f(x,y)}{\partial x \partial y} + v_2^2 \dfrac{\partial^2 f(x,y)}{\partial y^2} < 0$ for all v_1, v_2 not both zero, then $f(x, y)$ is concave
- If $f(x, y)$ is convex; stationary point is minimum point
- If $v_1^2 \dfrac{\partial^2 f(x,y)}{\partial x^2} + 2v_1 v_2 \dfrac{\partial^2 f(x,y)}{\partial x \partial y} + v_2^2 \dfrac{\partial^2 f(x,y)}{\partial y^2} > 0$ for all v_1, v_2 not both zero, then $f(x, y)$ is convex

Matrix Differentiation (Optimization)

The expression

$$v_1^2 \frac{\partial^2 f(x,y)}{\partial x^2} + 2v_1 v_2 \frac{\partial^2 f(x,y)}{\partial x \partial y} + v_2^2 \frac{\partial^2 f(x,y)}{\partial y^2}$$

is the quadratic form involving the Hessian

$$\begin{bmatrix} v_1 & v_2 \end{bmatrix} \begin{bmatrix} \dfrac{\partial^2 f(x,y)}{\partial x^2} & \dfrac{\partial^2 f(x,y)}{\partial x \partial y} \\ \dfrac{\partial^2 f(x,y)}{\partial x \partial y} & \dfrac{\partial^2 f(x,y)}{\partial y^2} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad \text{i.e.,} \quad v^{\mathrm{T}} H_f v$$

- If Hessian is neg. def., i.e., if $v^{\mathrm{T}} H_f v < 0$ for all $v \neq 0_{2\times 1}$, then $f(x, y)$ is concave
- If Hessian is pos. def., i.e., if $v^{\mathrm{T}} H_f v > 0$ for all $v \neq 0_{2\times 1}$, then $f(x, y)$ is convex

Matrix Differentiation (Optimization)

Extends to functions of many variables $f(x)$ where x is $n \times 1$

- f concave implies stationary pt is max, f convex implies stationary pt is min
- Stationary point: x^* such that $\dfrac{df(x^*)}{dx} = 0$
  - If $f(x)$ is convex; stationary point is minimum point
  - If $f(x)$ is concave; stationary point is maximum point
- If Hessian is positive definite, then $f(x)$ is convex
- If Hessian is negative definite, then $f(x)$ is concave

How to check for pos./neg. def'ness? We'll consider only a simple case in the next section

Session 5.2

Session 5.2 OLS Estimation of the MLR

- Setting Up the Multiple Linear Regression Model in Matrix Form
- OLS Estimation of the MLR
  - Estimator Formulas
  - Estimator Variance Formulas
  - R^2 and Adjusted-R^2 (same as before)

Setting Up The MLR Model

Population satisfies $E(Y \mid X_1, X_2, \ldots, X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

i.e.,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \epsilon \,,\; E(\epsilon \mid X_1, \ldots, X_k) = 0$$

$X_1, \ldots, X_k$ refer to k different regressors, not k obs. of a variable X

Representative iid sample from the population:

$$\{Y_i, X_{i1}, X_{i2}, \ldots, X_{ik}\}_{i=1}^{n}$$

No perfect collinearity in the sample:

$$c_0 + c_1 X_{i1} + \cdots + c_k X_{ik} = 0 \text{ for all } i = 1, \ldots, n \iff c_0 = c_1 = \cdots = c_k = 0$$

Setting Up The MLR Model

Sample satisfies

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i \,,$$

$$E(\epsilon_i \mid X_{11}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1k}, \ldots, X_{nk}) = 0$$

$$E(\epsilon_i^2 \mid X_{11}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1k}, \ldots, X_{nk}) = \sigma_i^2$$

$$E(\epsilon_i \epsilon_j \mid X_{11}, \ldots, X_{n1}, X_{12}, \ldots, X_{n2}, \ldots, X_{1k}, \ldots, X_{nk}) = 0$$

for all $i, j = 1, \ldots, n$, $i \neq j$

More compactly expressed in matrix form

Setting Up The MLR Model

The equations

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \epsilon_i \quad \text{for all} \quad i = 1, \dots, n$$

can be written

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\text{i.e.,} \quad y = X\beta + \epsilon$$

X_{ij} is ith sample of the jth regressor

Setting Up The MLR Model

The assumption

$$c_0 + c_1 X_{i1} + \cdots + c_k X_{ik} = 0 \;\; \text{for all} \;\; i = 1, \dots, n \quad \Longleftrightarrow \quad c_0 = c_1 = \cdots = c_k = 0$$

can be written $Xc = 0_{n\times 1} \Leftrightarrow c = 0_{n\times 1}$ or $Xc \neq 0_{n\times 1} \Leftrightarrow c \neq 0_{n\times 1}$

The assumption

$$E(\epsilon_i \mid X_{11}, \dots, X_{n1}, X_{12}, \dots, X_{n2}, \dots, X_{1k}, \dots, X_{nk}) = 0 \;\; \text{for} \;\; i = 1, \dots, n$$

can be written

$$\begin{bmatrix} E(\epsilon_1 \mid X) \\ E(\epsilon_2 \mid X) \\ \vdots \\ E(\epsilon_n \mid X) \end{bmatrix} = 0_{n\times 1} \quad \Longleftrightarrow \quad E(\epsilon \mid X) = 0_{n\times 1}$$

Setting Up The MLR Model

The assumptions

$$E(\epsilon_i^2 \mid X_{11}, \dots, X_{n1}, X_{12}, \dots, X_{n2}, \dots, X_{1k}, \dots, X_{nk}) = \sigma_i^2 \quad \text{for} \quad i = 1, \dots, n,$$

$$E(\epsilon_i \epsilon_j \mid X_{11}, \dots, X_{n1}, X_{12}, \dots, X_{n2}, \dots, X_{1k}, \dots, X_{nk}) = 0 \quad \text{for} \quad i, j = 1, \dots, n \ , i \neq j$$

can be written as

$$\begin{bmatrix} E(\epsilon_1^2 \mid X) & E(\epsilon_1\epsilon_2 \mid X) & \cdots & E(\epsilon_1\epsilon_n \mid X) \\ E(\epsilon_2\epsilon_1 \mid X) & E(\epsilon_2^2 \mid X) & \cdots & E(\epsilon_2\epsilon_n \mid X) \\ \vdots & \vdots & \ddots & \vdots \\ E(\epsilon_n\epsilon_1 \mid X) & E(\epsilon_n\epsilon_2 \mid X) & \cdots & E(\epsilon_n^2 \mid X) \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix}$$

or

$$Var\epsilon \mid X = E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \text{diag}(\sigma_1^2, \sigma_2^2, \dots, \sigma_n^2)$$

