Week 7: Multiple Regression

Brandon Stewart[1]

Princeton

October 24, 26, 2016

[1]These slides are heavily influenced by Matt Blackwell, Adam Glynn, Jens Hainmueller and Danny Hidalgo.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - matrix form of linear regression

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - matrix form of linear regression
  - Wednesday:

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - matrix form of linear regression
  - Wednesday:
    - hypothesis tests

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - matrix form of linear regression
  - Wednesday:
    - hypothesis tests
- Next Week
  - break!

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - matrix form of linear regression
  - Wednesday:
    - hypothesis tests
- Next Week
  - break!
  - then ... regression in social science
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression

Questions?

Why Matrices and Vectors?

Why Matrices and Vectors?

Here's one way to write the full multiple regression model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + u_i$$

Why Matrices and Vectors?

Here's one way to write the full multiple regression model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + u_i$$

- Notation is going to get needlessly messy as we add variables

Why Matrices and Vectors?

Here's one way to write the full multiple regression model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + u_i$$

- Notation is going to get needlessly messy as we add variables
- Matrices are clean, but they are like a foreign language

Why Matrices and Vectors?

Here's one way to write the full multiple regression model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + u_i$$

- Notation is going to get needlessly messy as we add variables
- Matrices are clean, but they are like a foreign language
- You need to build intuitions over a long period of time (and they will return in Soc504)

Why Matrices and Vectors?

Here's one way to write the full multiple regression model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + u_i$$

- Notation is going to get needlessly messy as we add variables
- Matrices are clean, but they are like a foreign language
- You need to build intuitions over a long period of time (and they will return in Soc504)
- Reminder of Parameter Interpretation:
  β_1 is the effect of a one-unit change in x_{i1} conditional on all other x_{ik}.

Why Matrices and Vectors?

Here's one way to write the full multiple regression model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + u_i$$

- Notation is going to get needlessly messy as we add variables
- Matrices are clean, but they are like a foreign language
- You need to build intuitions over a long period of time (and they will return in Soc504)
- Reminder of Parameter Interpretation:
  β_1 is the effect of a one-unit change in x_{i1} conditional on all other x_{ik}.

Why Matrices and Vectors?

Here's one way to write the full multiple regression model:

$$y_i = \beta_0 + x_{i1}\beta_1 + x_{i2}\beta_2 + \cdots + x_{iK}\beta_K + u_i$$

- Notation is going to get needlessly messy as we add variables
- Matrices are clean, but they are like a foreign language
- You need to build intuitions over a long period of time (and they will return in Soc504)
- Reminder of Parameter Interpretation:
  β_1 is the effect of a one-unit change in x_{i1} conditional on all other x_{ik}.

We are going to review the key points quite quickly just to refresh the basics.

Matrices and Vectors

- A matrix is just a rectangular array of numbers.

Matrices and Vectors

- A matrix is just a rectangular array of numbers.
- We say that a matrix is $n \times K$ ("n by K") if it has n rows and K columns.

Matrices and Vectors

- A matrix is just a rectangular array of numbers.
- We say that a matrix is $n \times K$ ("n by K") if it has n rows and K columns.
- Uppercase bold denotes a matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nK} \end{bmatrix}$$

Matrices and Vectors

- A matrix is just a rectangular array of numbers.
- We say that a matrix is $n \times K$ ("n by K") if it has n rows and K columns.
- Uppercase bold denotes a matrix:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1K} \\ a_{21} & a_{22} & \cdots & a_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nK} \end{bmatrix}$$

- Generic entry: a_{ik} where this is the entry in row i and column k

Design Matrix

Design Matrix

One example of a matrix that we'll use a lot is the design matrix, which has a column of ones, and then each of the subsequent columns is each independent variable in the regression.

Design Matrix

One example of a matrix that we'll use a lot is the design matrix, which has a column of ones, and then each of the subsequent columns is each independent variable in the regression.

$$\mathbf{X} = \begin{bmatrix} 1 & \text{exports}_1 & \text{age}_1 & \text{male}_1 \\ 1 & \text{exports}_2 & \text{age}_2 & \text{male}_2 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \text{exports}_n & \text{age}_n & \text{male}_n \end{bmatrix}$$

Vectors

Vectors

- A vector is just a matrix with only one row or one column.

Vectors

- A vector is just a matrix with only one row or one column.
- A row vector is a vector with only one row, sometimes called a $1 \times K$ vector:

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_K \end{bmatrix}$$

Vectors

- A vector is just a matrix with only one row or one column.
- A row vector is a vector with only one row, sometimes called a $1 \times K$ vector:

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_K \end{bmatrix}$$

- A column vector is a vector with one column and more than one row. Here is a $n \times 1$ vector:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Vectors

- A vector is just a matrix with only one row or one column.
- A row vector is a vector with only one row, sometimes called a $1 \times K$ vector:

$$\boldsymbol{\alpha} = \left[\begin{array}{ccccc} \alpha_1 & \alpha_2 & \alpha_3 & \cdots & \alpha_K \end{array} \right]$$

- A column vector is a vector with one column and more than one row. Here is a $n \times 1$ vector:

$$\mathbf{y} = \left[\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array} \right]$$

- Convention: we'll assume that a vector is column vector and vectors will be written with lowercase bold lettering (**b**)

Vector Examples

Vector Examples

One common vector that we will work with are individual variables, such as the dependent variable, which we will represent as **y**:

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Transpose

Transpose

- There are many operations we'll do on vectors and matrices, but one is very fundamental: the transpose.

Transpose

- There are many operations we'll do on vectors and matrices, but one is very fundamental: the transpose.
- The transpose of a matrix $\mathbf{A}$ is the matrix created by switching the rows and columns of the data and is denoted $\mathbf{A}'$. That is, the kth column becomes the kth row.

Transpose

- There are many operations we'll do on vectors and matrices, but one is very fundamental: the transpose.
- The transpose of a matrix $\mathbf{A}$ is the matrix created by switching the rows and columns of the data and is denoted $\mathbf{A}'$. That is, the kth column becomes the kth row.

$$\mathbf{Q} = \begin{bmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \\ q_{31} & q_{32} \end{bmatrix} \quad \mathbf{Q}' = \begin{bmatrix} q_{11} & q_{21} & q_{31} \\ q_{12} & q_{22} & q_{32} \end{bmatrix}$$

If $\mathbf{A}$ is $j \times k$, then $\mathbf{A}'$ will be $k \times j$.

Transposing Vectors

Transposing Vectors

Transposing will turn a $k \times 1$ column vector into a $1 \times k$ row vector and vice versa:

$$\boldsymbol{\omega} = \begin{bmatrix} 1 \\ 3 \\ 2 \\ -5 \end{bmatrix}$$

Transposing Vectors

Transposing will turn a $k \times 1$ column vector into a $1 \times k$ row vector and vice versa:

$$\boldsymbol{\omega} = \begin{bmatrix} 1 \\ 3 \\ 2 \\ -5 \end{bmatrix} \qquad \boldsymbol{\omega}' = \begin{bmatrix} 1 & 3 & 2 & -5 \end{bmatrix}$$

Addition and Subtraction

- To perform addition/subtraction the matrices/vectors need to be conformable, meaning that the dimensions have to be the same

Addition and Subtraction

- To perform addition/subtraction the matrices/vectors need to be conformable, meaning that the dimensions have to be the same
- Let $\mathbf{A}$ and $\mathbf{B}$ both be 2×2 matrices. Then, let $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where we add each cell together:

Addition and Subtraction

- To perform addition/subtraction the matrices/vectors need to be conformable, meaning that the dimensions have to be the same
- Let $\mathbf{A}$ and $\mathbf{B}$ both be 2×2 matrices. Then, let $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where we add each cell together:

Addition and Subtraction

- To perform addition/subtraction the matrices/vectors need to be conformable, meaning that the dimensions have to be the same
- Let $\mathbf{A}$ and $\mathbf{B}$ both be 2×2 matrices. Then, let $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where we add each cell together:

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}$$

Addition and Subtraction

- To perform addition/subtraction the matrices/vectors need to be conformable, meaning that the dimensions have to be the same
- Let $\mathbf{A}$ and $\mathbf{B}$ both be 2×2 matrices. Then, let $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where we add each cell together:

$$\begin{aligned}\mathbf{A} + \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}\end{aligned}$$

Addition and Subtraction

- To perform addition/subtraction the matrices/vectors need to be conformable, meaning that the dimensions have to be the same
- Let $\mathbf{A}$ and $\mathbf{B}$ both be 2×2 matrices. Then, let $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where we add each cell together:

$$\begin{aligned} \mathbf{A} + \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\ &= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix} \\ &= \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \end{aligned}$$

Addition and Subtraction

- To perform addition/subtraction the matrices/vectors need to be conformable, meaning that the dimensions have to be the same
- Let $\mathbf{A}$ and $\mathbf{B}$ both be 2×2 matrices. Then, let $\mathbf{C} = \mathbf{A} + \mathbf{B}$, where we add each cell together:

$$
\begin{aligned}
\mathbf{A} + \mathbf{B} &= \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} + \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \\
&= \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix} \\
&= \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \\
&= \mathbf{C}
\end{aligned}
$$

Scalar Multiplication

Scalar Multiplication

- A scalar is just a single number: you can think of it sort of like a 1 by 1 matrix.

Scalar Multiplication

- A scalar is just a single number: you can think of it sort of like a 1 by 1 matrix.
- When we multiply a scalar by a matrix, we just multiply each element/cell by that scalar:

Scalar Multiplication

- A scalar is just a single number: you can think of it sort of like a 1 by 1 matrix.
- When we multiply a scalar by a matrix, we just multiply each element/cell by that scalar:

$$\alpha \mathbf{A} = \alpha \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} \alpha \times a_{11} & \alpha \times a_{12} \\ \alpha \times a_{21} & \alpha \times a_{22} \end{bmatrix}$$

The Linear Model with New Notation

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

- Imagine we had an n of 4. We could write out each formula:

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

- Imagine we had an n of 4. We could write out each formula:

$$y_1 = \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)}$$

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

- Imagine we had an n of 4. We could write out each formula:

$$\begin{aligned} y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\ y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \end{aligned}$$

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

- Imagine we had an n of 4. We could write out each formula:

$$\begin{aligned} y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\ y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \\ y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)} \end{aligned}$$

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

- Imagine we had an n of 4. We could write out each formula:

$$\begin{aligned}
y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\
y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \\
y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)} \\
y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)}
\end{aligned}$$

The Linear Model with New Notation

- Remember that we wrote the linear model as the following for all $i \in [1, \ldots, n]$:

$$y_i = \beta_0 + x_i\beta_1 + z_i\beta_2 + u_i$$

- Imagine we had an n of 4. We could write out each formula:

$$\begin{aligned} y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\ y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \\ y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)} \\ y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)} \end{aligned}$$

The Linear Model with New Notation

$$
\begin{aligned}
y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\
y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \\
y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)} \\
y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)}
\end{aligned}
$$

The Linear Model with New Notation

$$
\begin{aligned}
y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)}\\
y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)}\\
y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)}\\
y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)}
\end{aligned}
$$

- We can write this as:

The Linear Model with New Notation

$$
\begin{aligned}
y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\
y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \\
y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)} \\
y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)}
\end{aligned}
$$

- We can write this as:

$$
\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}
$$

The Linear Model with New Notation

$$\begin{aligned} y_1 &= \beta_0 + x_1\beta_1 + z_1\beta_2 + u_1 \quad \text{(unit 1)} \\ y_2 &= \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad \text{(unit 2)} \\ y_3 &= \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad \text{(unit 3)} \\ y_4 &= \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad \text{(unit 4)} \end{aligned}$$

- We can write this as:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}$$

- Outcome is a linear combination of the the $\mathbf{x}$, $\mathbf{z}$, and $\mathbf{u}$ vectors

Grouping Things into Matrices

Grouping Things into Matrices

- Can we write this in a more compact form?

Grouping Things into Matrices

- Can we write this in a more compact form?
  Yes! Let **X** and β be the following:

Grouping Things into Matrices

- Can we write this in a more compact form?
  Yes! Let $\mathbf{X}$ and β be the following:

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ 1 & x_4 & z_4 \end{bmatrix}$$

Grouping Things into Matrices

- Can we write this in a more compact form?
  Yes! Let $\mathbf{X}$ and $\boldsymbol{\beta}$ be the following:

$$\underset{(4\times 3)}{\mathbf{X}} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ 1 & x_4 & z_4 \end{bmatrix} \quad \underset{(3\times 1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

Matrix multiplication by a vector

Matrix multiplication by a vector

- We can write this more compactly as a matrix (post-)multiplied by a vector:

Matrix multiplication by a vector

- We can write this more compactly as a matrix (post-)multiplied by a vector:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 = \mathbf{X}\boldsymbol{\beta}$$

Matrix multiplication by a vector

- We can write this more compactly as a matrix (post-)multiplied by a vector:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 = \mathbf{X}\boldsymbol{\beta}$$

- Multiplication of a matrix by a vector is just the linear combination of the columns of the matrix with the vector elements as weights/coefficients.

Matrix multiplication by a vector

- We can write this more compactly as a matrix (post-)multiplied by a vector:

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \beta_1 + \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} \beta_2 = \mathbf{X}\boldsymbol{\beta}$$

- Multiplication of a matrix by a vector is just the linear combination of the columns of the matrix with the vector elements as weights/coefficients.
- And the left-hand side here only uses scalars times vectors, which is easy!

