

# Lecture 18

## Quantitative Political Science

Prof. Bisbee

Vanderbilt University

Lecture Date: 2023/11/09

Slides Updated: 2023-11-13

# Agenda

1. Matrix Algebra fun!
2. Multiple Regression
3. Controls

# Matrix Algebra Fun! Thanks BK!

- Vectors: ordered arrays denoted  $\mathbf{v} = (v_1, v_2, \dots, v_k)$  or

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix}$$

- (Note that some will denote vectors with bold letters, or with  $\vec{v}$ )
- Addition and subtraction require two vectors of the same length,  $\mathbf{u}$  and  $\mathbf{v}$ , but are then just adding or subtracting the elements

$$\mathbf{u} \pm \mathbf{v} = \begin{pmatrix} u_1 \pm v_1 \\ u_2 \pm v_2 \\ \vdots \\ u_k \pm v_k \end{pmatrix}$$

# Vectors

- Multiplication by a constant  $c$  is just multiplying each element by  $c$

$$c\mathbf{v} = \begin{pmatrix} cv_1 \\ cv_2 \\ \vdots \\ cv_k \end{pmatrix}$$

- Multiplication of two vectors is called a **dot product**, written  $\mathbf{u} \cdot \mathbf{v}$ , and translates to multiplying each element in  $\mathbf{u}$  by the corresponding element in  $\mathbf{v}$  and then adding them all up

$$\begin{aligned} \mathbf{u} \cdot \mathbf{v} &= u_1v_1 + u_2v_2 + \cdots + u_kv_k \\ &= \sum_{m=1}^k u_mv_m \end{aligned}$$

# Matrices

- A matrix is a two-dimensional array with entries in  $n$  rows and  $m$  columns, called an  $n \times m$  matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

- As with vectors, matrices can be added and subtracted *as long as they are the same dimensions*

$$\mathbf{A} \pm \mathbf{B} = \begin{bmatrix} a_{11} \pm b_{11} & a_{12} \pm b_{12} & a_{13} \pm b_{13} \\ a_{21} \pm b_{21} & a_{22} \pm b_{22} & a_{23} \pm b_{23} \\ a_{31} \pm b_{31} & a_{32} \pm b_{32} & a_{33} \pm b_{33} \end{bmatrix}$$

- As with vectors, matrices multiplied by a constant are straightforward

$$c\mathbf{A} = \begin{bmatrix} ca_{11} & ca_{12} & ca_{13} \\ ca_{21} & ca_{22} & ca_{23} \\ ca_{31} & ca_{32} & ca_{33} \end{bmatrix}$$

# Matrices: Transpose

- Transposing: we can "rotate"  $n \times m$  matrices into  $m \times n$  matrices
  - Meaning that the first row becomes the first column, the second row becomes the second column, etc.
  - Denoted with  $\mathbf{A}^\top$  (or sometimes  $\mathbf{A}'$ )
- For example:

$$\mathbf{A} = \begin{bmatrix} 99 & 73 & 2 \\ 13 & 40 & 41 \end{bmatrix} \Leftrightarrow \mathbf{A}^\top = \begin{bmatrix} 99 & 13 \\ 73 & 40 \\ 2 & 41 \end{bmatrix}$$

# Matrices: Transpose

- Properties of transposes

$$(\mathbf{A}^\top)^\top = \mathbf{A},$$

$$(c\mathbf{A})^\top = c(\mathbf{A}^\top),$$

$$(\mathbf{A} + \mathbf{B})^\top = \mathbf{A}^\top + \mathbf{B}^\top,$$

$$(\mathbf{A} - \mathbf{B})^\top = \mathbf{A}^\top - \mathbf{B}^\top,$$

$$(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top.$$

- Note that it doesn't make sense to transpose a scalar
  - But also that this means a scalar is always equal to its transpose:  $a = a^\top$