Setting Up The MLR Model

If the noise terms are homoskedastic, i.e.,

$$E(\epsilon_i^2 \mid X_{11}, \dots, X_{n1}, X_{12}, \dots, X_{n2}, \dots, X_{1k}, \dots, X_{nk}) = \sigma^2 \quad \text{for} \quad i = 1, \dots, n,$$

then we have

$$E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

which we will assume for now

OLS Estimation of MLR

In matrix form, the MLR is

$$Y = X\beta + \epsilon\,,\ \ E(\epsilon \mid X) = 0\,,\ \ E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$$

For any potential estimator $\widehat{\beta}$, define

- Fitted values: $\widehat{y} = X\widehat{\beta}$
- Residuals: $\widehat{\epsilon} = y - \widehat{y} = y - X\widehat{\beta}$
- Sum of Squared Residuals $SSR(\widehat{\beta})$: $\sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \widehat{\epsilon}^{\mathrm{T}}\widehat{\epsilon}$

Agenda | A Bit More Matrix Algebra | OLS Estimation of MLR | Geometric Persp. | OLS Estimator Properties | Example | Asymptotics | Roadmap

OLS Estimation of MLR

We have

$$\begin{aligned} SSR(\widehat{\beta}) = \widehat{\epsilon}^{\mathrm{T}}\widehat{\epsilon} &= (y - X\widehat{\beta})^{\mathrm{T}}(y - X\widehat{\beta}) \\ &= (y^{\mathrm{T}} - \widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}})(y - X\widehat{\beta}) \\ &= y^{\mathrm{T}}y - \widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y - y^{\mathrm{T}}X\widehat{\beta} + \widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\widehat{\beta} \\ &= y^{\mathrm{T}}y - 2\widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y + \widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\widehat{\beta} \end{aligned}$$

OLS:

$$\begin{aligned} \widehat{\beta}^{ols} &= \arg\min_{\widehat{\beta}} SSR(\widehat{\beta}) \\ &= \arg\min_{\widehat{\beta}} y^{\mathrm{T}}y - 2\widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}}y + \widehat{\beta}^{\mathrm{T}}X^{\mathrm{T}}X\widehat{\beta} \end{aligned}$$

OLS Estimation of MLR

First and second derivatives are

$$\frac{dSSR(\hat{\beta})}{d\beta} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\hat{\beta} \quad \text{and} \quad \frac{d^2SSR(\hat{\beta})}{d\beta\, d\beta^{\mathrm{T}}} = 2X^{\mathrm{T}}X$$

FOC: $\left.\dfrac{dSSR(\hat{\beta})}{d\beta}\right|_{\hat{\beta}^{ols}} = -2X^{\mathrm{T}}y + 2X^{\mathrm{T}}X\hat{\beta}^{ols} = 0_{(k+1)\times 1}$ which implies

$$\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$$

- $Xc \neq 0 \Leftrightarrow c \neq 0$ means that $X^{\mathrm{T}}X$ in non-singular, and
- $c^{\mathrm{T}}X^{\mathrm{T}}Xc = (Xc)^{\mathrm{T}}(Xc) > 0$ (Hessian is pos. def.) so $\hat{\beta}^{ols}$ minimizes $SSR(\hat{\beta})$

OLS Estimation of MLR

Another way of expressing $\hat{\beta}^{ols}$

Writing

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} = \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix}$$

which emphasizes observations.

Note X_{i*} is $1 \times (k+1)$ vector comprising ith observation of all variables (including '1' for the intercept term)

OLS Estimation of MLR

Then

$$\begin{aligned}
\hat{\beta}^{ols} &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \\
&= \left\{ \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \cdots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \right\}^{-1} \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \cdots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \\
&= \underbrace{\left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*} \right)^{-1}}_{\text{sum of } k\times k \text{ matrices}} \underbrace{\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} Y_i}_{\text{sum of } k\times 1 \text{ vectors}}
\end{aligned}$$

OLS Estimation of MLR (Quick Summary)

- MLR: $Y = X\beta + \epsilon$, $E(\epsilon \mid X) = 0$, $Var(\epsilon \mid X) = E(\epsilon\epsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$
- OLS Estimator for β: $\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$
- OLS Fitted values:

$$\hat{y}^{ols} = X\hat{\beta}^{ols} = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = Py \quad \text{where} \quad P = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$$

- OLS Residuals:

$$\begin{aligned}
\hat{\epsilon}^{ols} &= y - \hat{y}^{ols} = y - X\hat{\beta}^{ols} = y - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y \\
&= (I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})y = My \quad \text{where} \quad M = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}
\end{aligned}$$

Where helpful to do so, I will place "ols" marker in subscript instead of superscript

OLS Estimation of MLR (More on the FOC)

The FOC can be written as: $X^{\mathrm{T}}\hat{\epsilon}^{ols} = 0_{(k+1)\times 1}$

$$\text{Since} \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{bmatrix} = \begin{bmatrix} i_n & X_{*1} & X_{*2} & \dots & X_{*k} \end{bmatrix}$$

$$\text{the FOC says:} \quad X^{\mathrm{T}}\hat{\epsilon}_{ols} = \begin{bmatrix} i_n^{\mathrm{T}} \\ X_{*1}^{\mathrm{T}} \\ X_{*2}^{\mathrm{T}} \\ \vdots \\ X_{*k}^{\mathrm{T}} \end{bmatrix} \hat{\epsilon}^{ols} = \begin{bmatrix} i_n^{\mathrm{T}}\hat{\epsilon}^{ols} \\ X_{*1}^{\mathrm{T}}\hat{\epsilon}^{ols} \\ X_{*2}^{\mathrm{T}}\hat{\epsilon}^{ols} \\ \vdots \\ X_{*k}^{\mathrm{T}}\hat{\epsilon}^{ols} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \hat{\epsilon}_i^{ols} \\ \sum_{i=1}^{n} X_{i1}\hat{\epsilon}_i^{ols} \\ \sum_{i=1}^{n} X_{i2}\hat{\epsilon}_i^{ols} \\ \vdots \\ \sum_{i=1}^{n} X_{ik}\hat{\epsilon}_i^{ols} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

OLS Estimation of MLR (More on the FOC)