General Matrix by Vector Multiplication

General Matrix by Vector Multiplication

- **A** is a $n \times K$ matrix

General Matrix by Vector Multiplication

- $\mathbf{A}$ is a $n \times K$ matrix
- $\mathbf{b}$ is a $K \times 1$ column vector

General Matrix by Vector Multiplication

- $\mathbf{A}$ is a $n \times K$ matrix
- $\mathbf{b}$ is a $K \times 1$ column vector
- Columns of $\mathbf{A}$ have to match rows of $\mathbf{b}$

General Matrix by Vector Multiplication

- $\mathbf{A}$ is a $n \times K$ matrix
- $\mathbf{b}$ is a $K \times 1$ column vector
- Columns of $\mathbf{A}$ have to match rows of $\mathbf{b}$
- Let $\mathbf{a}_k$ be the kth column of A. Then we can write:

General Matrix by Vector Multiplication

- **A** is a $n \times K$ matrix
- **b** is a $K \times 1$ column vector
- Columns of **A** have to match rows of **b**
- Let $\mathbf{a}_k$ be the kth column of A. Then we can write:

$$\underset{(j \times 1)}{\mathbf{c}} = \mathbf{Ab} = b_1\mathbf{a}_1 + b_2\mathbf{a}_2 + \cdots + b_K\mathbf{a}_K$$

General Matrix by Vector Multiplication

- $\mathbf{A}$ is a $n \times K$ matrix
- $\mathbf{b}$ is a $K \times 1$ column vector
- Columns of $\mathbf{A}$ have to match rows of $\mathbf{b}$
- Let $\mathbf{a}_k$ be the kth column of A. Then we can write:

$$\underset{(j \times 1)}{\mathbf{c}} = \mathbf{Ab} = b_1\mathbf{a}_1 + b_2\mathbf{a}_2 + \cdots + b_K\mathbf{a}_K$$

- $\mathbf{c}$ is linear combination of the columns of $\mathbf{A}$

Back to Regression

Back to Regression

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables

Back to Regression

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- β be the $(K+1) \times 1$ column vector of coefficients.

Back to Regression

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- $\boldsymbol{\beta}$ be the $(K+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\boldsymbol{\beta}$ will be $n \times 1$:

Back to Regression

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- $\boldsymbol{\beta}$ be the $(K+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\boldsymbol{\beta}$ will be $n \times 1$:

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_K \mathbf{x}_K$$

Back to Regression

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- $\boldsymbol{\beta}$ be the $(K+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\boldsymbol{\beta}$ will be $n \times 1$:

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_K \mathbf{x}_K$$

- We can compactly write the linear model as the following:

Back to Regression

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- $\boldsymbol{\beta}$ be the $(K+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\boldsymbol{\beta}$ will be $n \times 1$:

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1 \mathbf{x}_1 + \beta_2 \mathbf{x}_2 + \cdots + \beta_K \mathbf{x}_K$$

- We can compactly write the linear model as the following:

$$\underset{(n \times 1)}{\mathbf{y}} = \underset{(n \times 1)}{\mathbf{X}\boldsymbol{\beta}} + \underset{(n \times 1)}{\mathbf{u}}$$

Back to Regression

- $\mathbf{X}$ is the $n \times (K+1)$ design matrix of independent variables
- $\boldsymbol{\beta}$ be the $(K+1) \times 1$ column vector of coefficients.
- $\mathbf{X}\boldsymbol{\beta}$ will be $n \times 1$:

$$\mathbf{X}\boldsymbol{\beta} = \beta_0 + \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \cdots + \beta_K\mathbf{x}_K$$

- We can compactly write the linear model as the following:

$$\underset{(n\times 1)}{\mathbf{y}} = \underset{(n\times 1)}{\mathbf{X}\boldsymbol{\beta}} + \underset{(n\times 1)}{\mathbf{u}}$$

- We can also write this at the individual level, where $\mathbf{x}_i'$ is the ith row of $\mathbf{X}$:

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

Matrix Multiplication

- What if, instead of a column vector b, we have a matrix $\mathbf{B}$ with dimensions $K \times M$.

Matrix Multiplication

- What if, instead of a column vector b, we have a matrix $\mathbf{B}$ with dimensions $K \times M$.
- How do we do multiplication like so $\mathbf{C} = \mathbf{AB}$?

Matrix Multiplication

- What if, instead of a column vector b, we have a matrix $\mathbf{B}$ with dimensions $K \times M$.
- How do we do multiplication like so $\mathbf{C} = \mathbf{AB}$?
- Each column of the new matrix is just matrix by vector multiplication:

$$\mathbf{C} = [\mathbf{c}_1 \quad \mathbf{c}_2 \quad \cdots \quad \mathbf{c}_M] \qquad \mathbf{c}_k = \mathbf{A}\mathbf{b}_k$$

Matrix Multiplication

- What if, instead of a column vector b, we have a matrix $\mathbf{B}$ with dimensions $K \times M$.
- How do we do multiplication like so $\mathbf{C} = \mathbf{AB}$?
- Each column of the new matrix is just matrix by vector multiplication:

$$\mathbf{C} = [\mathbf{c}_1 \quad \mathbf{c}_2 \quad \cdots \quad \mathbf{c}_M] \qquad \mathbf{c}_k = \mathbf{A}\mathbf{b}_k$$

- Thus, each column of $\mathbf{C}$ is a linear combination of the columns of $\mathbf{A}$.

Special Multiplications

- The inner product of a two column vectors **a** and **b** (of equal dimension, $K \times 1$):

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_K b_K$$

Special Multiplications

- The inner product of a two column vectors **a** and **b** (of equal dimension, $K \times 1$):

$$\mathbf{a}'\mathbf{b} = a_1 b_1 + a_2 b_2 + \cdots + a_K b_K$$

- Special case of above: $\mathbf{a}'$ is a matrix with K columns and just 1 row, so the "columns" of $\mathbf{a}'$ are just scalars.

Sum of the Squared Residuals

Sum of the Squared Residuals

- Example: let's say that we have a vector of residuals, $\widehat{\mathbf{u}}$, then the inner product of the residuals is:

Sum of the Squared Residuals

- Example: let's say that we have a vector of residuals, $\widehat{\mathbf{u}}$, then the inner product of the residuals is:

$$\hat{\mathbf{u}}'\hat{\mathbf{u}}$$

Sum of the Squared Residuals

- Example: let's say that we have a vector of residuals, $\widehat{\mathbf{u}}$, then the inner product of the residuals is:

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \begin{bmatrix} \widehat{u}_1 & \widehat{u}_2 & \cdots & \widehat{u}_n \end{bmatrix} \begin{bmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \vdots \\ \widehat{u}_n \end{bmatrix}$$

Sum of the Squared Residuals

- Example: let's say that we have a vector of residuals, $\widehat{\mathbf{u}}$, then the inner product of the residuals is:

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \begin{bmatrix} \widehat{u}_1 & \widehat{u}_2 & \cdots & \widehat{u}_n \end{bmatrix} \begin{bmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \vdots \\ \widehat{u}_n \end{bmatrix}$$

$$\hat{\mathbf{u}}'\hat{\mathbf{u}} = \widehat{u}_1\widehat{u}_1 + \widehat{u}_2\widehat{u}_2 + \cdots + \widehat{u}_n\widehat{u}_n = \sum_{i=1}^{n} \widehat{u}_i^2$$

- It's just the sum of the squared residuals!

Square Matrices and the Diagonal

Square Matrices and the Diagonal

- A square matrix has equal numbers of rows and columns.

Square Matrices and the Diagonal

- A square matrix has equal numbers of rows and columns.
- The diagonal of a square matrix are the values a_{jj}:

Square Matrices and the Diagonal

- A square matrix has equal numbers of rows and columns.
- The diagonal of a square matrix are the values a_{jj}:

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

Square Matrices and the Diagonal

- A square matrix has equal numbers of rows and columns.
- The diagonal of a square matrix are the values a_{jj}:

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

- The identity matrix, $\mathbf{I}$ is a square matrix, with 1s along the diagonal and 0s everywhere else.

Square Matrices and the Diagonal

- A square matrix has equal numbers of rows and columns.
- The diagonal of a square matrix are the values a_{jj}:

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

- The identity matrix, $\mathbf{I}$ is a square matrix, with 1s along the diagonal and 0s everywhere else.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Square Matrices and the Diagonal

- A square matrix has equal numbers of rows and columns.
- The diagonal of a square matrix are the values a_{jj}:

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

- The identity matrix, $\mathbf{I}$ is a square matrix, with 1s along the diagonal and 0s everywhere else.

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

- The identity matrix multiplied by any matrix returns the matrix: $\mathbf{AI} = \mathbf{A}$.

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix}$$

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} \qquad \widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} \qquad \widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- It might be helpful to see this again more written out:

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} \qquad \widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- It might be helpful to see this again more written out:

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix}$$

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} \qquad \widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- It might be helpful to see this again more written out:

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \mathbf{X}\widehat{\boldsymbol{\beta}} =$$

Multiple Linear Regression in Matrix Form

- Let $\widehat{\boldsymbol{\beta}}$ be the matrix of estimated regression coefficients and $\widehat{\mathbf{y}}$ be the vector of fitted values:

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix} \qquad \widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- It might be helpful to see this again more written out:

$$\widehat{\mathbf{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix} = \mathbf{X}\widehat{\boldsymbol{\beta}} = \begin{bmatrix} 1\widehat{\beta}_0 + x_{11}\widehat{\beta}_1 + x_{12}\widehat{\beta}_2 + \cdots + x_{1K}\widehat{\beta}_K \\ 1\widehat{\beta}_0 + x_{21}\widehat{\beta}_1 + x_{22}\widehat{\beta}_2 + \cdots + x_{2K}\widehat{\beta}_K \\ \vdots \\ 1\widehat{\beta}_0 + x_{n1}\widehat{\beta}_1 + x_{n2}\widehat{\beta}_2 + \cdots + x_{nK}\widehat{\beta}_K \end{bmatrix}$$

Residuals

Residuals

- We can easily write the residuals in matrix form:

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

Residuals

- We can easily write the residuals in matrix form:

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- Our goal as usual is to minimize the sum of the squared residuals, which we saw earlier we can write:

Residuals

- We can easily write the residuals in matrix form:

$$\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\beta}$$

- Our goal as usual is to minimize the sum of the squared residuals, which we saw earlier we can write:

$$\widehat{\mathbf{u}}'\widehat{\mathbf{u}} = (\mathbf{y} - \mathbf{X}\widehat{\beta})'(\mathbf{y} - \mathbf{X}\widehat{\beta})$$

OLS Estimator in Matrix Form

OLS Estimator in Matrix Form

- Goal: minimize the sum of the squared residuals

OLS Estimator in Matrix Form

- Goal: minimize the sum of the squared residuals
- Take (matrix) derivatives, set equal to 0

OLS Estimator in Matrix Form

- Goal: minimize the sum of the squared residuals
- Take (matrix) derivatives, set equal to 0
- Resulting first order conditions:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\beta}) = 0$$

OLS Estimator in Matrix Form

- Goal: minimize the sum of the squared residuals
- Take (matrix) derivatives, set equal to 0
- Resulting first order conditions:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\beta}) = 0$$

- Rearranging:

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{y}$$

OLS Estimator in Matrix Form

- Goal: minimize the sum of the squared residuals
- Take (matrix) derivatives, set equal to 0
- Resulting first order conditions:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0$$

- Rearranging:

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

- In order to isolate $\widehat{\boldsymbol{\beta}}$, we need to move the $\mathbf{X}'\mathbf{X}$ term to the other side of the equals sign.

OLS Estimator in Matrix Form

- Goal: minimize the sum of the squared residuals
- Take (matrix) derivatives, set equal to 0
- Resulting first order conditions:

$$\mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = 0$$

- Rearranging:

$$\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$$

- In order to isolate $\widehat{\boldsymbol{\beta}}$, we need to move the $\mathbf{X}'\mathbf{X}$ term to the other side of the equals sign.
- We've learned about matrix multiplication, but what about matrix "division"?

Scalar Inverses

- What is division in its simplest form? $\frac{1}{a}$ is the value such that $a\frac{1}{a} = 1$:

Scalar Inverses

- What is division in its simplest form? $\frac{1}{a}$ is the value such that $a\frac{1}{a} = 1$:
- For some algebraic expression: $au = b$, let's solve for u:

$$\frac{1}{a}au = \frac{1}{a}b$$

Scalar Inverses

- What is division in its simplest form? $\frac{1}{a}$ is the value such that $a\frac{1}{a} = 1$:
- For some algebraic expression: $au = b$, let's solve for u:

$$\frac{1}{a}au = \frac{1}{a}b$$

Scalar Inverses

- What is division in its simplest form? $\frac{1}{a}$ is the value such that $a\frac{1}{a} = 1$:
- For some algebraic expression: $au = b$, let's solve for u:

$$\frac{1}{a}au = \frac{1}{a}b$$
$$u = \frac{b}{a}$$

Scalar Inverses

- What is division in its simplest form? $\frac{1}{a}$ is the value such that $a\frac{1}{a} = 1$:
- For some algebraic expression: $au = b$, let's solve for u:

$$\frac{1}{a}au = \frac{1}{a}b$$
$$u = \frac{b}{a}$$

- Need a matrix version of this: $\frac{1}{a}$.

Matrix Inverses

Definition (Matrix Inverse)

If it exists, the **inverse** of square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

- We can use the inverse to solve (systems of) equations:

$$\mathbf{Au} = \mathbf{b}$$

Matrix Inverses

Definition (Matrix Inverse)

If it exists, the **inverse** of square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

- We can use the inverse to solve (systems of) equations:

$$\mathbf{Au} = \mathbf{b}$$
$$\mathbf{A}^{-1}\mathbf{Au} = \mathbf{A}^{-1}\mathbf{b}$$

Matrix Inverses

Definition (Matrix Inverse)

If it exists, the **inverse** of square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

- We can use the inverse to solve (systems of) equations:

$$\begin{aligned} \mathbf{A}\mathbf{u} &= \mathbf{b} \\ \mathbf{A}^{-1}\mathbf{A}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{I}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$

Matrix Inverses

Definition (Matrix Inverse)

If it exists, the **inverse** of square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

- We can use the inverse to solve (systems of) equations:

$$\begin{aligned} \mathbf{A}\mathbf{u} &= \mathbf{b} \\ \mathbf{A}^{-1}\mathbf{A}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{I}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$

Matrix Inverses

Definition (Matrix Inverse)

If it exists, the **inverse** of square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

- We can use the inverse to solve (systems of) equations:

$$\begin{aligned} \mathbf{A}\mathbf{u} &= \mathbf{b} \\ \mathbf{A}^{-1}\mathbf{A}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{I}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$

- If the inverse exists, we say that $\mathbf{A}$ is invertible or nonsingular.