# Matrix Multiplication

- Refresher: need to multiply an  $n \times m$  matrix by an  $m \times p$  matrix.
  - **NOTE:** the number of rows in the second matrix must be equal to the number of columns in the first matrix!
- Resulting matrix is an  $n \times p$  matrix whose  $ij$ 'th element is the **dot product** of the  $i$ 'th row of the first matrix and the  $j$ 'th column of the second matrix
- Try it: solve  $\mathbf{AB}$  where

$$\mathbf{A} = \begin{bmatrix} 2 & 10 \\ 0 & 1 \\ -1 & 5 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 1 & 4 \\ -1 & 10 \end{bmatrix}$$

$$\mathbf{AB} = \begin{bmatrix} 2 \cdot 1 + 10 \cdot (-1) & 2 \cdot 4 + 10 \cdot 10 \\ 0 \cdot 1 + 1 \cdot (-1) & 0 \cdot 4 + 1 \cdot 10 \\ (-1) \cdot 1 + 5 \cdot (-1) & (-1) \cdot 4 + 5 \cdot 10 \end{bmatrix} = \begin{bmatrix} -8 & 108 \\ -1 & 10 \\ -6 & 46 \end{bmatrix}$$



# Matrix Multiplication

- Properties of matrix multiplication
  - **Associative:**  $(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC})$
  - **Distributive:**  $\mathbf{A}(\mathbf{B} + \mathbf{C}) = \mathbf{AB} + \mathbf{AC}$
  - **NOT** commutative:  $\mathbf{AB} \neq \mathbf{BA}$
  - **Transpose Rule:**  $(\mathbf{AB})^\top = \mathbf{B}^\top \mathbf{A}^\top$

# Matrix Expectations

- Expectations are easily distributed throughout a matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & x_{13} \\ x_{21} & x_{22} & x_{23} \\ x_{31} & x_{32} & x_{33} \end{bmatrix} \quad ; \quad E(\mathbf{X}) = \begin{bmatrix} E(x_{11}) & E(x_{12}) & E(x_{13}) \\ E(x_{21}) & E(x_{22}) & E(x_{23}) \\ E(x_{31}) & E(x_{32}) & E(x_{33}) \end{bmatrix}$$

# Matrix Derivatives

- Consider a matrix equation of the form  $\mathbf{y} = \mathbf{A}\mathbf{x}$ , meaning that each row is  $y_i = a_{1i}x_1 + a_{2i}x_2 + \cdots + a_{ki}x_k$
- In matrix notation:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{k1} \\ a_{12} & a_{22} & \cdots & a_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{kn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_k \end{bmatrix}$$

- To take the partial derivative with respect to  $\mathbf{x}$ , we go element by element in  $\mathbf{y}$ :  $\frac{\partial y_1}{\partial \mathbf{x}}, \frac{\partial y_2}{\partial \mathbf{x}}, \dots, \frac{\partial y_n}{\partial \mathbf{x}}$
- But to do THIS, we again go element by element through each value of  $\mathbf{x}$ , noting that  $\frac{\partial y_1}{\partial x_1} = a_{11}$  and  $\frac{\partial y_1}{\partial x_2} = a_{21}$ , and that  $\frac{\partial y_2}{\partial x_1} = a_{12}$  and  $\frac{\partial y_2}{\partial x_2} = a_{22}$

# Matrix Derivatives

- We can write these in vector form as follows:

$$\frac{\partial y_1}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} \\ \vdots \\ \frac{\partial y_1}{\partial x_k} \end{bmatrix} = \begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{k1} \end{bmatrix}; \quad \frac{\partial y_2}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_2}{\partial x_1} \\ \frac{\partial y_2}{\partial x_2} \\ \vdots \\ \frac{\partial y_2}{\partial x_k} \end{bmatrix} = \begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{k2} \end{bmatrix}; \quad \dots \quad \frac{\partial y_n}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_n}{\partial x_2} \\ \vdots \\ \frac{\partial y_n}{\partial x_k} \end{bmatrix} = \begin{bmatrix} a_{1n} \\ a_{2n} \\ \vdots \\ a_{kn} \end{bmatrix}$$

- Now let's just combine each of these vectors of derivatives into its own matrix to yield:

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{k1} & a_{k2} & \dots & a_{kn} \end{bmatrix} = \mathbf{A}^\top$$

# Matrix Derivatives

- Thus  $\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial(\mathbf{Ax})}{\partial \mathbf{x}} = \mathbf{A}^\top$
- From this, we can also note that, given  $y = \mathbf{a}^\top \mathbf{x}$ ,  $\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^\top \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a}$
- And also, given  $y = \mathbf{x}^\top \mathbf{Ax}$ ,  $\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^\top \mathbf{Ax}}{\partial \mathbf{x}} = 2\mathbf{Ax}$
- And finally, given  $y = \mathbf{x}^\top \mathbf{Ax}$ ,  $\frac{\partial y}{\partial \mathbf{A}} = \frac{\partial \mathbf{x}^\top \mathbf{Ax}}{\partial \mathbf{A}} = \mathbf{xx}^\top$