The first line $\sum_{i=1}^{n} \hat{\epsilon}_i^{ols} = 0$ means that $\overline{\hat{\epsilon}^{ols}} = 0$ so

$$\sum_{i=1}^{n} X_{ij}\hat{\epsilon}_i^{ols} = 0 \iff \text{smp. cov.}(X_{ij}, \hat{\epsilon}_i^{ols}) = 0$$

for each variable X_j, $j = 1, \dots, k$

OLS estimator $\hat{\beta}^{ols}$ makes OLS residuals

- mean zero and
- uncorrelated with each regressor

Goodness of Fit

$SST = SSE + SSR$ decomposition continues to hold in the general MLR case

Let $M_0 = I_n - (1/n)i_n i_n^{\mathrm{T}}$ which is symmetric and idempotent

- $M_0^{\mathrm{T}} = M_0$ and $M_0 M_0 = M_0$

We have

- $M_0 y = \begin{bmatrix} Y_1 - \overline{Y} \\ Y_2 - \overline{Y} \\ \vdots \\ Y_n - \overline{Y} \end{bmatrix}$ and $y^{\mathrm{T}} M_0^{\mathrm{T}} M_0 y = y^{\mathrm{T}} M_0 y = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$
- $M_0 \hat{\epsilon}^{ols} = \hat{\epsilon}^{ols}$

Agenda | A Bit More Matrix Algebra | OLS Estimation of MLR | Geometric Persp. | OLS Estimator Properties | Example | Asymptotics | Roadmap

Goodness of Fit

Therefore

$$y = \hat{y}_{ols} + \hat{\epsilon}_{ols}$$

$$M_0 y = M_0 \hat{y}_{ols} + \hat{\epsilon}_{ols}$$

$$y^{\mathrm{T}} M_0^{\mathrm{T}} M_0 y = (M_0 \hat{y}_{ols} + \hat{\epsilon}_{ols})^{\mathrm{T}} (M_0 \hat{y}_{ols} + \hat{\epsilon}_{ols})$$

$$y^{\mathrm{T}} M_0 y = \hat{y}_{ols}^{\mathrm{T}} M_0 \hat{y}_{ols} + \hat{\epsilon}_{ols}^{\mathrm{T}} \hat{y}_{ols} + \hat{y}_{ols}^{\mathrm{T}} \hat{\epsilon}_{ols} + \hat{\epsilon}_{ols}^{\mathrm{T}} \hat{\epsilon}_{ols}$$

$$y^{\mathrm{T}} M_0 y = \hat{y}_{ols}^{\mathrm{T}} M_0 \hat{y}_{ols} + \hat{\epsilon}_{ols}^{\mathrm{T}} \hat{\epsilon}_{ols}$$

$$SST = SSE + SSR$$

Goodness of Fit

We can use this to define the goodness-of-fit measure:

$$R^2 = 1 - \frac{SSR}{SST}$$

and "Adjusted R^2"

$$\text{Adj.-}R^2 = 1 - \frac{\frac{1}{n-k-1}SSR}{\frac{1}{n-1}SST} = 1 - \frac{SSR}{SST}\frac{n-1}{n-k-1}$$

where k is no. of slope coefficients

Session 5.3

Session 5.3 Geometric Perspectives on OLS Estimation of MLR

Before discussing properties, we take a digression and consider OLS estimation of the MLR from a geometric perspective

- A vector space view of OLS estimation
  - Builds intuition
  - Help you understand some of the language used in OLS discussions

Geometric Perspective

Consider one regressor and $n = 3$, i.e.,

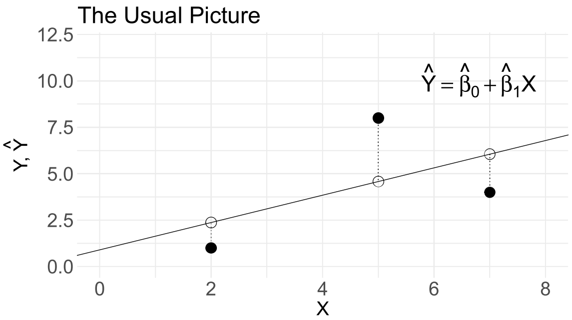
$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 & X_{11} \\ 1 & X_{21} \\ 1 & X_{31} \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} + \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \hat{\epsilon}_3 \end{bmatrix}$$

where

$$(X_{i1}, Y_i) = (2, 1), (5, 8), (7, 4)$$

OLS: find $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize

$$\hat{\epsilon}_1^2 + \hat{\epsilon}_2^2 + \hat{\epsilon}_3^2$$



Geometric Perspective

We consider now a different perspective

Write regression as $y = i_3\hat{\beta}_0 + X_{*1}\hat{\beta}_1 + \hat{\epsilon}$, i.e.,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \hat{\beta}_0 + \begin{bmatrix} X_{11} \\ X_{21} \\ X_{31} \end{bmatrix} \hat{\beta}_1 + \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \hat{\epsilon}_3 \end{bmatrix}$$

For our data,

$$\begin{bmatrix} 1 \\ 8 \\ 4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \hat{\beta}_0 + \begin{bmatrix} 2 \\ 5 \\ 7 \end{bmatrix} \hat{\beta}_1 + \begin{bmatrix} \hat{\epsilon}_1 \\ \hat{\epsilon}_2 \\ \hat{\epsilon}_3 \end{bmatrix}$$

The vectors $\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}^{\mathrm{T}}$ and $\begin{bmatrix} 2 & 5 & 7 \end{bmatrix}^{\mathrm{T}}$ are vectors in $\mathbb{R}^3$

Geometric Perspective

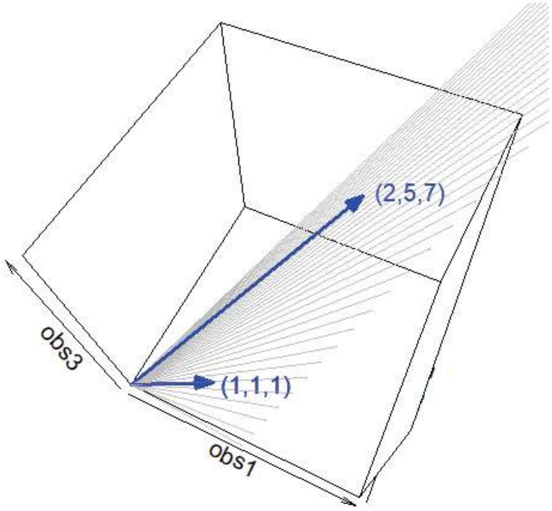
The vectors $\begin{bmatrix}1 & 1 & 1\end{bmatrix}^{\mathrm{T}}$ and $\begin{bmatrix}2 & 5 & 7\end{bmatrix}^{\mathrm{T}}$ are vectors in $\mathbb{R}^3$ (drawn in blue)

The set of all vectors

$$\begin{bmatrix}1\\1\\1\end{bmatrix}\hat{\beta}_0 + \begin{bmatrix}2\\5\\7\end{bmatrix}\hat{\beta}_1$$

(using all possible values of $\hat{\beta}_0$ and $\hat{\beta}_1$) forms a plane in $\mathbb{R}^3$