Matrix Inverses

Definition (Matrix Inverse)

If it exists, the **inverse** of square matrix $\mathbf{A}$, denoted $\mathbf{A}^{-1}$, is the matrix such that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$.

- We can use the inverse to solve (systems of) equations:

$$\begin{aligned} \mathbf{A}\mathbf{u} &= \mathbf{b} \\ \mathbf{A}^{-1}\mathbf{A}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{I}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$

- If the inverse exists, we say that $\mathbf{A}$ is invertible or nonsingular.

Back to OLS

Back to OLS

- Let's assume, for now, that the inverse of $\mathbf{X}'\mathbf{X}$ exists

Back to OLS

- Let's assume, for now, that the inverse of $\mathbf{X}'\mathbf{X}$ exists
- Then we can write the OLS estimator as the following:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Back to OLS

- Let's assume, for now, that the inverse of $\mathbf{X}'\mathbf{X}$ exists
- Then we can write the OLS estimator as the following:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Memorize this: "ex prime ex inverse ex prime y"

Back to OLS

- Let's assume, for now, that the inverse of $\mathbf{X}'\mathbf{X}$ exists
- Then we can write the OLS estimator as the following:

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Memorize this: "ex prime ex inverse ex prime y" sear it into your soul.

Intuition for the OLS in Matrix Form

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Intuition for the OLS in Matrix Form

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?

Intuition for the OLS in Matrix Form

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?
- "Numerator" $\mathbf{X}'\mathbf{y}$: is roughly composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$

Intuition for the OLS in Matrix Form

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?
- "Numerator" $\mathbf{X}'\mathbf{y}$: is roughly composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is roughly composed of the sample variances and covariances of variables within $\mathbf{X}$

Intuition for the OLS in Matrix Form

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?
- "Numerator" $\mathbf{X}'\mathbf{y}$: is roughly composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is roughly composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

$$\widehat{\boldsymbol{\beta}} \approx (\text{variance of } \mathbf{X})^{-1}(\text{covariance of } \mathbf{X} \ \& \ \mathbf{y})$$

Intuition for the OLS in Matrix Form

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- What's the intuition here?
- "Numerator" $\mathbf{X}'\mathbf{y}$: is roughly composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is roughly composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

$$\widehat{\boldsymbol{\beta}} \approx (\text{variance of } \mathbf{X})^{-1}(\text{covariance of } \mathbf{X} \;\&\; \mathbf{y})$$

- This is a rough sketch and isn't strictly true, but it can provide intuition.

General OLS Assumptions

General OLS Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$

General OLS Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.

General OLS Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$

General OLS Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$

General OLS Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$

General OLS Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

No Perfect Collinearity

No Perfect Collinearity

Definition (Rank)

The rank of a matrix is the maximum number of linearly independent columns.

No Perfect Collinearity

Definition (Rank)

The rank of a matrix is the maximum number of linearly independent columns.

- In matrix form: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$

No Perfect Collinearity

Definition (Rank)

The rank of a matrix is the maximum number of linearly independent columns.

- In matrix form: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
- If $\mathbf{X}$ has rank $K+1$, then all of its columns are linearly independent

No Perfect Collinearity

Definition (Rank)

The rank of a matrix is the maximum number of linearly independent columns.

- In matrix form: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
- If $\mathbf{X}$ has rank $K+1$, then all of its columns are linearly independent
- ...and none of its columns are linearly dependent $\implies$ no perfect collinearity

No Perfect Collinearity

Definition (Rank)

The rank of a matrix is the maximum number of linearly independent columns.

- In matrix form: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
- If $\mathbf{X}$ has rank $K+1$, then all of its columns are linearly independent
- ... and none of its columns are linearly dependent $\implies$ no perfect collinearity
- $\mathbf{X}$ has rank $K+1 \implies (\mathbf{X}'\mathbf{X})$ is invertible

No Perfect Collinearity

Definition (Rank)

The rank of a matrix is the maximum number of linearly independent columns.

- In matrix form: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
- If $\mathbf{X}$ has rank $K+1$, then all of its columns are linearly independent
- ...and none of its columns are linearly dependent $\implies$ no perfect collinearity
- $\mathbf{X}$ has rank $K+1 \implies (\mathbf{X}'\mathbf{X})$ is invertible
- Just like variation in X led us to be able to divide by the variance in simple OLS

Expected Values of Vectors

Expected Values of Vectors

- The expected value of the vector is just the expected value of its entries.

Expected Values of Vectors

- The expected value of the vector is just the expected value of its entries.
- Using the zero mean conditional error assumptions:

$$\mathbb{E}[\mathbf{u}|\mathbf{X}] = \begin{bmatrix} \mathbb{E}[u_1|\mathbf{X}] \\ \mathbb{E}[u_2|\mathbf{X}] \\ \vdots \\ \mathbb{E}[u_n|\mathbf{X}] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \mathbf{0}$$

OLS is Unbiased

Under matrix assumptions 1-4, OLS is unbiased for $\boldsymbol{\beta}$:

$$\mathbb{E}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$$

Unbiasedness of $\hat{\boldsymbol{\beta}}$

Is $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$?

Unbiasedness of $\hat{\boldsymbol{\beta}}$

Is $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$?

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)}$$

Unbiasedness of $\hat{\beta}$

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

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)}$$
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})$$

Unbiasedness of $\hat{\beta}$

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

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)}$$
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})$$
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

Unbiasedness of $\hat{\beta}$

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

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)}$$
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})$$
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$
$$\hat{\beta} = \mathbf{I}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

Unbiasedness of $\hat{\boldsymbol{\beta}}$

Is $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$?

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)}$$
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})$$
$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$
$$\hat{\boldsymbol{\beta}} = \mathbf{I}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$
$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

Unbiasedness of $\hat{\beta}$

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

$$\begin{aligned}
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)} \\
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \mathbf{I}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
E[\hat{\beta}|\mathbf{X}] &= E[\beta|\mathbf{X}] + E[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}]
\end{aligned}$$

Unbiasedness of $\hat{\beta}$

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

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)}$$
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})$$
$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$
$$\hat{\beta} = \mathbf{I}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$
$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$
$$E[\hat{\beta}|\mathbf{X}] = E[\beta|\mathbf{X}] + E[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}]$$
$$E[\hat{\beta}|\mathbf{X}] = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}]$$

Unbiasedness of $\hat{\beta}$

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

$$\begin{aligned}
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)} \\
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\
\hat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \mathbf{I}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u} \\
E[\hat{\beta}|\mathbf{X}] &= E[\beta|\mathbf{X}] + E[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
E[\hat{\beta}|\mathbf{X}] &= \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}] \\
E[\hat{\beta}|\mathbf{X}] &= \beta \text{ (zero conditional mean)}
\end{aligned}$$

Unbiasedness of $\hat{\beta}$

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

$$\begin{aligned}
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \text{ (linearity and no collinearity)} \\
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u}) \\
\hat{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \mathbf{I}\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\
\hat{\beta} &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\
E[\hat{\beta}|\mathbf{X}] &= E[\beta|\mathbf{X}] + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
E[\hat{\beta}|\mathbf{X}] &= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E[\mathbf{u}|\mathbf{X}] \\
E[\hat{\beta}|\mathbf{X}] &= \beta \text{ (zero conditional mean)}
\end{aligned}$$

So, yes!

A Much Shorter Proof of Unbiasedness of $\hat{\beta}$

A shorter but perhaps less informative proof of unbiasedness,

A Much Shorter Proof of Unbiasedness of $\hat{\boldsymbol{\beta}}$

A shorter but perhaps less informative proof of unbiasedness,

$$\begin{aligned} E[\hat{\boldsymbol{\beta}}] &= E[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}] \text{ (definition of the estimator)} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \text{ (expectation of y)} \\ &= \boldsymbol{\beta} \end{aligned}$$

Variance-Covariance Matrix

Variance-Covariance Matrix

- The homoskedasticity assumption is different: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2\mathbf{I}_n$

Variance-Covariance Matrix

- The homoskedasticity assumption is different: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- In order to investigate this, we need to know what the variance of a vector is.

Variance-Covariance Matrix

- The homoskedasticity assumption is different: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- In order to investigate this, we need to know what the variance of a vector is.
- The variance of a vector is actually a matrix:

$$\text{var}[\mathbf{u}] = \Sigma_u = \begin{bmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \dots & \text{cov}(u_1, u_n) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \dots & \text{cov}(u_2, u_n) \\ \vdots & & \ddots & \\ \text{cov}(u_n, u_1) & \text{cov}(u_n, u_2) & \dots & \text{var}(u_n) \end{bmatrix}$$

Variance-Covariance Matrix

- The homoskedasticity assumption is different: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- In order to investigate this, we need to know what the variance of a vector is.
- The variance of a vector is actually a matrix:

$$\text{var}[\mathbf{u}] = \Sigma_u = \begin{bmatrix} \text{var}(u_1) & \text{cov}(u_1, u_2) & \dots & \text{cov}(u_1, u_n) \\ \text{cov}(u_2, u_1) & \text{var}(u_2) & \dots & \text{cov}(u_2, u_n) \\ \vdots & & \ddots & \\ \text{cov}(u_n, u_1) & \text{cov}(u_n, u_2) & \dots & \text{var}(u_n) \end{bmatrix}$$

- This matrix is symmetric since $\text{cov}(u_i, u_j) = \text{cov}(u_j, u_i)$

Matrix Version of Homoskedasticity

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- $\mathbf{I}_n$ is the $n \times n$ identity matrix

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- $\mathbf{I}_n$ is the $n \times n$ identity matrix
- Visually:

$$\text{var}[\mathbf{u}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_u^2 & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \sigma_u^2 \end{bmatrix}$$

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- $\mathbf{I}_n$ is the $n \times n$ identity matrix
- Visually:

$$\text{var}[\mathbf{u}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_u^2 & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \sigma_u^2 \end{bmatrix}$$

- In less matrix notation:

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- $\mathbf{I}_n$ is the $n \times n$ identity matrix
- Visually:

$$\text{var}[\mathbf{u}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_u^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_u^2 \end{bmatrix}$$

- In less matrix notation:
  - $\text{var}(u_i) = \sigma_u^2$ for all i (constant variance)

Matrix Version of Homoskedasticity

- Once again: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- $\mathbf{I}_n$ is the $n \times n$ identity matrix
- Visually:

$$\text{var}[\mathbf{u}] = \sigma_u^2 \mathbf{I}_n = \begin{bmatrix} \sigma_u^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_u^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_u^2 \end{bmatrix}$$

- In less matrix notation:
  - $\text{var}(u_i) = \sigma_u^2$ for all i (constant variance)
  - $\text{cov}(u_i, u_j) = 0$ for all $i \neq j$ (implied by iid)

Sampling Variance for OLS Estimates

Sampling Variance for OLS Estimates

- Under assumptions 1-5, the sampling variance of the OLS estimator can be written in matrix form as the following:

$$\text{var}[\widehat{\beta}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$$

Sampling Variance for OLS Estimates

- Under assumptions 1-5, the sampling variance of the OLS estimator can be written in matrix form as the following:

$$\text{var}[\widehat{\beta}] = \sigma_u^2(\mathbf{X}'\mathbf{X})^{-1}$$

- This matrix looks like this:

Sampling Variance for OLS Estimates

- Under assumptions 1-5, the sampling variance of the OLS estimator can be written in matrix form as the following:

$$\text{var}[\widehat{\beta}] = \sigma_u^2 (\mathbf{X}'\mathbf{X})^{-1}$$

- This matrix looks like this:

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\cdots$	$\widehat{\beta}_K$
$\widehat{\beta}_0$	$\text{var}[\widehat{\beta}_0]$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_K]$
$\widehat{\beta}_1$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\text{var}[\widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_K]$
$\widehat{\beta}_2$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\text{cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\text{var}[\widehat{\beta}_2]$	$\cdots$	$\text{cov}[\widehat{\beta}_2, \widehat{\beta}_K]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\widehat{\beta}_K$	$\text{cov}[\widehat{\beta}_0, \widehat{\beta}_K]$	$\text{cov}[\widehat{\beta}_K, \widehat{\beta}_1]$	$\text{cov}[\widehat{\beta}_K, \widehat{\beta}_2]$	$\cdots$	$\text{var}[\widehat{\beta}_K]$

Sampling Distribution for $\widehat{\beta}_j$

Under the first four assumptions,

$$\hat{\beta}_j | X \sim N\left(\beta_j, SE(\hat{\beta}_j)^2\right)$$

$$SE(\hat{\beta}_j)^2 = \frac{1}{1 - R_j^2} \frac{\sigma_u^2}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

where R_j^2 is from the regression of x_j on all other explanatory variables.