# Special Matrices

- **Zero** matrices:  $\mathbf{0}$  has all entries as zero
  - NB:  $\mathbf{A}_{r \times c} \cdot \mathbf{0}_{c \times n} = \mathbf{0}_{r \times n}$  and  $\mathbf{0}_{n \times r} \cdot \mathbf{A}_{r \times c} = \mathbf{0}_{n \times c}$
- **Square** matrices:  $n \times n$  size, meaning the same number of rows as columns
- **Symmetric** square matrices:  $\mathbf{A} = \mathbf{A}^\top$
- **Diagonal** symmetric square matrices: zeros everywhere except the diagonal: if  $i$  are rows and  $j$  are columns,  $i \neq j$ , then  $a_{ij} = 0$ .
- **Identity** diagonal symmetric square matrices:  $\mathbf{I}_n$  is a diagonal matrix where the diagonals are 1s
  - What is

$$\mathbf{A} = \begin{bmatrix} 99 & 73 & 2 \\ 13 & 40 & 41 \end{bmatrix} \quad . \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

# Matrix Inversion

- In the scalar world, we know we can rewrite a division problem  $\frac{a}{b}$  as a multiplication problem  $a \times \frac{1}{b} = a \times b^{-1}$ 
  - $b^{-1}$  is the inverse of  $b$
  - The (obvious) requirement for the inverse is that  $b \times b^{-1} = \frac{b}{1} \times \frac{1}{b} = \frac{b}{b} = 1$
- In the matrix world, the inverse of a matrix  $\mathbf{A}$  is denoted  $\mathbf{A}^{-1}$  and must also satisfy:  $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}_n$
- Some properties!
  - If  $\mathbf{C}$  is an inverse of  $\mathbf{A}$ , then  $\mathbf{A}$  is also the inverse of  $\mathbf{C}$
  - If  $\mathbf{C}$  and  $\mathbf{D}$  are both inverses of  $\mathbf{A}$ , then  $\mathbf{C} = \mathbf{D}$
  - The inverse of an inverse of  $\mathbf{A}$  is just  $\mathbf{A}$ :  $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
  - The inverse of  $\mathbf{A}^\top$  is the same as the inverse of  $\mathbf{A}$ , transposed:  $(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$
  - If you have a scalar  $c$  multiplied by a matrix  $\mathbf{A}$ , then  $(c\mathbf{A})^{-1} = \frac{1}{c}\mathbf{A}^{-1}$

# Matrix Inversion

- To invert a  $2 \times 2$  matrix, follow this rule:
- For

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

- Invert using

$$\mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

- where  $ad - bc$  is known as the **determinant** of the matrix  $\mathbf{A}$ , so named because it "determines" whether a matrix is invertible.
  - Why would it not be invertible? If  $ad - bc = 0$  or  $ad = bc$ !



# Matrix Inversion

- Matrix inversion gets harder with larger matrices...you can learn how to do it manually, but this is where software like R comes in handy!

```
A <- matrix(c(2, 1, 3, 4),  
            nrow = 2,  
            ncol = 2)
```

A

```
##      [,1] [,2]  
## [1,]    2    3  
## [2,]    1    4
```

- Use the `solve()` function to get the inverse of A

```
A_inv <- solve(A)  
A_inv
```

```
##      [,1] [,2]  
## [1,]  0.8 -0.6  
## [2,] -0.2  0.4
```

# Matrix Math in R

- R also can make our lives easier for matrix multiplication...just use `%%` instead of the standard `*`

```
# Use %% to do matrix multiplication
A*A_inv # Doesn't work...just does element-by-element multiplication
```

```
##      [,1] [,2]
## [1,]  1.6 -1.8
## [2,] -0.2  1.6
```

```
A %% A_inv # Works! We've proved that A_inv is the inverse of A!
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

# Why all this!?

- It helps us solve systems of equations!
- Back in the day, you probably had lots of practice with these types of things:

$$\begin{aligned}2x_1 + x_2 &= 10, \\ 2x_1 - x_2 &= -10\end{aligned}$$

- You probably learned to solve it various ways (i.e., solve for  $x_1$  first then plug in)
- We can solve with matrix math instead!

$$\begin{aligned}\mathbf{A} &= \begin{bmatrix} 2 & 1 \\ 2 & -1 \end{bmatrix}, \\ \mathbf{x} &= \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \\ \mathbf{b} &= \begin{bmatrix} 10 \\ -10 \end{bmatrix}\end{aligned}$$

# Systems of Equations

- We can re-write the two equations with matrix notation as  $\mathbf{Ax} = \mathbf{b}$
- To solve for  $\mathbf{x}$ , we just invert  $\mathbf{A}$  and write  $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$