Part of plane shown in grey



Geometric Perspective

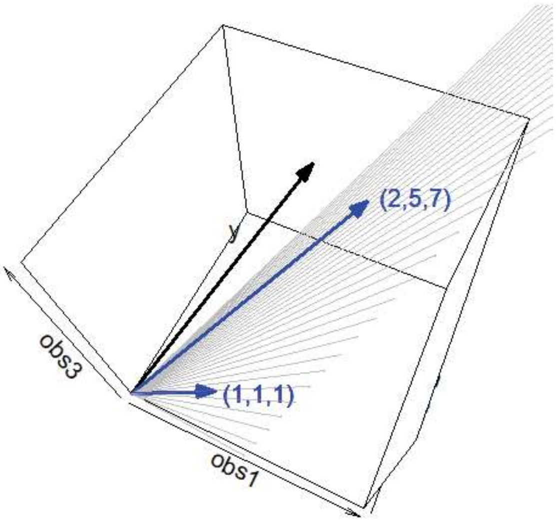
The vector $y = \begin{bmatrix}1 & 8 & 4\end{bmatrix}^{\mathrm{T}}$ is also a vector in $\mathbb{R}^3$ but it does not lie on the grey plane

y cannot be written as a linear combination of the blue vectors

If it did, the points

$$(X_{i1}, Y_i) = (2,1), (5,8), (7,4)$$

would lie in a straight line when plotted on x-y plane



Geometric Perspective

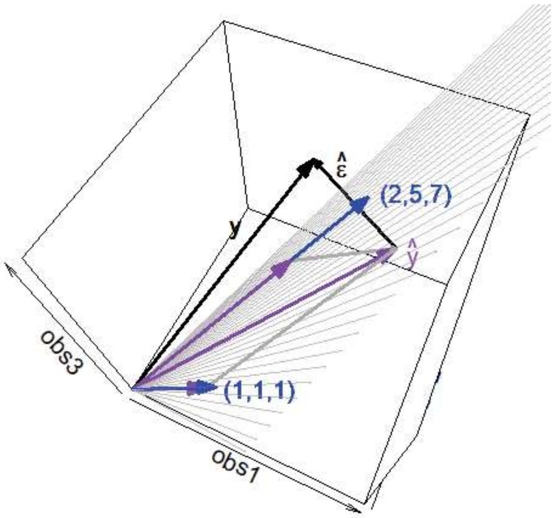
Let $\widehat{y}$ be any linear combination of the blue vectors, and $\widehat{\epsilon}$ be such that

$$y = \widehat{y} + \widehat{\epsilon}$$

Qn: Which linear combination of the blue vectors makes $\widehat{y}$ "closest" to y (in the sense that $\widehat{\epsilon}$ is as "short" as possible)?

Ans: The linear combination such that $\widehat{\epsilon}$ is perpendicular to grey plane

- $\widehat{\epsilon}$ is perpendicular to $[1 \quad 1 \quad 1]^{\mathrm{T}}$ and to $[2 \quad 5 \quad 7]^{\mathrm{T}}$



Geometric Perspective

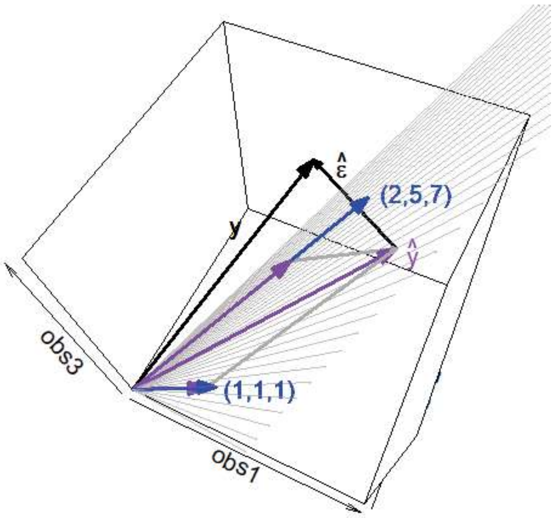
OLS chooses $\widehat{\beta}_0$ and $\widehat{\beta}_1$ such that

$$\underbrace{\widehat{y} = \begin{bmatrix}1\\1\\1\end{bmatrix}\widehat{\beta}_0 + \begin{bmatrix}2\\5\\7\end{bmatrix}\widehat{\beta}_1}_{\text{in Purple}} \quad \text{and} \quad \underbrace{\widehat{\epsilon} = \begin{bmatrix}\widehat{\epsilon}_1\\\widehat{\epsilon}_2\\\widehat{\epsilon}_3\end{bmatrix}}_{\text{in Black}}$$

are orthogonal (i.e., perpendicular), thereby minimizing

$$\|\widehat{\epsilon}\|^2 = \widehat{\epsilon}_1^2 + \widehat{\epsilon}_2^2 + \widehat{\epsilon}_3^2$$

$\widehat{y}$ is called the "orthogonal" projection of Y on to the column space of X



Geometric Perspective

Intuition holds for general MLR case

$$y = X\hat{\beta} + \hat{\epsilon}$$

where y is $n \times 1$ vector, X is $n \times (k+1)$ matrix whose columns space a $k+1$ dimensional subspace of $\mathbb{R}^n$

Cannot draw pictures in n-dimensions!

Nonetheless, OLS minimization solution shows that $\|\hat{\epsilon}\|$ is minimized when $\hat{\beta} = \hat{\beta}^{ols}$, which makes $\hat{y}$ and $\hat{\epsilon}$ orthogonal

$$\hat{y} = X\hat{\beta}^{ols} = \underbrace{X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}}_{\text{Projection Matrix}}y$$

Session 5.4

Session 5.4 Properties of OLS Estimators

- Unbiasedness under basic assumptions
- Standard errors under homoskedasticity and heteroskedasticity
- Best linear unbiasedness under homoskedasticity
- t-tests, F-tests

OLS Estimator Properties: Unbiasedness

- OLS estimator $\hat{\beta}^{ols}$ is unbiased:

$$\hat{\beta}^{ols} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(X\beta + \epsilon) = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\epsilon$$

$$E(\hat{\beta}^{ols} \mid X) = \beta + E((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\epsilon \mid X) = \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}E(\epsilon \mid X) = \beta$$

$$E(\hat{\beta}^{ols}) = \beta$$

  But remember points about causal interpretation
- you have estimated $E(Y \mid X) = X\beta$
- causal interpretation of coefficient only if you have included all relevant determinants of Y
- assuming no issues such as simultaneity, sampling problems, misspecification

OLS Estimator Variance (Homoskedasticity Case)