Inference in the General Setting

- Under assumption 1-5 in large samples:

$$\frac{\widehat{\beta}_k - \beta_k}{\widehat{SE}[\widehat{\beta}_k]} \sim N(0, 1)$$

Inference in the General Setting

- Under assumption 1-5 in large samples:

$$\frac{\widehat{\beta}_k - \beta_k}{\widehat{SE}[\widehat{\beta}_k]} \sim N(0, 1)$$

- In small samples, under assumptions 1-6,

$$\frac{\widehat{\beta}_k - \beta_k}{\widehat{SE}[\widehat{\beta}_k]} \sim t_{n-(K+1)}$$

Inference in the General Setting

- Under assumption 1-5 in large samples:

$$\frac{\widehat{\beta}_k - \beta_k}{\widehat{SE}[\widehat{\beta}_k]} \sim N(0,1)$$

- In small samples, under assumptions 1-6,

$$\frac{\widehat{\beta}_k - \beta_k}{\widehat{SE}[\widehat{\beta}_k]} \sim t_{n-(K+1)}$$

- Thus, under the null of $H_0 : \beta_k = 0$, we know that

$$\frac{\widehat{\beta}_k}{\widehat{SE}[\widehat{\beta}_k]} \sim t_{n-(K+1)}$$

Inference in the General Setting

- Under assumption 1-5 in large samples:

$$\frac{\widehat{\beta}_k - \beta_k}{\widehat{SE}[\widehat{\beta}_k]} \sim N(0,1)$$

- In small samples, under assumptions 1-6,

$$\frac{\widehat{\beta}_k - \beta_k}{\widehat{SE}[\widehat{\beta}_k]} \sim t_{n-(K+1)}$$

- Thus, under the null of $H_0 : \beta_k = 0$, we know that

$$\frac{\widehat{\beta}_k}{\widehat{SE}[\widehat{\beta}_k]} \sim t_{n-(K+1)}$$

- Here, the estimated SEs come from:

$$\widehat{\text{var}}[\widehat{\boldsymbol{\beta}}] = \widehat{\sigma}_u^2(\mathbf{X}'\mathbf{X})^{-1}$$
$$\widehat{\sigma}_u^2 = \frac{\widehat{\mathbf{u}}'\widehat{\mathbf{u}}}{n-(k+1)}$$

Properties of the OLS Estimator: Summary

Properties of the OLS Estimator: Summary

Theorem

Under Assumptions 1–6, the $(k+1) \times 1$ vector of OLS estimators $\hat{\boldsymbol{\beta}}$, conditional on $\mathbf{X}$, follows a multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$:

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

Properties of the OLS Estimator: Summary

Theorem

Under Assumptions 1–6, the $(k+1) \times 1$ *vector of OLS estimators* $\hat{\boldsymbol{\beta}}$*, conditional on* $\mathbf{X}$*, follows a multivariate normal distribution with mean* $\boldsymbol{\beta}$ *and variance-covariance matrix* $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$*:*

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

- *Each element of* $\hat{\boldsymbol{\beta}}$ *(i.e.* $\hat{\beta}_0, ..., \hat{\beta}_{k+1}$*) is normally distributed, and* $\hat{\boldsymbol{\beta}}$ *is an unbiased estimator of* $\boldsymbol{\beta}$ *as* $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$

Properties of the OLS Estimator: Summary

Theorem

Under Assumptions 1–6, the $(k+1) \times 1$ *vector of OLS estimators* $\hat{\boldsymbol{\beta}}$*, conditional on* $\mathbf{X}$*, follows a multivariate normal distribution with mean* $\boldsymbol{\beta}$ *and variance-covariance matrix* $\sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$*:*

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right)$$

- *Each element of* $\hat{\boldsymbol{\beta}}$ *(i.e.* $\hat{\beta}_0, ..., \hat{\beta}_{k+1}$*) is normally distributed, and* $\hat{\boldsymbol{\beta}}$ *is an unbiased estimator of* $\boldsymbol{\beta}$ *as* $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$
- *Variances and covariances are given by* $V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1}$

Properties of the OLS Estimator: Summary

Theorem

Under Assumptions 1–6, the $(k+1) \times 1$ vector of OLS estimators $\hat{\boldsymbol{\beta}}$, conditional on $\mathbf{X}$, follows a multivariate normal distribution *with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$:*

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

- *Each element of $\hat{\boldsymbol{\beta}}$ (i.e. $\hat{\beta}_0, ..., \hat{\beta}_{k+1}$) is normally distributed, and $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ as $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$*
- *Variances and covariances are given by $V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$*
- *An unbiased estimator for the error variance σ^2 is given by*

$$\widehat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

Properties of the OLS Estimator: Summary

Theorem

Under Assumptions 1–6, the $(k+1) \times 1$ vector of OLS estimators $\hat{\boldsymbol{\beta}}$, conditional on $\mathbf{X}$, follows a multivariate normal distribution with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$:

$$\hat{\boldsymbol{\beta}}|\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}\right)$$

- *Each element of $\hat{\boldsymbol{\beta}}$ (i.e. $\hat{\beta}_0, ..., \hat{\beta}_{k+1}$) is normally distributed, and $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$ as $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$*
- *Variances and covariances are given by $V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$*
- *An unbiased estimator for the error variance σ^2 is given by*

$$\widehat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

- *With a large sample, $\hat{\boldsymbol{\beta}}$ approximately follows the same distribution under Assumptions 1–5 only, i.e., without assuming the normality of $\mathbf{u}$.*

Implications of the Variance-Covariance Matrix

Implications of the Variance-Covariance Matrix

- Note that the sampling distribution of $\hat{\beta}$ has covariance terms

Implications of the Variance-Covariance Matrix

- Note that the sampling distribution of $\hat{\beta}$ has covariance terms
- In a practical sense, this means that our uncertainty about coefficients is correlated across variables.

Implications of the Variance-Covariance Matrix

- Note that the sampling distribution of $\hat{\beta}$ has covariance terms
- In a practical sense, this means that our uncertainty about coefficients is correlated across variables.
- Let's go to the board and discuss!

Motivation for the Bootstrap

Sometimes it is hard to calculate the sampling distribution.

Motivation for the Bootstrap

Sometimes it is hard to calculate the sampling distribution.

Bootstrapping provides an alternative way to calculate the sampling distribution of a function of a sample when that function is smooth.

Motivation for the Bootstrap

Sometimes it is hard to calculate the sampling distribution.

Bootstrapping provides an alternative way to calculate the sampling distribution of a function of a sample when that function is smooth.

Let's work through an example.

Sample

Suppose that a week before the 2012 election, you contacted a sample of $n = 625$ potential Florida voters, randomly selected (with replacement) from the population of $N = 11,900,000$ on the public voters register, to ask whether they planned to vote.

Sample

Suppose that a week before the 2012 election, you contacted a sample of $n = 625$ potential Florida voters, randomly selected (with replacement) from the population of $N = 11,900,000$ on the public voters register, to ask whether they planned to vote.

Suppose also,

- voters register is completely up to date
- all potential voters can be contacted, will respond honestly to your questions, and will not change their minds about voting

Sample

Suppose that a week before the 2012 election, you contacted a sample of $n = 625$ potential Florida voters, randomly selected (with replacement) from the population of $N = 11,900,000$ on the public voters register, to ask whether they planned to vote.

Suppose also,

- voters register is completely up to date
- all potential voters can be contacted, will respond honestly to your questions, and will not change their minds about voting

Table: Sample

i	1	2	3	4	...	625	$\bar{y}_{625}$
y_i	1	1	0	1	...	0	.68

Sample versus Population

Table: Sample

i	1	2	3	4	...	625	$\bar{y}_{625}$
y_i	1	1	0	1	...	0	.68

Sample versus Population

Table: Sample

i	1	2	3	4	...	625	$\bar{y}_{625}$
y_i	1	1	0	1	...	0	.68

After election day, we found that in fact 71% of the registered voters turned out to vote.

Table: Population

j	1	2	3	4	...	...	...	...	11.9 mil	$\bar{y}_{11.9mil}$
y_j	0	1	0	1	...	...	...	...	1	.71

Sampling Distribution

Table: Sampling Distribution of $\overline{Y}_{625}$

i	1		2		...	625		
s	J_1	Y_1	J_2	Y_2	...	J_{625}	Y_{625}	$\overline{Y}_{625}$
1	9562350	1	8763351	1	...	1294801	0	.68
2	5331704	0	4533839	1	...	3342359	1	.70
3	5129936	0	10981600	0	...	4096184	1	.75
4	803605	0	7036389	1	...	803605	0	.73
5	148567	0	3833847	1	...	4769869	1	.69
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1 mil	4163458	0	8384613	1	...	377981	1	.74
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
f.	Be(.71)		Be(.71)			Be(.71)		$\frac{Bin(625,.71)}{625}$

The Sampling Distribution in R

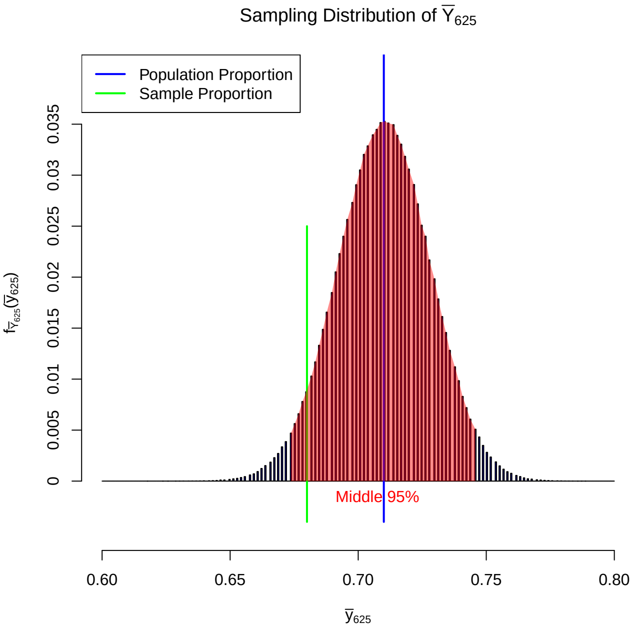
```
# Resample the number of voters 1,000,000
# times and store these 1,000,000
# numbers in a vector.

sumY_vec <- rbinom(1000000, size=625, prob=.71)
```

The Sampling Distribution in R

```
# Resample the number of voters 1,000,000
# times and store these 1,000,000
# numbers in a vector.

sumY_vec <- rbinom(1000000, size=625, prob=.71)

# Divide all of these numbers
# by the sample size.

Ybar_vec <- sumY_vec/625
```

The Sampling Distribution in R

```
# Resample the number of voters 1,000,000
# times and store these 1,000,000
# numbers in a vector.

sumY_vec <- rbinom(1000000, size=625, prob=.71)

# Divide all of these numbers
# by the sample size.

Ybar_vec <- sumY_vec/625

# Plot a histogram

hist(Ybar_vec)
```


Sampling Distribution of $\overline{Y}_{625}$
Population Proportion
Sample Proportion
$f_{\overline{Y}_{625}}(\overline{y}_{625})$
0
0.005
0.01
0.015
0.02
0.025
0.03
0.035
Middle 95%
0.60
0.65
0.70
0.75
0.80
$\overline{y}_{625}$

Bootstrapping

At the time of our sample, we don't observe the population or population proportion (.71), so we cannot construct the sampling distribution.

Bootstrapping

At the time of our sample, we don't observe the population or population proportion (.71), so we cannot construct the sampling distribution.

However, we can take repeated random samples with replacement of size 625 from the sample of size 625.

Table: Sample

i	1	2	3	4	$\dots$	625	$\bar{y}_{625}$
y_i	1	1	0	1	$\dots$	0	.68

Bootstrapping

At the time of our sample, we don't observe the population or population proportion (.71), so we cannot construct the sampling distribution.

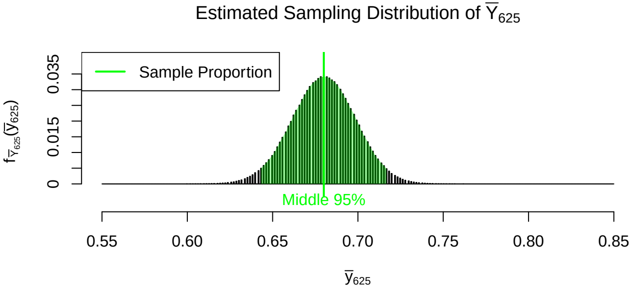
However, we can take repeated random samples with replacement of size 625 from the sample of size 625.

Table: Sample

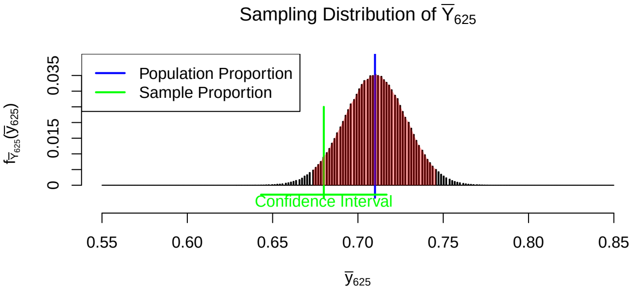
i	1	2	3	4	...	625	$\bar{y}_{625}$
y_i	1	1	0	1	...	0	.68

The is equivalent to replacing .71 with .68 in the R code.

```
sumY_vec <- rbinom(1000000, size=625, prob=.68)
```



Estimated Sampling Distribution of $\overline{Y}_{625}$
Sample Proportion
Middle 95%
0.55
0.60
0.65
0.70
0.75
0.80
0.85
$\overline{y}_{625}$
$f_{\overline{Y}_{625}}(\overline{y}_{625})$
0
0.015
0.035



Sampling Distribution of $\overline{Y}_{625}$
Population Proportion
Sample Proportion
Confidence Interval
0.55
0.60
0.65
0.70
0.75
0.80
0.85
$\overline{y}_{625}$
$f_{\overline{Y}_{625}}(\overline{y}_{625})$
0
0.015
0.035

Example 2: Linear Regression

This works with regression too!