```
A <- matrix(c(2, 1, 2, -1),  
            nrow = 2,  
            ncol = 2)  
b <- matrix(c(10, -10), nrow = 2, ncol = 1)  
  
solve(A)%*%b
```

```
##      [,1]  
## [1,] -2.5  
## [2,]  7.5
```

- $x_1 = -2.5$  and  $x_2 = 7.5$ ! So easy!
- Note that there is a unique solution for  $x_1$  and  $x_2$  iff  $\mathbf{A}$  is invertible
  - If not, there is either no solution or infinitely many solutions

# Multiple Regression

- We can use matrix algebra to help us with **multiple regression** (one outcome with multiple predictors)
  - Note: **multivariate regression** (multiple outcomes)  $\neq$  multiple regression
- Let's start with familiar notation and then break it down:  $y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + u_i$
- What does  $y$  look like? I mean this literally...what is it in a dataset?
  - It is an  $n$ -length vector of values  $\mathbf{y}$ , one for each row in our dataset!
- $\mathbf{x}$  and  $\mathbf{z}$  are the same

```
##      respondent_id      y      x      z
## 1      1  1.48840379 -1.270882210 -0.560475647
## 2      2  1.56929669  0.026706220 -0.230177489
## 3      3 -0.51183694  1.312016436  1.558708314
## 4      4  0.19565146 -0.277034208  0.070508391
## 5      5 -1.36595852 -0.822330832  0.129287735
## 6      6 -0.52127462  1.670037262  1.715064987
## 7      7 -1.57350731 -0.323988263  0.460916206
## 8      8 -2.26255920 -2.933003171 -1.265061235
## 9      9  1.270680095 -1.067079372 -0.686852852
```

# Multiple Regression

- Let's now look at the data in a different way, from the perspective of a single unit of observation
  - I.e., if we are dealing with a survey of individuals, our data might have some respondent 7 for whom we observe both  $y_7$  as well as  $x_7$  and  $z_7$
- From this perspective, unit 7 is associated with an outcome  $y_7$  (a single value) and then a vector of predictors:  $\mathbf{x}_7 = (x_7, z_7)$