Under homoskedasticity $Var(\epsilon\epsilon^{\mathrm{T}} \mid X) = \sigma^2 I_n$, we have

$$\begin{aligned} Var(\hat{\beta}^{ols} \mid X) &= Var(\beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\epsilon \mid X) \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\,Var(\epsilon \mid X)X(X^{\mathrm{T}}X)^{-1} \\ &= (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}(\sigma^2 I_n)X(X^{\mathrm{T}}X)^{-1} \\ &= \sigma^2(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}I_nX(X^{\mathrm{T}}X)^{-1} \\ &= \sigma^2(X^{\mathrm{T}}X)^{-1} \end{aligned}$$

but must estimate σ^2

OLS Estimator Variance (Homoskedasticity Case)

Unbiased estimator of σ^2 is $\widehat{\sigma^2} = \dfrac{1}{n-k-1}\sum_{i=1}^{n}\hat{\epsilon}_{i,ols}^2 = \dfrac{\hat{\epsilon}_{ols}^{\mathrm{T}}\hat{\epsilon}_{ols}}{n-k-1}$

Proof: $\hat{\epsilon}_{ols} = My$ where $M = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$

note that

- $MX = (I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})X = X - X = 0_{n\times(k+1)}$
- M is symmetric (Exercise!)
- M is idempotent, i.e., $MM = M$ (Exercise!)

Therefore $\hat{\epsilon}_{ols} = My = M(X\beta + \epsilon) = M\epsilon$ and

$$\hat{\epsilon}_{ols}^{\mathrm{T}}\hat{\epsilon}_{ols} = (M\epsilon)^{\mathrm{T}}M\epsilon = \epsilon^{T}M^{\mathrm{T}}M\epsilon = \epsilon^{\mathrm{T}}MM\epsilon = \epsilon^{\mathrm{T}}M\epsilon$$

OLS Estimator Variance (Homoskedasticity Case)

$$\begin{aligned}
E(\hat{\epsilon}_{ols}^{\mathrm{T}}\hat{\epsilon}_{ols} \mid X) &= E(\epsilon^{\mathrm{T}}M\epsilon \mid X) \\
&= E(\text{trace}(\epsilon^{\mathrm{T}}M\epsilon) \mid X) \quad \text{because } \epsilon^{\mathrm{T}}M\epsilon \text{ is a scalar} \\
&= E(\text{trace}(\epsilon\epsilon^{\mathrm{T}}M) \mid X) \\
&= \text{trace}(E(\epsilon\epsilon^{\mathrm{T}}M) \mid X) \quad \text{since trace is a sum} \\
&= \text{trace}(E(\epsilon\epsilon^{\mathrm{T}} \mid X)M) = \text{trace}(\sigma^2 I_n M) = \sigma^2\text{trace}(M) \\
&= \sigma^2\text{trace}(I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}) = \sigma^2(\text{trace}(I_n) - \text{trace}(X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})) \\
&= \sigma^2(n - \text{trace}((X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}X)) = \sigma^2(n - \text{trace}(I_{k+1})) \\
&= \sigma^2(n-k-1)
\end{aligned}$$

Therefore $E\left(\dfrac{\hat{\epsilon}_{ols}^{\mathrm{T}}\hat{\epsilon}_{ols}}{n-k-1} \,\middle|\, X\right) = \sigma^2$, which implies $E\left(\dfrac{\hat{\epsilon}_{ols}^{\mathrm{T}}\hat{\epsilon}_{ols}}{n-k-1}\right) = \sigma^2$

OLS Estimator Variance (Heteroskedasticity Case)

$$Var(\hat{\beta}^{ols} \mid X) = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}Var(\epsilon \mid X)X(X^{\mathrm{T}}X)^{-1}$$

$$= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \begin{bmatrix} X_{1*}^{\mathrm{T}} & X_{2*}^{\mathrm{T}} & \cdots & X_{n*}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{bmatrix} \begin{bmatrix} X_{1*} \\ X_{2*} \\ \vdots \\ X_{n*} \end{bmatrix} \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}$$

$$= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \sigma_i^2 X_{i*}^{\mathrm{T}}X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}$$

Heteroskedasticity-Robust Estimator Variance-Covariance Matrix: replace σ_i^2 with $\hat{\epsilon}_{i,ols}^2$

$$\widehat{Var(\hat{\beta}^{ols})} = \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_{i,ols}^2 X_{i*}^{\mathrm{T}}X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}}X_{i*}\right)^{-1}$$

OLS (Best Linear Unbiased, under Homoskedasticity)

We have shown that $\hat{\beta}^{ols}$ is unbiased

It is also "linear"

- A linear estimator is one with the form $\tilde{\beta} = Ay$
- Each $\tilde{\beta}_j = \sum_{i=1}^{n} a_{ji}Y_i$
- OLS estimator is $\hat{\beta}^{ols} = \underbrace{(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}}_{A}y$

Now we show: OLS estimators are BLU (they have the smallest variance among all linear unbiased estimators)

OLS (Best Linear Unbiased, under Homoskedasticity)

BLU in the sense that

$$Var(c^{\mathrm{T}}\hat{\beta}^{ols} \mid X) \leq Var(c^{\mathrm{T}}\tilde{\beta} \mid X)$$

for all $(k+1) \times 1$ vectors c, and for all unbiased estimators of the form $\tilde{\beta} = By$

- each individual $\hat{\beta}_k$ is BLU
- all linear combinations of $\hat{\beta}$ are BLU

Consider predicting Y at the new observation $X_{0*} = [1 \quad X_{01} \quad \ldots \quad X_{0k}]$. OLS predictor is

$$\hat{Y}(X_{0*}) = X_{0*}\hat{\beta}^{ols}$$

which is a linear combination of the parameter estimates in $\hat{\beta}^{ols}$, i.e., OLS prediction rule gives us the most precise linear unbiased prediction of Y at X_{0*}

OLS (Best Linear Unbiased, under Homoskedasticity)

Proof of Efficiency: let $\tilde{\beta} = By$ be an unbiased estimator where $B \neq (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$

Let D be such that $B = D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$, so that

$$\begin{aligned}
\tilde{\beta} = By &= (D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})y \\
&= (D + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(X\beta + \epsilon) \\
&= DX\beta + D\epsilon + \beta + (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}\epsilon
\end{aligned}$$

To ensure unbiasedness of $\tilde{\beta}$, we must assume $DX = 0$

OLS (Best Linear Unbiased, under Homoskedasticity)