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.
It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.
It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.
It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.
But instead we'll use Bootstrap:

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.
It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.
But instead we'll use Bootstrap:

1) Sample from data set, with replacement n times, $\tilde{\boldsymbol{X}}$

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.
It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.
But instead we'll use Bootstrap:

1) Sample from data set, with replacement n times, $\tilde{\boldsymbol{X}}$
2) Calculate $f(\tilde{\boldsymbol{X}})$ (in this case a regression)

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.
It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.
But instead we'll use Bootstrap:

1) Sample from data set, with replacement n times, $\tilde{\boldsymbol{X}}$
2) Calculate $f(\tilde{\boldsymbol{X}})$ (in this case a regression)
3) Repeat M times, form distribution of statistics

Example 2: Linear Regression

This works with regression too!

We skimmed over the sampling distribution of the variance parameter earlier.
It turns out that $\widehat{\sigma}^2 \sim \chi^2_{n-(K+1)}$.
But instead we'll use Bootstrap:

1) Sample from data set, with replacement n times, $\tilde{\boldsymbol{X}}$
2) Calculate $f(\tilde{\boldsymbol{X}})$ (in this case a regression)
3) Repeat M times, form distribution of statistics
4) Calculate confidence interval by identifying $\alpha/2$ and $1-\alpha/2$ value of statistic. (percentile method)

Confidence Intervals, via Bootstrap

Suppose we draw 20 realizations of

$$X_i \sim \text{Normal}(1, 10)$$

Confidence Intervals, via Bootstrap

Suppose we draw 20 realizations of

$$X_i \sim \text{Normal}(1, 10)$$

Bootstrapped 95% confidence interval for σ^2:

Confidence Intervals, via Bootstrap

Suppose we draw 20 realizations of

$$X_i \sim \text{Normal}(1, 10)$$

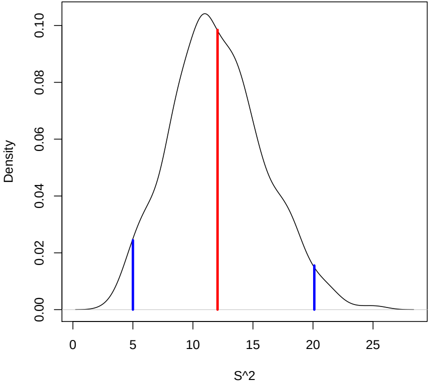
Bootstrapped 95% confidence interval for σ^2:
$[5.00, 20.11]$ (with mean 12.05)

Confidence Intervals, via Bootstrap

Suppose we draw 20 realizations of

$$X_i \sim \text{Normal}(1, 10)$$

Bootstrapped 95% confidence interval for σ^2:
$[5.00, 20.11]$ (with mean 12.05)



The Bootstrap More Formally

The Bootstrap More Formally

- What we are discussing is the nonparametric bootstrap
- $y_1, \ldots, y_n$ are the outcomes of independent and identically distributed random variables $Y_1, \ldots, Y_n$ whose PDF and CDF are denoted by f and F.

The Bootstrap More Formally

- What we are discussing is the nonparametric bootstrap
- $y_1, \dots, y_n$ are the outcomes of independent and identically distributed random variables $Y_1, \dots, Y_n$ whose PDF and CDF are denoted by f and F.
- The sample is used to make inferences about an estimand, denoted by θ using a statistic T whose value in the sample is t.

The Bootstrap More Formally

- What we are discussing is the nonparametric bootstrap
- $y_1, \ldots, y_n$ are the outcomes of independent and identically distributed random variables $Y_1, \ldots, Y_n$ whose PDF and CDF are denoted by f and F.
- The sample is used to make inferences about an estimand, denoted by θ using a statistic T whose value in the sample is t.
- If we observed F, statistical inference would be very easy, but instead we observe $\hat{F}$, which is the empirical distribution that put equal probabilities n^{-1} at each sample value y_i.

The Bootstrap More Formally

- What we are discussing is the nonparametric bootstrap
- $y_1, \ldots, y_n$ are the outcomes of independent and identically distributed random variables $Y_1, \ldots, Y_n$ whose PDF and CDF are denoted by f and F.
- The sample is used to make inferences about an estimand, denoted by θ using a statistic T whose value in the sample is t.
- If we observed F, statistical inference would be very easy, but instead we observe $\hat{F}$, which is the empirical distribution that put equal probabilities n^{-1} at each sample value y_i.
  - Estimates are constructed by the plug-in principle, which says that the parameter $\theta = t(F)$ is estimated by $\hat{\theta} = t(\hat{F})$. (i.e. we plug in the ECDF for the CDF)

The Bootstrap More Formally

- What we are discussing is the nonparametric bootstrap
- $y_1, \ldots, y_n$ are the outcomes of independent and identically distributed random variables $Y_1, \ldots, Y_n$ whose PDF and CDF are denoted by f and F.
- The sample is used to make inferences about an estimand, denoted by θ using a statistic T whose value in the sample is t.
- If we observed F, statistical inference would be very easy, but instead we observe $\hat{F}$, which is the empirical distribution that put equal probabilities n^{-1} at each sample value y_i.
  - Estimates are constructed by the plug-in principle, which says that the parameter $\theta = t(F)$ is estimated by $\hat{\theta} = t(\hat{F})$. (i.e. we plug in the ECDF for the CDF)
  - Why does this work? Sampling distribution entirely determined by the CDF and n, WLLN says the ECDF will look more and more like the CDF as n gets large.

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

- Bootstrap fails when the sampling distribution of the estimator is non-smooth. (e.g. max and min).

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

- Bootstrap fails when the sampling distribution of the estimator is non-smooth. (e.g. max and min).
- Dependent data: nonparametric bootstrap assumes data so independent so will not work with time series data or other dependent structures.

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

- Bootstrap fails when the sampling distribution of the estimator is non-smooth. (e.g. max and min).
- Dependent data: nonparametric bootstrap assumes data so independent so will not work with time series data or other dependent structures.
  - For clustered data, standard bootstrap will not work, but the block bootstrap will work. In the block bootstrap, clusters are resampled (not necessarily units) with replacement.

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

- Bootstrap fails when the sampling distribution of the estimator is non-smooth. (e.g. max and min).
- Dependent data: nonparametric bootstrap assumes data so independent so will not work with time series data or other dependent structures.
  - For clustered data, standard bootstrap will not work, but the block bootstrap will work. In the block bootstrap, clusters are resampled (not necessarily units) with replacement.
  - More on this later.

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

- Bootstrap fails when the sampling distribution of the estimator is non-smooth. (e.g. max and min).
- Dependent data: nonparametric bootstrap assumes data so independent so will not work with time series data or other dependent structures.
  - For clustered data, standard bootstrap will not work, but the block bootstrap will work. In the block bootstrap, clusters are resampled (not necessarily units) with replacement.
  - More on this later.
- Many other variants that may be right for certain situations: studentized intervals, jackknife, parametric bootstrap, bag of little bootstraps, bootstrapping for complex survey designs, etc.

When Does the Bootstrap Fail?

Bootstrap works in a wide variety of circumstances, but it does require some regularity conditions and it can fail with certain types of data and estimators:

- Bootstrap fails when the sampling distribution of the estimator is non-smooth. (e.g. max and min).
- Dependent data: nonparametric bootstrap assumes data so independent so will not work with time series data or other dependent structures.
  - For clustered data, standard bootstrap will not work, but the block bootstrap will work. In the block bootstrap, clusters are resampled (not necessarily units) with replacement.
  - More on this later.
- Many other variants that may be right for certain situations: studentized intervals, jackknife, parametric bootstrap, bag of little bootstraps, bootstrapping for complex survey designs, etc.

Fox Chapter 21 has a nice section on the bootstrap, Aronow and Miller (2016) covers the theory well.

This Section

This Section

- The next few slides have some technical details of vector/matrix calculus

This Section

- The next few slides have some technical details of vector/matrix calculus
- You won't be tested on this material but its necessary for the proofs in the appendix

This Section

- The next few slides have some technical details of vector/matrix calculus
- You won't be tested on this material but its necessary for the proofs in the appendix
- It will also come back in Soc504 where you will need to know this stuff, so its worth thinking about now (but I will reintroduce it next semester).

This Section

- The next few slides have some technical details of vector/matrix calculus
- You won't be tested on this material but its necessary for the proofs in the appendix
- It will also come back in Soc504 where you will need to know this stuff, so its worth thinking about now (but I will reintroduce it next semester).
- We will just preview this stuff now, but I'm happy to answer questions for those who want to engage it more.

Gradient

Gradient

Let $v = v(\mathbf{u})$ be a scalar-valued function $\mathbb{R}_n \to \mathbb{R}_1$ where $\mathbf{u}$ is a $(n \times 1)$ column vector.

Gradient

Let $v = v(\mathbf{u})$ be a scalar-valued function $\mathbb{R}_n \rightarrow \mathbb{R}_1$ where $\mathbf{u}$ is a $(n \times 1)$ column vector. For example: $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$

Gradient

Let $v = v(\mathbf{u})$ be a scalar-valued function $\mathbb{R}_n \to \mathbb{R}_1$ where $\mathbf{u}$ is a $(n \times 1)$ column vector. For example: $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$

Definition (Gradient)

We can define the column vector of partial derivatives

$$\frac{\partial v(\mathbf{u})}{\partial \mathbf{u}} = \begin{bmatrix} \partial v/\partial u_1 \\ \partial v/\partial u_2 \\ \vdots \\ \partial v/\partial u_n \end{bmatrix}$$

This vector of partial derivatives is called the gradient.

Vector Derivative Rule I (linear functions)

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ *of an* $(n \times 1)$ *vector* $\mathbf{u}$, *the derivative of* $v(\mathbf{u})$ *w.r.t.* $\mathbf{u}$ *is given by*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ *is a matrix and therefore* v *is a vector-valued function.*

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ *of an* $(n \times 1)$ *vector* $\mathbf{u}$, *the derivative of* $v(\mathbf{u})$ *w.r.t.* $\mathbf{u}$ *is given by*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ *is a matrix and therefore* v *is a vector-valued function.*

For example, let $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ *of an* $(n \times 1)$ *vector* $\mathbf{u}$*, the derivative of* $v(\mathbf{u})$ *w.r.t.* $\mathbf{u}$ *is given by*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ *is a matrix and therefore* v *is a vector-valued function.*

For example, let $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$, then

$$v = \mathbf{c}'\mathbf{u} = 0 \cdot u_1 + 1 \cdot u_2 + 3 \cdot u_3$$

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ of an $(n \times 1)$ vector $\mathbf{u}$, the derivative of $v(\mathbf{u})$ w.r.t. $\mathbf{u}$ is given by

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ is a matrix and therefore v is a vector-valued function.

For example, let $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$, then

$$v = \mathbf{c}'\mathbf{u} = 0 \cdot u_1 + 1 \cdot u_2 + 3 \cdot u_3$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v/\partial u_1 \\ \partial v/\partial u_2 \\ \partial v/\partial u_3 \end{bmatrix} =$$

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ *of an* $(n \times 1)$ *vector* $\mathbf{u}$, *the derivative of* $v(\mathbf{u})$ *w.r.t.* $\mathbf{u}$ *is given by*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ *is a matrix and therefore* v *is a vector-valued function.*

For example, let $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$, then

$$v = \mathbf{c}'\mathbf{u} = 0 \cdot u_1 + 1 \cdot u_2 + 3 \cdot u_3$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v/\partial u_1 \\ \partial v/\partial u_2 \\ \partial v/\partial u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} = \mathbf{c}$$

Vector Derivative Rule I (linear functions)

Theorem (differentiation of linear functions)

Given a linear function $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ of an $(n \times 1)$ vector $\mathbf{u}$, the derivative of $v(\mathbf{u})$ w.r.t. $\mathbf{u}$ is given by

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

This also works when $\mathbf{c}$ is a matrix and therefore v is a vector-valued function.

For example, let $v(\mathbf{u}) = \mathbf{c}'\mathbf{u}$ where $\mathbf{c} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}$, then

$$v = \mathbf{c}'\mathbf{u} = 0 \cdot u_1 + 1 \cdot u_2 + 3 \cdot u_3$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v/\partial u_1 \\ \partial v/\partial u_2 \\ \partial v/\partial u_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 3 \end{bmatrix} = \mathbf{c}$$

Hence,

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{c}$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$, *we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$*, we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$*, we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$, *we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

$$\begin{aligned} v &= [3 \cdot u_1 + u_2,\ u_1 + 5 \cdot u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= 3u_1^2 + 2u_1 u_2 + 5u_2^2 \end{aligned}$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \end{bmatrix} = \begin{bmatrix} 6u_1 + 2u_2 \\ 2u_1 + 10u_2 \end{bmatrix} = 2 \cdot \begin{bmatrix} 3u_1 + 1u_2 \\ 1u_1 + 5u_2 \end{bmatrix} = 2\mathbf{A}\mathbf{u}$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$, *we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

$$\begin{aligned} v &= [3 \cdot u_1 + u_2,\ u_1 + 5 \cdot u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= 3u_1^2 + 2u_1 u_2 + 5u_2^2 \end{aligned}$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$, *we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

$$\begin{aligned} v &= [3 \cdot u_1 + u_2,\ u_1 + 5 \cdot u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= 3u_1^2 + 2u_1u_2 + 5u_2^2 \end{aligned}$$

and

$$\frac{\partial v}{\partial \mathbf{u}} =$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$, *we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

$$\begin{aligned} v &= [3 \cdot u_1 + u_2,\ u_1 + 5 \cdot u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= 3u_1^2 + 2u_1 u_2 + 5u_2^2 \end{aligned}$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \end{bmatrix} =$$

Vector Derivative Rule II (quadratic form)

Theorem (quadratic form)

Given a $(n \times n)$ *symmetric matrix* $\mathbf{A}$ *and a scalar-valued function* $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ *of* $(n \times 1)$ *vector* $\mathbf{u}$*, we have*

$$\frac{\partial v}{\partial \mathbf{u}} = \mathbf{A}'\mathbf{u} + \mathbf{A}\mathbf{u} = 2\mathbf{A}\mathbf{u}$$