```
dat %>% slice(7)
```

```
##      respondent_id      y      x      z
## 1                7 -1.573507 -0.3239883 0.4609162
```

- We can write our regression equation for this specific respondent as  $y_7 = \beta_0 + \beta_1 x_7 + \beta_2 z_7 + u_7$ , or we can write it as  $y_7 = \mathbf{x}_7 \cdot \boldsymbol{\beta} + u_7$  (note that this is the same as writing )
  - $\boldsymbol{\beta}$  is now itself a **vector** of coefficients:  $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$
  - $\mathbf{x}_7$  now needs to include the number 1:  $\mathbf{x}_7 = (1, x_7, z_7)$  in order to capture the  $\beta_0$  coefficient.

# Multiple Regression

- We can then think of  $\beta$  as a  $k \times 1$  vector (where  $k$  is the number of predictors) and  $\mathbf{x}_7$  as a  $1 \times k$  vector, and then matrix multiply them!

$$\begin{aligned}y_7 &= \mathbf{x}_7 \cdot \beta + u_7 \\&= \begin{bmatrix} 1 & x_7 & z_7 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix} + u_7 \\&= \beta_0 + \beta_1 x_7 + \beta_2 z_7 + u_7,\end{aligned}$$

- Now this was just one observation in our data, but we can imagine doing this for every single row, and then stacking our equations on top of each other

$$\begin{aligned}y_1 &= \beta \cdot \mathbf{x}_1 + u_1, \\y_2 &= \beta \cdot \mathbf{x}_2 + u_2, \\&\vdots \\y_n &= \beta \cdot \mathbf{x}_N + u_n.\end{aligned}$$

# Multiple Regression

- As with any system of equations, we can re-write as vectors and matrices

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \\ 1 & x_n & z_n \end{bmatrix}, \quad \mathbf{u} = \begin{bmatrix} u_1 & u_2 & \vdots & u_n \end{bmatrix}$$

- Plugging in:  $\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u}$
- Note that this is the same as writing:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ \vdots & \vdots & \\ 1 & x_n & z_n \end{bmatrix}_{n \times k} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}_{k \times 1} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n \times 1}$$

- where  $k$  is the number of parameters (in this case, 3) and  $n$  is the number of observations



# Multiple Regression

- Note that  $\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u}$  is assumed to be a reflection of the real world
  - Aside: prove to yourself that  $\mathbf{y} = \mathbf{X} \cdot \boldsymbol{\beta} + \mathbf{u}$  and  $\mathbf{y} = \boldsymbol{\beta}^\top \cdot \mathbf{X} + \mathbf{u}$  are equivalent
- We estimate these, as before, with our OLS estimators  $\hat{\boldsymbol{\beta}}$
- To do so, we first calculate our residuals as  $\mathbf{u} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ , and then add them up and square them.
  - In the **scalar** world, we would write this as  $\sum u_i^2$ .
  - In the **vector** world, we write this as  $\mathbf{u}^\top \mathbf{u}$ . Take a moment and try to see why!

$$\mathbf{u}^\top \mathbf{u} = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix}_{1 \times n} \cdot \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} u_1 * u_1 + u_2 * u_2 + \dots + u_n * u_n \end{bmatrix}_{1 \times n}$$

- Note that  $\mathbf{u}^\top \mathbf{u} = \begin{bmatrix} u_1 * u_1 + u_2 * u_2 + \dots + u_n * u_n \end{bmatrix}$  is the same as  $\sum u_i^2$ !!

# Multiple Regression

- We can re-write the sum of squared residuals as  $\mathbf{u}^\top \mathbf{u} = (\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta})$  by plugging in
- Now let's try doing some reorganizing of this

$$\begin{aligned}(\mathbf{y} - \mathbf{X}\hat{\beta})^\top (\mathbf{y} - \mathbf{X}\hat{\beta}) &= (\mathbf{y}^\top - \hat{\beta}^\top \mathbf{X}^\top)(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \mathbf{y}^\top \mathbf{y} - \mathbf{y}^\top \mathbf{X}\hat{\beta} - \hat{\beta}^\top \mathbf{X}^\top \mathbf{y} + \hat{\beta}^\top \mathbf{X}^\top \mathbf{X}\hat{\beta}\end{aligned}$$

- To subtract, it must be that  $\mathbf{y}^\top \mathbf{y}$  is conformable with  $\mathbf{y}^\top \mathbf{X}\hat{\beta}$ , meaning they must have the same dimensions
  - Note that  $\mathbf{y}^\top \mathbf{y}$  is a scalar, meaning that  $\mathbf{y}^\top \mathbf{X}\hat{\beta}$  must also be a scalar (by conformability)
  - Thus we can re-write  $\mathbf{y}^\top \mathbf{X}\hat{\beta} = (\mathbf{y}^\top \mathbf{X}\hat{\beta})^\top = \hat{\beta}^\top \mathbf{X}^\top \mathbf{y}$  (by transpose of a scalar)
- Substitute this in to reduce to:

$$\mathbf{u}^\top \mathbf{u} = \mathbf{y}^\top \mathbf{y} - 2\hat{\beta}^\top \mathbf{X}^\top \mathbf{y} + \hat{\beta}^\top \mathbf{X}^\top \mathbf{X}\hat{\beta}$$