Then

$$\begin{aligned}
Var(\tilde{\beta} \mid X) &= E((\tilde{\beta}-\beta)(\tilde{\beta}-\beta)^{\mathrm{T}} \mid X) \\
&= E((D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})\varepsilon\varepsilon^{\mathrm{T}}(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \mid X) \\
&= (D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})E(\varepsilon\varepsilon^{\mathrm{T}} \mid X)(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \\
&= \sigma^2(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})^{\mathrm{T}} \\
&= \sigma^2(D+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}})(D^{\mathrm{T}}+X(X^{\mathrm{T}}X)^{-1}) \\
&= \sigma^2[DD^{\mathrm{T}}+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}D^{\mathrm{T}}+DX(X^{\mathrm{T}}X)^{-1}+(X^{\mathrm{T}}X)^{-1}] \\
&= \sigma^2[DD^{\mathrm{T}}+(X^{\mathrm{T}}X)^{-1}] = \sigma^2DD^{\mathrm{T}}+\sigma^2(X^{\mathrm{T}}X)^{-1} = \sigma^2DD^{\mathrm{T}}+Var(\hat{\beta} \mid X)
\end{aligned}$$

Agenda | A Bit More Matrix Algebra | OLS Estimation of MLR | Geometric Persp. | OLS Estimator Properties | Example | Asymptotics | Roadmap

OLS (Best Linear Unbiased, under Homoskedasticity)

Therefore

$$\begin{aligned}
Var(c^{\mathrm{T}}\tilde{\beta} \mid X) &= c^{\mathrm{T}}Var(\tilde{\beta} \mid X)c \\
&= c^{\mathrm{T}}(\sigma^2DD^{\mathrm{T}}+Var(\hat{\beta} \mid X))c \\
&= \sigma^2c^{\mathrm{T}}DD^{\mathrm{T}}c+c^{\mathrm{T}}Var(\hat{\beta} \mid X))c \\
&= \sigma^2(D^{\mathrm{T}}c)^{\mathrm{T}}D^{\mathrm{T}}c+Var(c^{\mathrm{T}}\hat{\beta} \mid X)) \geq Var(c^{\mathrm{T}}\hat{\beta} \mid X))
\end{aligned}$$

NB: $D^{\mathrm{T}}c$ is a vector, therefore $(D^{\mathrm{T}}c)^{\mathrm{T}}D^{\mathrm{T}}c$ is a sum of squares, which cannot be negative

Session 5.5

Session 5.5 Empirical Example

- Estimate Multiple Linear Regression using `earnings.csv` data

$$
\begin{aligned}
100\ln(earn) = \beta_0 + &\beta_1 educ + \beta_2 educ^2 + \beta_3 height + \beta_4 male + \\
&\beta_5 \ln(wexp) + \beta_6 \ln(wexp)^2 + \beta_7 \ln(tenure) + \beta_8 \ln(tenure)^2 + \\
&\beta_9 age + \beta_{10} age^2 + \epsilon
\end{aligned}
$$

- Everything computed twice
  - Using formulas derived in this session
  - Using existing R functions and packages

```r
library(tidyverse)
library(car)       # using linearHypothesis()
library(sandwich)  # using vcovHC()
library(lmtest)    # using coeftest()
```


Empirical Example

```r
dat <- read_csv("data\\earnings2019.csv", show_col_types=FALSE) %>%
         mutate(learn100 = 100*log(earn), educsq = educ^2, lwexp = log(wexp), lwexp2 = log(wexp)^2,
                          ltenure=log(tenure), ltenure2=log(tenure)^2, agesq = age^2)
y <- dat %>% select(learn100) %>% as.matrix()
X <- dat  %>% mutate(intercept = 1) %>%
  select(c(intercept, educ, educsq, height, male, lwexp, lwexp2, ltenure, ltenure2, age, agesq)) %>%
  as.matrix()
XTX <- t(X)%*%X                                     # (X'X)
XTXinv <- solve(XTX)                                # (X'X)^{-1}
XTy <- t(X)%*%y                                     # X'y
betahat <- XTXinv %*% XTy                           # beta estimate
yhat <- X%*%betahat                                 # fitted value
ehat <- y - yhat                                    # residuals
n <- nrow(X); kplus1 <- ncol(X); df <- n - kplus1   # n - k - 1
SST <- sum((y - mean(y))^2); SSE <- sum((yhat - mean(yhat))^2); SSR <- sum(ehat^2)
s2hat <- SSR/df
betavar <- s2hat*XTXinv
betase <- sqrt(diag(betavar))
tstats <- betahat/betase
pvals <- pt(abs(tstats), df=df, lower.tail=FALSE)*2
R2 <- 1 - SSR/SST; AdjR2 <- 1 - (SSR/df)/(SST/(n-1))
```

Empirical Example

Result using `lm()` function

```
mdl1 <- lm(learn100 ~ educ + educsq + height + male
  lwexp + lwexp2 + ltenure + ltenure2 + age + agesq,
sum1 <- summary(mdl1); sum1$coef %>% round(5)
cat("\nR2:", sum1$r.squared,
    " Adj.R2:", sum1$adj.r.squared)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	81.12348	37.53689	2.16117	0.03073
educ	-15.26088	4.81587	-3.16887	0.00154
educsq	1.00645	0.17185	5.85653	0.00000
height	1.34326	0.26387	5.09054	0.00000
male	18.07162	2.14105	8.44054	0.00000
lwexp	-1.58201	2.71219	-0.58330	0.55972
lwexp2	-0.22626	0.77976	-0.29016	0.77170
ltenure	5.22349	2.66119	1.96284	0.04972
ltenure2	2.31435	0.76945	3.00781	0.00264
age	5.82374	0.47516	12.25627	0.00000
agesq	-0.06125	0.00517	-11.84342	0.00000

```
R2: 0.303907  Adj.R2: 0.3024965
```

Result via direct computation

```
results <- cbind(betahat, betase, tstats, pvals)
colnames(results) <- c("Estimate", "Std. Error",
                       "t values", "Pr(>|t|)")
results %>% round(5)
cat("\nR2:", R2, " Adj.R2:", AdjR2)
```

	Estimate	Std. Error	t values	Pr(>|t|)
intercept	81.12348	37.53689	2.16117	0.03073
educ	-15.26088	4.81587	-3.16887	0.00154
educsq	1.00645	0.17185	5.85653	0.00000
height	1.34326	0.26387	5.09054	0.00000
male	18.07162	2.14105	8.44054	0.00000
lwexp	-1.58201	2.71219	-0.58330	0.55972
lwexp2	-0.22626	0.77976	-0.29016	0.77170
ltenure	5.22349	2.66119	1.96284	0.04972
ltenure2	2.31435	0.76945	3.00781	0.00264
age	5.82374	0.47516	12.25627	0.00000
agesq	-0.06125	0.00517	-11.84342	0.00000

```
R2: 0.303907  Adj.R2: 0.3024965
```


Empirical Example: Var-Cov under Homosked. (self-calc.)