For example, let $\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 5 \end{bmatrix}$ and $\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}$.Then $v(\mathbf{u}) = \mathbf{u}'\mathbf{A}\mathbf{u}$ is equal to

$$\begin{aligned} v &= [3 \cdot u_1 + u_2,\ u_1 + 5 \cdot u_2] \begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \\ &= 3u_1^2 + 2u_1u_2 + 5u_2^2 \end{aligned}$$

and

$$\frac{\partial v}{\partial \mathbf{u}} = \begin{bmatrix} \partial v / \partial u_1 \\ \partial v / \partial u_2 \end{bmatrix} = \begin{bmatrix} 6u_1 + 2u_2 \\ 2u_1 + 10u_2 \end{bmatrix} = 2 \cdot \begin{bmatrix} 3u_1 + 1u_2 \\ 1u_1 + 5u_2 \end{bmatrix} = 2\mathbf{A}\mathbf{u}$$

Hessian

Hessian

Suppose v is a scalar-valued function $v = f(\mathbf{u})$ of a $(k+1) \times 1$ column vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{k+1} \end{bmatrix}'$

Hessian

Suppose v is a scalar-valued function $v = f(\mathbf{u})$ of a $(k+1) \times 1$ column vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{k+1} \end{bmatrix}'$

Definition (Hessian)

The $(k+1) \times (k+1)$ matrix of second-order partial derivatives of $v = f(\mathbf{u})$ is called the Hessian matrix and denoted

$$\frac{\partial v^2}{\partial \mathbf{u} \partial \mathbf{u}'} = \begin{bmatrix} \frac{\partial v^2}{\partial u_1 \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_1 \partial u_{k+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial v^2}{\partial u_{k+1} \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_{k+1} \partial u_{k+1}} \end{bmatrix}$$

Hessian

Suppose v is a scalar-valued function $v = f(\mathbf{u})$ of a $(k+1) \times 1$ column vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{k+1} \end{bmatrix}'$

Definition (Hessian)

The $(k+1) \times (k+1)$ matrix of second-order partial derivatives of $v = f(\mathbf{u})$ is called the Hessian matrix and denoted

$$\frac{\partial v^2}{\partial \mathbf{u} \partial \mathbf{u}'} = \begin{bmatrix} \frac{\partial v^2}{\partial u_1 \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_1 \partial u_{k+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial v^2}{\partial u_{k+1} \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_{k+1} \partial u_{k+1}} \end{bmatrix}$$

Note: The Hessian is symmetric.

Hessian

Suppose v is a scalar-valued function $v = f(\mathbf{u})$ of a $(k+1) \times 1$ column vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \cdots & u_{k+1} \end{bmatrix}'$

Definition (Hessian)

The $(k+1) \times (k+1)$ matrix of second-order partial derivatives of $v = f(\mathbf{u})$ is called the Hessian matrix and denoted

$$\frac{\partial v^2}{\partial \mathbf{u} \partial \mathbf{u}'} = \begin{bmatrix} \frac{\partial v^2}{\partial u_1 \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_1 \partial u_{k+1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial v^2}{\partial u_{k+1} \partial u_1} & \cdots & \frac{\partial v^2}{\partial u_{k+1} \partial u_{k+1}} \end{bmatrix}$$

Note: The Hessian is symmetric.

The above rules are used to derive the optimal estimators in the appendix slides.

Conclusion

Conclusion

- Multiple regression is much like the regression formulations we have already seen

Conclusion

- Multiple regression is much like the regression formulations we have already seen
- We showed how to estimate the coefficients and get the variance covariance matrix

Conclusion

- Multiple regression is much like the regression formulations we have already seen
- We showed how to estimate the coefficients and get the variance covariance matrix
- We discussed the bootstrap as an alternative strategy for estimating the sampling distribution

Conclusion

- Multiple regression is much like the regression formulations we have already seen
- We showed how to estimate the coefficients and get the variance covariance matrix
- We discussed the bootstrap as an alternative strategy for estimating the sampling distribution
- Appendix contains numerous additional topics worth knowing:
  - Systems of Equations
  - Details on the variance/covariance interpretation of estimator
  - Derivation for the estimator
  - Proof of consistency

Fun With Weights

Aronow, Peter M., and Cyrus Samii. "Does Regression Produce Representative Estimates of Causal Effects?." *American Journal of Political Science* (2015).[2]

[2] I'm grateful to Peter Aronow for sharing his slides, several of which are used here.

Fun With Weights

Aronow, Peter M., and Cyrus Samii. "Does Regression Produce Representative Estimates of Causal Effects?." *American Journal of Political Science* (2015).[2]

- Imagine we care about the possibly heterogeneous causal effect of a treatment D and we control for some covariates X?

[2] I'm grateful to Peter Aronow for sharing his slides, several of which are used here.

Fun With Weights

Aronow, Peter M., and Cyrus Samii. "Does Regression Produce Representative Estimates of Causal Effects?." *American Journal of Political Science* (2015).[2]

- Imagine we care about the possibly heterogeneous causal effect of a treatment D and we control for some covariates X?
- We can express the regression as a weighting over individual observation treatment effects where the weight depends only on X.

[2] I'm grateful to Peter Aronow for sharing his slides, several of which are used here.

Fun With Weights

Aronow, Peter M., and Cyrus Samii. "Does Regression Produce Representative Estimates of Causal Effects?." *American Journal of Political Science* (2015).[2]

- Imagine we care about the possibly heterogeneous causal effect of a treatment D and we control for some covariates X?
- We can express the regression as a weighting over individual observation treatment effects where the weight depends only on X.
- Useful technology for understanding what our models are identifying off of by showing us our effective sample.

[2] I'm grateful to Peter Aronow for sharing his slides, several of which are used here.

How this works

How this works

We start by asking what the estimate of the average causal effect of interest converges to in a large sample:

How this works

We start by asking what the estimate of the average causal effect of interest converges to in a large sample:

$$\hat{\beta} \xrightarrow{p} \frac{E[w_i \tau_i]}{E[w_i]} \text{ where } w_i = (D_i - E[D_i|X])^2 \, ,$$

so that $\hat{\beta}$ converges to a reweighted causal effect. As $E[w_i|X_i] = \text{Var}[D_i|X_i]$, we obtain an average causal effect reweighted by conditional variance of the treatment.

Estimation

A simple, consistent plug-in estimator of w_i is available: $\hat{w}_i = \tilde{D}_i^2$ where $\tilde{D}_i$ is the residualized treatment. (the proof is connected to the partialing out strategy we showed last week)

Estimation

A simple, consistent plug-in estimator of w_i is available: $\hat{w}_i = \tilde{D}_i^2$ where $\tilde{D}_i$ is the residualized treatment. (the proof is connected to the partialing out strategy we showed last week)

Easily implemented in R:

```
wts <- (d - predict(lm(d~x)))^2
```

Implications

- Unpacking the black box of regression gives us *substantive insight*

Implications

- Unpacking the black box of regression gives us *substantive insight*
- When some observations have no weight, this means that the covariates *completely* explain their treatment condition.

Implications

- Unpacking the black box of regression gives us *substantive insight*
- When some observations have no weight, this means that the covariates *completely* explain their treatment condition.
- This is a *feature*, not a *bug*, of regression: it can automatically handle issues of *common support*.

Implications

- Unpacking the black box of regression gives us *substantive insight*
- When some observations have no weight, this means that the covariates *completely* explain their treatment condition.
- This is a *feature*, not a *bug*, of regression: it can automatically handle issues of *common support*.

Application

Jensen (2003), "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment."

Application

Jensen (2003), "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment."

Jensen presents a large-N TSCS-analysis of the causal effects of governance (as measured by the Polity III score) on Foreign Direct Investment (FDI).

Application

Jensen (2003), "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment."

Jensen presents a large-N TSCS-analysis of the causal effects of governance (as measured by the Polity III score) on Foreign Direct Investment (FDI).

The nominal sample: 114 countries from 1970 to 1997.

Application

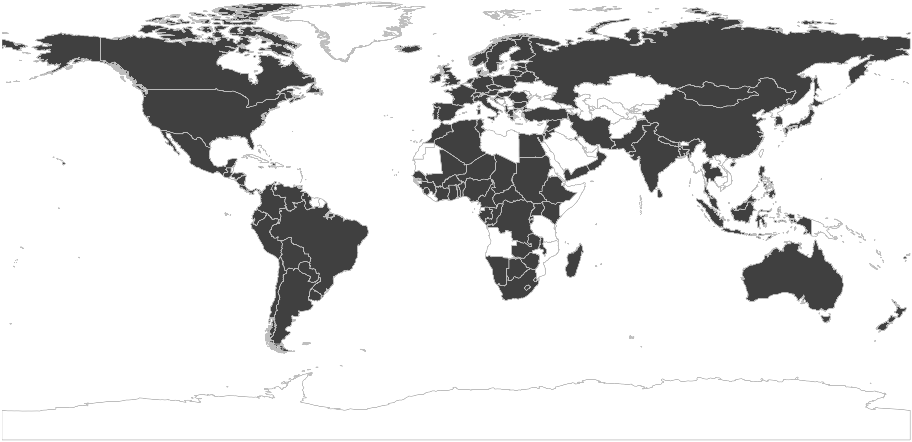
Jensen (2003), "Democratic Governance and Multinational Corporations: Political Regimes and Inflows of Foreign Direct Investment."

Jensen presents a large-N TSCS-analysis of the causal effects of governance (as measured by the Polity III score) on Foreign Direct Investment (FDI).

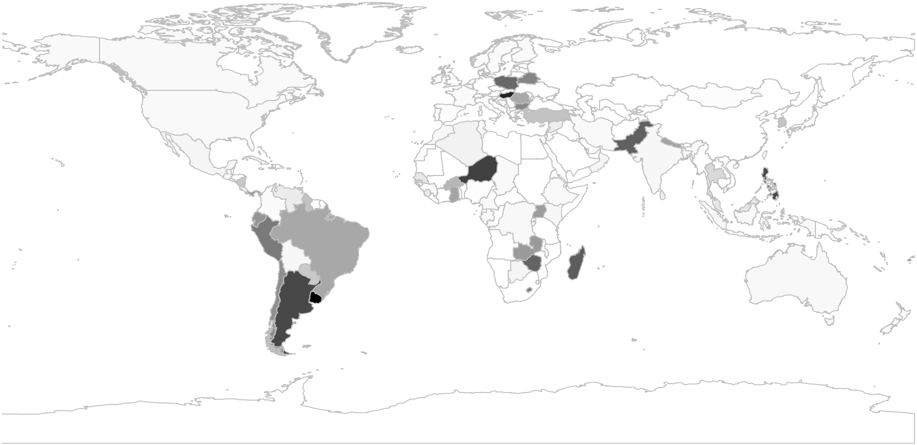
The nominal sample: 114 countries from 1970 to 1997.

Jensen estimates that a 1 unit increase in polity score corresponds to a 0.020 increase in net FDI inflows as a percentage of GDP ($p < 0.001$).

Nominal and Effective Samples



Nominal and Effective Samples



Nominal and Effective Samples

Over 50% of the weight goes to just 12 (out of 114) countries.

Broader Implications

When causal effects are heterogeneous, we can draw a distinction between "internally valid" and "externally valid" estimates of an Average Causal Effect (ACE). (See, e.g., Cook and Campbell 1979)

Broader Implications

When causal effects are heterogeneous, we can draw a distinction between "internally valid" and "externally valid" estimates of an Average Causal Effect (ACE). (See, e.g., Cook and Campbell 1979)

- "Internally valid": reliable estimates of ACEs, but perhaps not for the population you care about
  - randomized (lab, field, survey) experiments, instrumental variables, regression discontinuity designs, other natural experiments

Broader Implications

When causal effects are heterogeneous, we can draw a distinction between "internally valid" and "externally valid" estimates of an Average Causal Effect (ACE). (See, e.g., Cook and Campbell 1979)

- "Internally valid": reliable estimates of ACEs, but perhaps not for the population you care about
  - randomized (lab, field, survey) experiments, instrumental variables, regression discontinuity designs, other natural experiments
- "Externally valid": perhaps unreliable estimates of ACEs, but for the population of interest
  - large-N analyses, representative surveys

Broader Implications

Aronow and Samii argue that analyses which use regression, *even with a representative sample*, have no greater claim to external validity than do [natural] experiments.

Broader Implications

Aronow and Samii argue that analyses which use regression, *even with a representative sample*, have no greater claim to external validity than do [natural] experiments.

- When a treatment is "as-if" randomly assigned conditional on covariates, regression distorts the sample by implicitly applying weights.

Broader Implications

Aronow and Samii argue that analyses which use regression, *even with a representative sample*, have no greater claim to external validity than do [natural] experiments.

- When a treatment is "as-if" randomly assigned conditional on covariates, regression distorts the sample by implicitly applying weights.
- The *effective sample* (upon which causal effects are estimated) may have radically different properties than the nominal sample.

Broader Implications

Aronow and Samii argue that analyses which use regression, *even with a representative sample*, have no greater claim to external validity than do [natural] experiments.

- When a treatment is "as-if" randomly assigned conditional on covariates, regression distorts the sample by implicitly applying weights.
- The *effective sample* (upon which causal effects are estimated) may have radically different properties than the nominal sample.
- When there is an underlying natural experiment in the data, a properly specified regression model may reproduce the internally valid estimate associated with the natural experiment.

Solving Systems of Equations Using Matrices

Matrices are very useful to solve linear systems of equations, such as the first order conditions for our least squares estimates.

Here is an example with three equations and three unknowns:

$$\begin{aligned} x + 2y + z &= 3 \\ 3x - y - 3z &= -1 \\ 2x + 3y + z &= 4 \end{aligned}$$

How would one go about solving this?
There are various techniques, including substitution, and multiplying equations by constants and adding them to get single variables to cancel.