# Multiple Regression

- Take the derivative with respect to  $\hat{\beta}$  and set it equal to zero, just like we did in the bivariate case

$$\frac{\partial \mathbf{u}^\top \mathbf{u}}{\partial \hat{\beta}} = -2\mathbf{X}^\top \mathbf{y} + 2\mathbf{X}^\top \mathbf{X} \hat{\beta} = 0$$
$$(\mathbf{X}^\top \mathbf{X}) \hat{\beta} = \mathbf{X}^\top \mathbf{y}$$

- To solve for  $\hat{\beta}$ , we need to pre-multiply both the left and the right by the inverse of  $(\mathbf{X}^\top \mathbf{X})$ , assuming it exists

$$\begin{aligned}(\mathbf{X}^\top \mathbf{X}) \hat{\beta} &= \mathbf{X}^\top \mathbf{y} \\(\mathbf{X}^\top \mathbf{X})^{-1} (\mathbf{X}^\top \mathbf{X}) \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ \mathbf{I} \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ \hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}\end{aligned}$$

- **Welcome to the matrix definition of the OLS estimator!**

# Unbiasedness

- Is this unbiased?
- To start, let's fiddle with the preceding definition of  $\hat{\beta}$  a little bit by replacing  $\mathbf{y}$  with  $\mathbf{X}\beta + \mathbf{u}$ .
  - Note that this requires **Assumption 1**: that the population model can be written as  $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$

$$\begin{aligned}\hat{\beta} &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X}\beta + \mathbf{u}) \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X}\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \\ &= \mathbf{I}\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \\ &= \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}\end{aligned}$$

# Unbiasedness

- Now let's invoke **Assumption 2** that these observations are drawn from an i.i.d. random sample, allowing us take expectations

$$\begin{aligned} E(\hat{\beta}) &= E\left[\beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}\right] \\ &= E(\beta) + E\left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}\right] \\ &= \beta + E\left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}\right] \end{aligned}$$

- Note that this requires  $(\mathbf{X}^\top \mathbf{X})^{-1}$  to exist, so we'll invoke **Assumption 3**: there is no perfect multicollinearity among our  $X$  values
  - *Compare this to the non-zero variance when we were working with scalars in the bivariate case*

# Unbiasedness

- Finally, let's invoke our most heroic assumption **Assumption 4**:  $E(\mathbf{u}|\mathbf{X}) = \mathbf{0}$ , and then rely on the law of iterated expectations (LIE)

$$\begin{aligned} E(\hat{\beta} \mid \mathbf{X}) &= \beta + E\left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mid \mathbf{X}\right] \\ &= \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E(\mathbf{u} \mid \mathbf{X}) \\ &= \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{0} \\ &= \beta \end{aligned}$$

# Properties of the OLS Estimators

- $\mathbf{X}^\top \mathbf{u} = 0$ : To prove, substitute the definition of  $\mathbf{y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{u}$  into the normal equation

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top \mathbf{y}$$

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\top (\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{u})$$

$$(\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})\hat{\boldsymbol{\beta}} + \mathbf{X}^\top \mathbf{u}$$

$$0 = \mathbf{X}^\top \mathbf{u}$$

# Properties of the OLS Estimators

- If our regression specification includes a constant,  $\sum u_i = 0$ : To prove, look inside the matrices!

$$\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{k1} & x_{k2} & \dots & x_{kn} \end{bmatrix} \cdot \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} x_{11} * u_1 + x_{12} * u_2 + \dots + x_{1n} * u_n \\ x_{21} * u_1 + x_{22} * u_2 + \dots + x_{2n} * u_n \\ \vdots \\ x_{k1} * u_1 + x_{k2} * u_2 + \dots + x_{kn} * u_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

- If  $\mathbf{X}^\top \mathbf{u} = \mathbf{0}$ , then every column  $\mathbf{x}_k$ 's dot product with  $\mathbf{u}$  must be zero
- Since the first column of  $\mathbf{X}$  is all 1, then this first column reduces to  $\sum u_i = 0$
- Also note that therefore  $\bar{u} = 0$  since  $\bar{u} = \frac{\sum u_i}{n}$



# Properties of the OLS Estimators

- The regression **hyperplane** (no longer a single line, since we have multiple predictors) will pass through  $\bar{X}$  and  $\bar{y}$ 
  - We just showed that  $\bar{u} = 0$ , and we know that  $u = y - X\hat{\beta}$