```
options(width=400); round(betavar,3)  # VAR-COV UNDER HOMOSKEDASTICITY
```

	intercept	educ	educsq	height	male	lwexp	lwexp2	ltenure	ltenure2	age	agesq
intercept	1409.018	-154.411	5.479	-3.632	15.606	1.019	0.115	-3.887	1.708	-4.886	0.050
educ	-154.411	23.193	-0.825	-0.110	1.009	-0.360	0.076	0.215	-0.099	0.104	-0.001
educsq	5.479	-0.825	0.030	0.004	-0.033	0.014	-0.002	-0.010	0.004	-0.004	0.000
height	-3.632	-0.110	0.004	0.070	-0.385	-0.016	0.003	0.009	-0.002	-0.004	0.000
male	15.606	1.009	-0.033	-0.385	4.584	0.081	-0.020	-0.077	-0.002	0.038	0.000
lwexp	1.019	-0.360	0.014	-0.016	0.081	7.356	-1.988	-0.108	0.004	-0.136	0.002
lwexp2	0.115	0.076	-0.002	0.003	-0.020	-1.988	0.608	0.010	0.006	0.020	0.000
ltenure	-3.887	0.215	-0.010	0.009	-0.077	-0.108	0.010	7.082	-1.935	-0.120	0.002
ltenure2	1.708	-0.099	0.004	-0.002	-0.002	0.004	0.006	-1.935	0.592	0.012	0.000
age	-4.886	0.104	-0.004	-0.004	0.038	-0.136	0.020	-0.120	0.012	0.226	-0.002
agesq	0.050	-0.001	0.000	0.000	0.000	0.002	0.000	0.002	0.000	-0.002	0.000

```
betavar %>% diag %>% as.matrix %>% t %>% sqrt %>% round(3)  # STANDARD ERRORS
```

	intercept	educ	educsq	height	male	lwexp	lwexp2	ltenure	ltenure2	age	agesq
[1,]	37.537	4.816	0.172	0.264	2.141	2.712	0.78	2.661	0.769	0.475	0.005

Empirical Example: Var-Cov under Homoskedasticity with lm()

```
options(width=400); round(vcov(mdl1),3)  # VAR-COV UNDER HOMOSKEDASTICITY lm()
```

```
             (Intercept)     educ educsq height   male  lwexp lwexp2 ltenure ltenure2    age  agesq
(Intercept)     1409.018 -154.411  5.479 -3.632 15.606  1.019  0.115  -3.887    1.708 -4.886  0.050
educ            -154.411   23.193 -0.825 -0.110  1.009 -0.360  0.076   0.215   -0.099  0.104 -0.001
educsq             5.479   -0.825  0.030  0.004 -0.033  0.014 -0.002  -0.010    0.004 -0.004  0.000
height            -3.632   -0.110  0.004  0.070 -0.385 -0.016  0.003   0.009   -0.002 -0.004  0.000
male              15.606    1.009 -0.033 -0.385  4.584  0.081 -0.020  -0.077   -0.002  0.038  0.000
lwexp              1.019   -0.360  0.014 -0.016  0.081  7.356 -1.988  -0.108    0.004 -0.136  0.002
lwexp2             0.115    0.076 -0.002  0.003 -0.020 -1.988  0.608   0.010    0.006  0.020  0.000
ltenure           -3.887    0.215 -0.010  0.009 -0.077 -0.108  0.010   7.082   -1.935 -0.120  0.002
ltenure2           1.708   -0.099  0.004 -0.002 -0.002  0.004  0.006  -1.935    0.592  0.012  0.000
age               -4.886    0.104 -0.004 -0.004  0.038 -0.136  0.020  -0.120    0.012  0.226 -0.002
agesq              0.050   -0.001  0.000  0.000  0.000  0.002  0.000   0.002    0.000 -0.002  0.000
```

```
vcov(mdl1) %>% diag %>% as.matrix %>% t %>% sqrt %>% round(3)  # STANDARD ERRORS
```

```
     (Intercept)  educ educsq height  male lwexp lwexp2 ltenure ltenure2   age agesq
[1,]      37.537 4.816  0.172  0.264 2.141 2.712   0.78   2.661    0.769 0.475 0.005
```


Empirical Example: Het-Robust Var-Cov (self-calculated)

```
options(width=400);
varhat_HC0 <- XTXinv%*%t(X)%*%diag(as.vector(ehat^2))%*%X%*%XTXinv; round(varhat_HC0,3)
```

```
           intercept     educ educsq height   male  lwexp lwexp2 ltenure ltenure2    age  agesq
intercept   1345.536 -144.697  5.136 -3.296 14.791 -2.703  1.435 -10.416    3.477 -5.957  0.063
educ        -144.697   22.049 -0.785 -0.163  0.932  0.350 -0.173   0.829   -0.232  0.163 -0.002
educsq         5.136   -0.785  0.028  0.006 -0.028 -0.012  0.007  -0.032    0.009 -0.006  0.000
height        -3.296   -0.163  0.006  0.071 -0.382 -0.022  0.008   0.006   -0.001 -0.006  0.000
male          14.791    0.932 -0.028 -0.382  4.563  0.076 -0.048   0.098   -0.053  0.075 -0.001
lwexp         -2.703    0.350 -0.012 -0.022  0.076  8.468 -2.300  -0.121    0.041 -0.208  0.002
lwexp2         1.435   -0.173  0.007  0.008 -0.048 -2.300  0.705   0.029   -0.008  0.034 -0.001
ltenure      -10.416    0.829 -0.032  0.006  0.098 -0.121  0.029   7.554   -2.077 -0.030  0.001
ltenure2       3.477   -0.232  0.009 -0.001 -0.053  0.041 -0.008  -2.077    0.639 -0.022  0.000
age           -5.957    0.163 -0.006 -0.006  0.075 -0.208  0.034  -0.030   -0.022  0.266 -0.003
agesq          0.063   -0.002  0.000  0.000 -0.001  0.002 -0.001   0.001    0.000 -0.003  0.000
```

```
varhat_HC0 %>% diag %>% as.matrix %>% t %>% sqrt %>% round(3)  # STANDARD ERRORS
```

```
     intercept  educ educsq height  male lwexp lwexp2 ltenure ltenure2   age agesq
[1,]    36.682 4.696  0.168  0.266 2.136  2.91   0.84   2.748      0.8 0.516 0.006
```

Empirical Example : Het-Robust Var-Cov (using packages)

- Can use `hccm()` from car package or `vcovHC()` function from sandwich package

```
options(width=400); varhat_HC0a = vcovHC(mdl1,type="HC0"); round(varhat_HC0a,3)  # HET-ROBUST VAR
```