Solving Systems of Equations Using Matrices

An easier way is to use matrix algebra. Note that the system of equations

$$\begin{aligned} x + 2y + z &= 3 \\ 3x - y - 3z &= -1 \\ 2x + 3y + z &= 4 \end{aligned}$$

can be written as follows:

$$\begin{bmatrix} 1 & 2 & 1 \\ 3 & -1 & -3 \\ 2 & 3 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} \iff \mathbf{Au} = \mathbf{b}$$

How do we solve this for $\mathbf{u} = \begin{bmatrix} x \\ y \\ z \end{bmatrix}$? Let's look again at the scalar case first.

Solving Equations with Inverses (scalar case)

Let's go back to the scalar world of 8th grade algebra. How would you solve the following for u?

$$au = b$$

We multiply both sides of by the reciprocal $1/a$ (the inverse of a) and get:

$$\frac{1}{a} a\, u = \frac{1}{a} b$$

$$u = \frac{b}{a}$$

(Note that this technique only works if $a \neq 0$. If $a = 0$, then there are either an infinite number of solutions for u (when $b = 0$), or no solutions for u (when $b \neq 0$).)

So to solve our multiple equation problem in the matrix case we need a matrix equivalent of the inverse. This equivalent is the inverse matrix. The inverse of $\mathbf{A}$ is written as $\mathbf{A}^{-1}$.

Inverse of a Matrix

The inverse $\mathbf{A}^{-1}$ of $\mathbf{A}$ has the property that $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

- The inverse $\mathbf{A}^{-1}$ exists only if $\mathbf{A}$ is invertible or nonsingular (more on this soon)
- The inverse is unique if it exists and then the linear system has a unique solution.
- There are various methods for finding/computing the inverse of a matrix

The inverse matrix allows us to solve linear systems of equations.

$$\begin{aligned} \mathbf{A}\mathbf{u} &= \mathbf{b} \\ \mathbf{A}^{-1}\mathbf{A}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{I}\mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \\ \mathbf{u} &= \mathbf{A}^{-1}\mathbf{b} \end{aligned}$$

Given $\mathbf{A}$ we find that $\mathbf{A^{-1}}$ is:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 3 & -1 & -3 \\ 2 & 3 & 1 \end{bmatrix} ; \mathbf{A^{-1}} = \begin{bmatrix} 8 & 1 & -5 \\ -9 & -1 & 6 \\ 11 & 1 & -7 \end{bmatrix}$$

We can now solve our system of equations:

Given $\mathbf{A}$ we find that $\mathbf{A}^{-1}$ is:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 3 & -1 & -3 \\ 2 & 3 & 1 \end{bmatrix} ; \mathbf{A}^{-1} = \begin{bmatrix} 8 & 1 & -5 \\ -9 & -1 & 6 \\ 11 & 1 & -7 \end{bmatrix}$$

We can now solve our system of equations:

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{b} = \begin{bmatrix} 8 & 1 & -5 \\ -9 & -1 & 6 \\ 11 & 1 & -7 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \\ 4 \end{bmatrix}$$

Given $\mathbf{A}$ we find that $\mathbf{A}^{-1}$ is:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 1 \\ 3 & -1 & -3 \\ 2 & 3 & 1 \end{bmatrix} ; \mathbf{A}^{-1} = \begin{bmatrix} 8 & 1 & -5 \\ -9 & -1 & 6 \\ 11 & 1 & -7 \end{bmatrix}$$

We can now solve our system of equations:

$$\mathbf{u} = \mathbf{A}^{-1}\mathbf{b} = \begin{bmatrix} 8 & 1 & -5 \\ -9 & -1 & 6 \\ 11 & 1 & -7 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \\ 4 \end{bmatrix} = \begin{bmatrix} 3 \\ -2 \\ 4 \end{bmatrix}$$

So the solution vector is $x = 3$, $y = -2$, and $z = 4$. Verifying:

$$\begin{aligned} x + 2y + z &= 3 + 2 \cdot -2 + 4 &= 3 \\ 3x - y - 3z &= 3 \cdot 3 - -2 - 3 \cdot 4 &= -1 \\ 2x + 3y + z &= 2 \cdot 3 + 3 \cdot -2 + 4 &= 4 \end{aligned}$$

Computationally, this method is very convenient. We "just" compute the inverse, and perform a single matrix multiplication.

Singularity of a Matrix

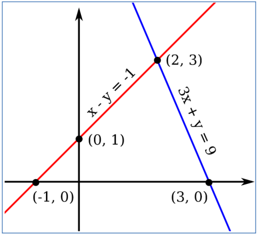
If the inverse of **A** exists, then the linear system has a unique (non-trivial) solution. If it exists, we say that **A** is nonsingular or invertible (these statements are equivalent).

A must be square to be invertible, but not all square matrices are invertible. More precisely, a square matrix **A** is invertible iff its column vectors (or equivalently its row vectors) are linearly independent.

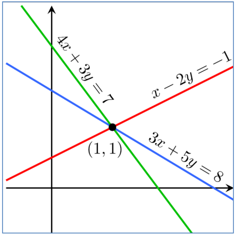
The column rank of a matrix **A** is the largest number of linearly independent columns of **A**. If the rank of **A** equals the number of columns of **A**, then we say that **A** has full column rank. This implies that all its column vectors are linearly independent.

If a column of **A** is a linear combination of the other columns, there are either no solutions to the system of equations or infinitely many solutions to the system of equations. The system is said to be underdetermined.

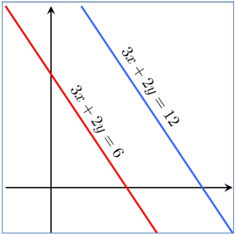
Geometric Example in 2D



Unique Solution



Redundant Equation



No Solution

$$A = \begin{bmatrix} 1 & -1 \\ 3 & 1 \end{bmatrix} \quad A = \begin{bmatrix} 4 & 3 \\ 1 & -2 \\ 3 & 5 \end{bmatrix} \quad A = \begin{bmatrix} 3 & 2 \\ 3 & 2 \end{bmatrix}$$

Why do we care about invertibility?

We have seen that OLS regression is defined by a system of linear equations

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1\hat{\beta}_0 + x_{11}\hat{\beta}_1 + x_{12}\hat{\beta}_2 + \cdots + x_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + x_{21}\hat{\beta}_1 + x_{22}\hat{\beta}_2 + \cdots + x_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + x_{n1}\hat{\beta}_1 + x_{n2}\hat{\beta}_2 + \cdots + x_{nk}\hat{\beta}_k \end{bmatrix}$$

with our data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

Why do we care about invertibility?

We have seen that OLS regression is defined by a system of linear equations

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1\hat{\beta}_0 + x_{11}\hat{\beta}_1 + x_{12}\hat{\beta}_2 + \cdots + x_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + x_{21}\hat{\beta}_1 + x_{22}\hat{\beta}_2 + \cdots + x_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + x_{n1}\hat{\beta}_1 + x_{n2}\hat{\beta}_2 + \cdots + x_{nk}\hat{\beta}_k \end{bmatrix}$$

with our data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

We have also learned that $\hat{\boldsymbol{\beta}}$ is obtained by solving normal equations, a linear system of equations.

Why do we care about invertibility?

We have seen that OLS regression is defined by a system of linear equations

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \mathbf{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} 1\hat{\beta}_0 + x_{11}\hat{\beta}_1 + x_{12}\hat{\beta}_2 + \cdots + x_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + x_{21}\hat{\beta}_1 + x_{22}\hat{\beta}_2 + \cdots + x_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + x_{n1}\hat{\beta}_1 + x_{n2}\hat{\beta}_2 + \cdots + x_{nk}\hat{\beta}_k \end{bmatrix}$$

with our data matrix

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{bmatrix}$$

We have also learned that $\hat{\boldsymbol{\beta}}$ is obtained by solving normal equations, a linear system of equations.

It turns out that to solve for $\hat{\boldsymbol{\beta}}$, we need to invert $\mathbf{X}'\mathbf{X}$, a $(k+1) \times (k+1)$ matrix.

Some Non-invertible Explanatory Data Matrices

$\mathbf{X}'\mathbf{X}$ is invertible iff $\mathbf{X}$ is full column rank (see Wooldridge D.4), so the collection of predictors need to be linearly independent (no perfect collinearity).

Some example of $\mathbf{X}$ that are not full column rank:

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 3 & -3 \\ 1 & 4 & -4 \\ 1 & 5 & -5 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 54 & 54,000 \\ 1 & 37 & 37,000 \\ 1 & 89 & 89,000 \\ 1 & 72 & 72,000 \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

Covariance/variance interpretation of matrix OLS

$$\mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \begin{bmatrix} y_i \\ y_i x_{i1} \\ y_i x_{i2} \\ \vdots \\ y_i x_{iK} \end{bmatrix} \approx \begin{bmatrix} n\overline{y} \\ \widehat{\text{cov}}(y_i, x_{i1}) \\ \widehat{\text{cov}}(y_i, x_{i2}) \\ \vdots \\ \widehat{\text{cov}}(y_i, x_{iK}) \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \begin{bmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{iK} \\ x_{i1} & x_{i1}^2 & x_{i2}x_{i1} & \cdots & x_{i1}x_{iK} \\ x_{i2} & x_{i1}x_{i2} & x_{i2}^2 & \cdots & x_{i2}x_{iK} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{iK} & x_{i1}x_{iK} & x_{i2}x_{iK} & \cdots & x_{iK}x_{iK} \end{bmatrix} \approx \begin{bmatrix} n & n\overline{x}_1 & n\overline{x}_2 \\ n\overline{x}_1 & \widehat{\text{var}}(x_{i1}) & \widehat{\text{cov}}(x_{i1}, x_{i2}) \\ n\overline{x}_2 & \widehat{\text{cov}}(x_{i2}, x_{i1}) & \widehat{\text{var}}(x_{i2}) \\ \vdots & \vdots & \vdots \\ n\overline{x}_K & \widehat{\text{cov}}(x_{iK}, x_{i1}) & \widehat{\text{cov}}(x_{iK}, x_{i2}) \end{bmatrix}$$

Derivatives with respect to $\tilde{\boldsymbol{\beta}}$

$$\begin{aligned} S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\ &= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \end{aligned}$$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} =$$

Derivatives with respect to $\tilde{\boldsymbol{\beta}}$

$$
\begin{aligned}
S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}
\end{aligned}
$$

$$
\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}
$$

- The first term does not contain $\tilde{\boldsymbol{\beta}}$
- The second term is an example of rule I from the derivative section
- The third term is an example of rule II from the derivative section

Derivatives with respect to $\tilde{\boldsymbol{\beta}}$

$$
\begin{aligned}
S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y}) &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}
\end{aligned}
$$

$$
\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}
$$

- The first term does not contain $\tilde{\boldsymbol{\beta}}$
- The second term is an example of rule I from the derivative section
- The third term is an example of rule II from the derivative section

And while we are at it the Hessian is:

$$
\frac{\partial^2 S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}} \partial \tilde{\boldsymbol{\beta}}'} = 2\mathbf{X}'\mathbf{X}
$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

Solving for $\hat{\beta}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\mathbf{0} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\begin{aligned} \mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ -2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} \end{aligned}$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\begin{aligned} \mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\ -2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} \\ \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \end{aligned}$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\begin{aligned}
\mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
-2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\
(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\begin{aligned}
\mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
-2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\
\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
\mathbf{I}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\begin{aligned}
\mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
-2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\
\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
\mathbf{I}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}$$

Solving for $\hat{\boldsymbol{\beta}}$

$$\frac{\partial S(\tilde{\boldsymbol{\beta}}, \mathbf{X}, \mathbf{y})}{\partial \tilde{\boldsymbol{\beta}}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}$$

Setting the vector of partial derivatives equal to zero and substituting $\hat{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$, we can solve for the OLS estimator.

$$\begin{aligned}
\mathbf{0} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \\
-2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= -2\mathbf{X}'\mathbf{y} \\
\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \mathbf{X}'\mathbf{y} \\
\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
\mathbf{I}\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}$$

Note that we implicitly assumed that $\mathbf{X}'\mathbf{X}$ is invertible.

Variance-Covariance Matrix of Random Vectors

Let's unpack the homoskedasticity assumption $V[\mathbf{u}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$.

Definition (variance-covariance matrix)

For a $(n \times 1)$ random vector $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \dots & u_n \end{bmatrix}'$, its variance-covariance matrix, denoted $V[\mathbf{u}]$ or also $\Sigma_{\mathbf{u}}$, is defined as:

$$V[\mathbf{u}] = \Sigma_{\mathbf{u}} = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 & \dots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_2^2 & \dots & \sigma_{2n}^2 \\ \vdots & & \dots & \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \dots & \sigma_n^2 \end{bmatrix}$$

where $\sigma_j^2 = V[u_j]$ and $\sigma_{ij}^2 = Cov[u_i, u_j]$.

Notice that this matrix is always symmetric.

Homoskedasticity in Matrix Notation

If $V[u_i] = \sigma^2$ for all $i = 1, ..., n$ and the units are independent then $V[\mathbf{u}] = \sigma^2 \mathbf{I}_n$.

Homoskedasticity in Matrix Notation

If $V[u_i] = \sigma^2$ for all $i = 1, ..., n$ and the units are independent then $V[\mathbf{u}] = \sigma^2 \mathbf{I}_n$.
More visually:

$$V[\mathbf{u}] = \sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Homoskedasticity in Matrix Notation

If $V[u_i] = \sigma^2$ for all $i = 1, ..., n$ and the units are independent then $V[\mathbf{u}] = \sigma^2 \mathbf{I}_n$.
More visually:

$$V[\mathbf{u}] = \sigma^2 \mathbf{I}_n = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

So homoskedasticity $V[\mathbf{u}|\mathbf{X}] = \sigma^2 \mathbf{I}_n$ implies that:

1. $V[u_i|\mathbf{X}] = \sigma^2$ for all i (the variance of the errors u_i does not depend on $\mathbf{X}$ and is constant across observations)
2. $Cov[u_i, u_j|\mathbf{X}] = 0$ for all $i \neq j$ (the errors are uncorrelated across observations). This holds under our random sampling assumption.