  - Thus  $\bar{u} = \bar{y} - \bar{x}\hat{\beta}$ , meaning  $\bar{y} = \bar{x}\hat{\beta}$
- The predicted values of  $y$  are uncorrelated with the residuals
  - $\hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$ , meaning that

$$\begin{aligned}\hat{\mathbf{y}}^\top \mathbf{u} &= \mathbf{X}\hat{\beta}^\top \mathbf{u} \\ &= \hat{\beta}^\top \mathbf{X}^\top \mathbf{u} \\ &= \hat{\beta}^\top \cdot \mathbf{0}\end{aligned}$$

# Errors

- The preceding has brought us to the statement that our OLS estimator is unbiased...but is it the "best"?
- We require **Assumption 5**:  $E(\mathbf{u}\mathbf{u}^\top \mid \mathbf{X}) = \sigma^2 \mathbf{I}$ . AKA: "spherical errors"
- Let's write it out:

$$\begin{aligned} E(\mathbf{u}\mathbf{u}^\top \mid \mathbf{X}) &= E\left( \begin{bmatrix} u_1 \mid \mathbf{X} \\ u_2 \mid \mathbf{X} \\ \vdots \\ u_n \mid \mathbf{X} \end{bmatrix} \begin{bmatrix} u_1 \mid \mathbf{X} & u_2 \mid \mathbf{X} & \dots & u_n \mid \mathbf{X} \end{bmatrix} \right) \\ &= E \begin{bmatrix} u_1^2 \mid \mathbf{X} & u_1 u_2 \mid \mathbf{X} & \dots & u_1 u_n \mid \mathbf{X} \\ u_2 u_1 \mid \mathbf{X} & u_2^2 \mid \mathbf{X} & \dots & u_2 u_n \mid \mathbf{X} \\ \vdots & \vdots & \ddots & \vdots \\ u_n u_1 \mid \mathbf{X} & u_n u_2 \mid \mathbf{X} & \dots & u_n^2 \mid \mathbf{X} \end{bmatrix} \end{aligned}$$

# Errors

- Distribute expectations to get:

$$E(\mathbf{u}\mathbf{u}^\top \mid \mathbf{X}) = \begin{bmatrix} E(u_1^2 \mid \mathbf{X}) & E(u_1 u_2 \mid \mathbf{X}) & \dots & E(u_1 u_n \mid \mathbf{X}) \\ E(u_2 u_1 \mid \mathbf{X}) & E(u_2^2 \mid \mathbf{X}) & \dots & E(u_2 u_n \mid \mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_n u_1 \mid \mathbf{X}) & E(u_n u_2 \mid \mathbf{X}) & \dots & E(u_n^2 \mid \mathbf{X}) \end{bmatrix}$$

- From **Assumption 5**:
  - Homoskedasticity states that the variance of  $u_i = \sigma^2$  for all  $i$ , or  $VAR(u_i \mid \mathbf{X}) = \sigma^2 \quad \forall i$
  - No autocorrelation states that  $cov(u_i, u_j \mid \mathbf{X}) = 0$

# Errors

- Thus, assumption 5 allows us to re-write:

$$\begin{aligned} E(\mathbf{u}\mathbf{u}^\top \mid \mathbf{X}) &= \begin{bmatrix} E(u_1^2 \mid \mathbf{X}) & E(u_1 u_2 \mid \mathbf{X}) & \dots & E(u_1 u_n \mid \mathbf{X}) \\ E(u_2 u_1 \mid \mathbf{X}) & E(u_2^2 \mid \mathbf{X}) & \dots & E(u_2 u_n \mid \mathbf{X}) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_n u_1 \mid \mathbf{X}) & E(u_n u_2 \mid \mathbf{X}) & \dots & E(u_n^2 \mid \mathbf{X}) \end{bmatrix} \\ &= \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} \end{aligned}$$

- which is the same as writing  $\sigma^2 \mathbf{I}$

# Errors

- Finally, let's calculate the variance of our OLS estimators,  $\hat{\beta}$
- But first, note that the variance of a **vector** is expressed as the covariance:  $E[(\hat{\beta} - E(\hat{\beta}))(\hat{\beta} - E(\hat{\beta}))^\top]$ 
  - Further note that we have already demonstrated that  $E(\hat{\beta}) = \beta$
- Also note that  $\hat{\beta} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}$ , or  $\hat{\beta} - \beta = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}$
- Plug in

$$\begin{aligned} E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top] &= E \left[ \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \right) \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \right)^\top \right] \\ &= E \left[ \left( (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \right) \left( \mathbf{u}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \right) \right] \\ &= E \left[ (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mathbf{u}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \right] \end{aligned}$$

# Errors

- Take the LIE conditional on  $\mathbf{X}$  to get

$$\begin{aligned} E[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top \mid \mathbf{X}] &= E\left[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mathbf{u}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \mid \mathbf{X}\right] \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top E(\mathbf{u} \mathbf{u}^\top \mid \mathbf{X}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\sigma^2 \mathbf{I}) \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma^2 \mathbf{I} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma^2 \mathbf{I} (\mathbf{X}^\top \mathbf{X})^{-1} \\ &= \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \end{aligned}$$

- In practice, we estimate the unknown  $\sigma^2$  with  $\hat{\sigma}^2 = \frac{\mathbf{u}^\top \mathbf{u}}{n-k}$