```
             (Intercept)     educ educsq height   male  lwexp lwexp2 ltenure ltenure2    age  agesq
(Intercept)     1345.536 -144.697  5.136 -3.296 14.791 -2.703  1.435 -10.416    3.477 -5.957  0.063
educ            -144.697   22.049 -0.785 -0.163  0.932  0.350 -0.173   0.829   -0.232  0.163 -0.002
educsq             5.136   -0.785  0.028  0.006 -0.028 -0.012  0.007  -0.032    0.009 -0.006  0.000
height            -3.296   -0.163  0.006  0.071 -0.382 -0.022  0.008   0.006   -0.001 -0.006  0.000
male              14.791    0.932 -0.028 -0.382  4.563  0.076 -0.048   0.098   -0.053  0.075 -0.001
lwexp             -2.703    0.350 -0.012 -0.022  0.076  8.468 -2.300  -0.121    0.041 -0.208  0.002
lwexp2             1.435   -0.173  0.007  0.008 -0.048 -2.300  0.705   0.029   -0.008  0.034 -0.001
ltenure          -10.416    0.829 -0.032  0.006  0.098 -0.121  0.029   7.554   -2.077 -0.030  0.001
ltenure2           3.477   -0.232  0.009 -0.001 -0.053  0.041 -0.008  -2.077    0.639 -0.022  0.000
age               -5.957    0.163 -0.006 -0.006  0.075 -0.208  0.034  -0.030   -0.022  0.266 -0.003
agesq              0.063   -0.002  0.000  0.000 -0.001  0.002 -0.001   0.001    0.000 -0.003  0.000
```

```
varhat_HC0a %>% diag %>% as.matrix %>% t %>% sqrt %>% round(3)  # STANDARD ERRORS
```

```
     (Intercept)  educ educsq height  male lwexp lwexp2 ltenure ltenure2   age agesq
[1,]      36.682 4.696  0.168  0.266 2.136  2.91   0.84   2.748      0.8 0.516 0.006
```


Session 5.6

Session 5.6 Asymptotic Properties of OLS Estimators

- Consistency
- Asymptotic Normality
- Theory behind Heteroskedasticity-Robust Standard Errors

Asymptotic Properties of OLS Estimators

Recall

- Khinchine's LLN: $If \{Z_i\}_{i=1}^n$ iid, $E(Z_i) = \mu < \infty$, then $\overline{Z} \xrightarrow{p} \mu$.
- Lindeberg-Levy CLT: $\{Z_i\}_{i=1}^n$ iid, $E(Z_i) = \mu$ and $Var(Z_i) = \sigma^2 < \infty$, then

$$\sqrt{n}(\overline{Z} - \mu) \xrightarrow{d} \text{Normal}(0, \sigma^2)$$

- if $\mu = 0$, then

$$\sqrt{n}\,\overline{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i \xrightarrow{d} \text{Normal}(0, \sigma^2).$$

Asymptotic Properties of OLS Estimators

Multivariate versions of these rules

If $\{Z_i\}_{i=1}^n$ iid vectors of random variables with $E(Z_i) = 0$ and $Var(Z_i) = \Omega$, for all i, then

$$\overline{Z} = \frac{1}{n} \sum_{i=1}^{n} Z_i = \begin{bmatrix} (1/n)\sum_{i=1}^n Z_{1i} \\ (1/n)\sum_{i=1}^n Z_{2i} \\ \vdots \\ (1/n)\sum_{i=1}^n Z_{ki} \end{bmatrix} \xrightarrow{p} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix} = \mu$$

$$\sqrt{n}\,\overline{Z} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Z_i = \begin{bmatrix} (1/\sqrt{n})\sum_{i=1}^n Z_{1i} \\ (1/\sqrt{n})\sum_{i=1}^n Z_{2i} \\ \vdots \\ (1/\sqrt{n})\sum_{i=1}^n Z_{ki} \end{bmatrix} \xrightarrow{d} \text{Normal}_k(0, \Omega)$$

Asymptotic Properties

To talk about limiting distributions, we have to scale $\hat{\beta}$. Use

$$\sqrt{n}(\hat{\beta}^{ols} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} \epsilon_i\right)$$

Our assumptions and the CLT imply

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i*}\epsilon_i \xrightarrow{d} \text{Normal}_{k+1}(0, S)$$

therefore

$$\sqrt{n}(\hat{\beta}^{ols} - \beta) = \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}}_{\xrightarrow{p} \Sigma_{xx}^{-1}} \underbrace{\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} \epsilon_i\right)}_{\xrightarrow{p} \text{Normal}_{k+1}(0,S)} \xrightarrow{d} \text{Normal}_{k+1}(0, \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1})$$

Asymptotic Properties

That is, $\hat{\beta}^{ols}$ is consistent, with asymptotic variance $Avar(\hat{\beta}^{ols}) = \Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}$. This result justifies the approximation

$$Var(\hat{\beta}^{ols}) \approx (1/n)\Sigma_{xx}^{-1} S \Sigma_{xx}^{-1}.$$

An obvious estimator for Σ_{xx} is

$$\hat{\Sigma}_{xx} = \frac{1}{n}\sum_{i=1}^{N} X_{i*}^{\mathrm{T}} X_{i*} = \frac{1}{n} X^{\mathrm{T}} X$$

Some additional assumptions (see advanced econometrics textbooks) guarantee

$$\hat{S} = \frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*} \xrightarrow{p} S.$$

Asymptotic Properties

This allows us to consistently estimate the asymptotic variance of $\widehat{\beta}$ by

$$\widehat{Avar}(\hat{\beta}^{ols}) = \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}$$

and justifies the use of

$$\begin{aligned}
\widehat{Var}_{HC0}(\hat{\beta}^{ols}) &= \frac{1}{n}\widehat{Avar}(\hat{\beta}^{ols}) \\
&= \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\frac{1}{n}\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \\
&= \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 X_{i*}^{\mathrm{T}} X_{i*}\right) \left(\sum_{i=1}^{n} X_{i*}^{\mathrm{T}} X_{i*}\right)^{-1}
\end{aligned}$$

Roadmap

- (Previous) Session 1: Statistics Review
- (Previous) Session 2: Simple Linear Regression
- (Previous) Session 3: Estimator Standard Errors; Multiple Linear Regression
- (Previous) Session 4: Matrix Algebra
- **This Session 5: OLS using Matrix Algebra**
- *Next Session 6: Hypothesis Testing*
- Session 7: Prediction
- Session 8: Instrumental Variable Regression
- Session 9: Logistic and Other Regressions
- Session 10: Panel Data Regressions
- Session 11: Introduction to Time Series
- Session 12: Time Series Regressions