Estimation of the Error Variance

Given our vector of regression error terms $\mathbf{u}$, what is $E[\mathbf{u}\mathbf{u}']$?

$E[\mathbf{u}\mathbf{u}'] =$

Estimation of the Error Variance

Given our vector of regression error terms $\mathbf{u}$, what is $E[\mathbf{u}\mathbf{u}']$?

$$E[\mathbf{u}\mathbf{u}'] = \begin{bmatrix} E[u_1^2] & E[u_1 u_2] & \dots & E[u_1 u_n] \\ E[u_2 u_1] & E[u_2^2] & \dots & E[u_2 u_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[u_n u_1] & E[u_n u_2] & \dots & E[u_n^2] \end{bmatrix} =$$

Estimation of the Error Variance

Given our vector of regression error terms $\mathbf{u}$, what is $E[\mathbf{u}\mathbf{u}']$?

$$E[\mathbf{u}\mathbf{u}'] = \begin{bmatrix} E[u_1^2] & E[u_1u_2] & \dots & E[u_1u_n] \\ E[u_2u_1] & E[u_2^2] & \dots & E[u_2u_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[u_nu_1] & E[u_nu_2] & \dots & E[u_n^2] \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Recall $E[u_i] = 0$ for all i. So $V[u_i] = E[u_i^2] - (E[u_i])^2 = E[u_i^2]$ and by independence $E[u_iu_j] = E[u_i] \cdot E[u_j] = 0$

$$Var(\mathbf{u}) = E[\mathbf{u}\mathbf{u}'] = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

Variance of Linear Function of Random Vector

Definition (Variance of Linear Transformation of Random Vector)

Recall that for a linear transformation of a random variable X we have $V[aX + b] = a^2 V[X]$ with constants a and b.

There is an analogous rule for linear functions of random vectors. Let $v(\mathbf{u}) = \mathbf{Au} + \mathbf{B}$ be a linear transformation of a random vector $\mathbf{u}$ with non-random vectors or matrices $\mathbf{A}$ and $\mathbf{B}$. Then the variance of the transformation is given by:

$$V[v(\mathbf{u})] = V[\mathbf{Au} + \mathbf{B}] = \mathbf{A} V[\mathbf{u}]\mathbf{A}' = \mathbf{A}\Sigma_{\mathbf{u}}\mathbf{A}'$$

Conditional Variance of $\hat{\beta}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

Conditional Variance of $\hat{\boldsymbol{\beta}}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}]$$

Conditional Variance of $\hat{\boldsymbol{\beta}}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$
\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}]
\end{aligned}
$$

Conditional Variance of $\hat{\beta}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$
\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X})
\end{aligned}
$$

Conditional Variance of $\hat{\beta}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$
\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

Conditional Variance of $\hat{\beta}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$
\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by homoskedasticity)}
\end{aligned}
$$

Conditional Variance of $\hat{\beta}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$
\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by homoskedasticity)} \\
&= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}
$$

Conditional Variance of $\hat{\beta}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by homoskedasticity)} \\
&= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

Conditional Variance of $\hat{\beta}$

$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\beta}|\mathbf{X}] = \beta + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \beta$ so the OLS estimator is a linear function of the errors. Thus:

$$\begin{aligned}
V[\hat{\beta}|\mathbf{X}] &= V[\beta|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by homoskedasticity)} \\
&= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

This gives the $(k+1) \times (k+1)$ variance-covariance matrix of $\hat{\beta}$.

Conditional Variance of $\hat{\boldsymbol{\beta}}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ and $E[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \boldsymbol{\beta} + E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] = \boldsymbol{\beta}$ so the OLS estimator is a linear function of the errors. Thus:

$$\begin{aligned}
V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \quad (\mathbf{X} \text{ is nonrandom given } \mathbf{X}) \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \text{(by homoskedasticity)} \\
&= \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
&= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

This gives the $(k+1) \times (k+1)$ variance-covariance matrix of $\hat{\boldsymbol{\beta}}$.

To estimate $V[\hat{\boldsymbol{\beta}}|\mathbf{X}]$, we replace σ^2 with its unbiased estimator $\hat{\sigma}^2$, which is now written using matrix notation as:

$$\hat{\sigma}^2 = \frac{SSR}{n-(k+1)} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-(k+1)}$$

Variance-covariance matrix of $\hat{\beta}$

The variance-covariance matrix of the OLS estimators is given by:

$$V[\hat{\beta}|\mathbf{X}] = \sigma^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} =$$

	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\cdots$	$\widehat{\beta}_k$
$\widehat{\beta}_0$	$V[\widehat{\beta}_0]$	$\text{Cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$\text{Cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\cdots$	$\text{Cov}[\widehat{\beta}_0, \widehat{\beta}_k]$
$\widehat{\beta}_1$	$\text{Cov}[\widehat{\beta}_0, \widehat{\beta}_1]$	$V[\widehat{\beta}_1]$	$\text{Cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$\cdots$	$\text{Cov}[\widehat{\beta}_1, \widehat{\beta}_k]$
$\widehat{\beta}_2$	$\text{Cov}[\widehat{\beta}_0, \widehat{\beta}_2]$	$\text{Cov}[\widehat{\beta}_1, \widehat{\beta}_2]$	$V[\widehat{\beta}_2]$	$\cdots$	$\text{Cov}[\widehat{\beta}_2, \widehat{\beta}_k]$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\widehat{\beta}_k$	$\text{Cov}[\widehat{\beta}_0, \widehat{\beta}_k]$	$\text{Cov}[\widehat{\beta}_k, \widehat{\beta}_1]$	$\text{Cov}[\widehat{\beta}_k, \widehat{\beta}_2]$	$\cdots$	$V[\widehat{\beta}_k]$

Consistency of $\hat{\beta}$

To show consistency, we rewrite the OLS estimator in terms of sample means so that we can apply LLN.

Consistency of $\hat{\beta}$

To show consistency, we rewrite the OLS estimator in terms of sample means so that we can apply LLN.

First, note that a matrix cross product can be written as a sum of vector products:

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i' y_i$$

where $\mathbf{x}_i$ is the $1 \times (k+1)$ row vector of predictor values for unit i.

Consistency of $\hat{\boldsymbol{\beta}}$

To show consistency, we rewrite the OLS estimator in terms of sample means so that we can apply LLN.

First, note that a matrix cross product can be written as a sum of vector products:

$$\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i \quad \text{and} \quad \mathbf{X}'\mathbf{y} = \sum_{i=1}^{n} \mathbf{x}_i' y_i$$

where $\mathbf{x}_i$ is the $1 \times (k+1)$ row vector of predictor values for unit i.

Now we can rewrite the OLS estimator as,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i' y_i\right) \\
&= \left(\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i'(\mathbf{x}_i\boldsymbol{\beta} + u_i)\right) \\
&= \boldsymbol{\beta} + \left(\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\sum_{i=1}^{n} \mathbf{x}_i' u_i\right) \\
&= \boldsymbol{\beta} + \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i'\mathbf{x}_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i' u_i\right)
\end{aligned}$$

Consistency of $\hat{\beta}$

Now let's apply the LLN to the sample means:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right) \xrightarrow{p} E[\mathbf{x}_i'\mathbf{x}_i], \text{ a } (k+1)\times(k+1) \text{ nonsingular matrix.}$$

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'u_i\right) \xrightarrow{p} E[\mathbf{x}_i'u_i] = 0, \text{ by the zero cond. mean assumption.}$$

Consistency of $\hat{\boldsymbol{\beta}}$

Now let's apply the LLN to the sample means:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right) \xrightarrow{p} E[\mathbf{x}_i'\mathbf{x}_i], \text{ a } (k+1)\times(k+1) \text{ nonsingular matrix.}$$
$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'u_i\right) \xrightarrow{p} E[\mathbf{x}_i'u_i] = 0, \text{ by the zero cond. mean assumption.}$$

Therefore, we have

$$\begin{aligned} \text{plim}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + \left(E[\mathbf{x}_i'\mathbf{x}_i]\right)^{-1}\cdot 0 \\ &= \boldsymbol{\beta}. \end{aligned}$$

Consistency of $\hat{\boldsymbol{\beta}}$

Now let's apply the LLN to the sample means:

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i'\mathbf{x}_i\right) \xrightarrow{p} E[\mathbf{x}_i'\mathbf{x}_i], \text{ a } (k+1)\times(k+1) \text{ nonsingular matrix.}$$
$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i' u_i\right) \xrightarrow{p} E[\mathbf{x}_i' u_i] = 0, \text{ by the zero cond. mean assumption.}$$

Therefore, we have

$$\begin{aligned} \text{plim}(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + \left(E[\mathbf{x}_i'\mathbf{x}_i]\right)^{-1}\cdot 0 \\ &= \boldsymbol{\beta}. \end{aligned}$$

We can also show the asymptotic normality of $\hat{\boldsymbol{\beta}}$ using a similar argument but with the CLT.

References

- Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane Jr, Michael C. Herron, and Henry E. Brady. "The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* (2001): 793-810.
- Lange, Peter, and Geoffrey Garrett. "The politics of growth: Strategic interaction and economic performance in the advanced industrial democracies, 19741980." *The Journal of Politics* 47, no. 03 (1985): 791-827.
- Jackman, Robert W. "The Politics of Economic Growth in the Industrial Democracies, 197480: Leftist Strength or North Sea Oil?." The *Journal of Politics* 49, no. 01 (1987): 242-256.
- Wooldridge, Jeffrey. *Introductory econometrics: A modern approach*. Cengage Learning, 2012.

Where We've Been and Where We're Going...

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - ⋆ a brief review of matrix algebra

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - a brief review of matrix algebra
    - matrix form of linear regression

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - a brief review of matrix algebra
    - matrix form of linear regression
  - Wednesday:

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - a brief review of matrix algebra
    - matrix form of linear regression
  - Wednesday:
    - hypothesis tests

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - a brief review of matrix algebra
    - matrix form of linear regression
  - Wednesday:
    - hypothesis tests
- Next Week
  - break!

Where We've Been and Where We're Going...

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - Monday:
    - ★ a brief review of matrix algebra
    - ★ matrix form of linear regression
  - Wednesday:
    - ★ hypothesis tests
- Next Week
  - break!
  - then ... regression in social science
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression

Questions?

Running Example: Chilean Referendum on Pinochet

- The 1988 Chilean national plebiscite was a national referendum held to determine whether or not dictator Augusto Pinochet would extend his rule for another eight-year term in office.

Running Example: Chilean Referendum on Pinochet

- The 1988 Chilean national plebiscite was a national referendum held to determine whether or not dictator Augusto Pinochet would extend his rule for another eight-year term in office.
- Data: national survey conducted in April and May of 1988 by FLACSO in Chile.

Running Example: Chilean Referendum on Pinochet

- The 1988 Chilean national plebiscite was a national referendum held to determine whether or not dictator Augusto Pinochet would extend his rule for another eight-year term in office.
- Data: national survey conducted in April and May of 1988 by FLACSO in Chile.
- Outcome: 1 if respondent intends to vote for Pinochet, 0 otherwise. We can interpret the β slopes as marginal "effects" on the probability that respondent votes for Pinochet.

Running Example: Chilean Referendum on Pinochet

- The 1988 Chilean national plebiscite was a national referendum held to determine whether or not dictator Augusto Pinochet would extend his rule for another eight-year term in office.
- Data: national survey conducted in April and May of 1988 by FLACSO in Chile.
- Outcome: 1 if respondent intends to vote for Pinochet, 0 otherwise. We can interpret the β slopes as marginal "effects" on the probability that respondent votes for Pinochet.
- Plebiscite was held on October 5, 1988. The No side won with 56% of the vote, with 44% voting Yes.

Running Example: Chilean Referendum on Pinochet

- The 1988 Chilean national plebiscite was a national referendum held to determine whether or not dictator Augusto Pinochet would extend his rule for another eight-year term in office.
- Data: national survey conducted in April and May of 1988 by FLACSO in Chile.
- Outcome: 1 if respondent intends to vote for Pinochet, 0 otherwise. We can interpret the β slopes as marginal "effects" on the probability that respondent votes for Pinochet.
- Plebiscite was held on October 5, 1988. The No side won with 56% of the vote, with 44% voting Yes.
- We model the intended Pinochet vote as a linear function of gender, education, and age of respondents.

Hypothesis Testing in R

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
~~~~~
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4875 on 1699 degrees of freedom
Multiple R-squared: 0.05112,       Adjusted R-squared: 0.04945
F-statistic: 30.51 on 3 and 1699 DF,  p-value: < 2.2e-16
```

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient β_j:

$$H_0: \ \beta_j \ = \ c$$

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient β_j:

$$H_0 : \ \beta_j \ = \ c$$

- In the simple linear regression we used the t-value to test this kind of hypothesis.

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient β_j:

$$H_0 : \beta_j = c$$

- In the simple linear regression we used the t-value to test this kind of hypothesis.
- We can consider the same t-value about β_j for the multiple regression:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}(\hat{\beta}_j)}$$

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient β_j:

$$H_0: \beta_j = c$$

- In the simple linear regression we used the t-value to test this kind of hypothesis.
- We can consider the same t-value about β_j for the multiple regression:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}(\hat{\beta}_j)}$$

- How do we compute $\hat{SE}(\hat{\beta}_j)$?

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient β_j:

$$H_0: \beta_j = c$$

- In the simple linear regression we used the t-value to test this kind of hypothesis.
- We can consider the same t-value about β_j for the multiple regression:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}(\hat{\beta}_j)}$$

- How do we compute $\hat{SE}(\hat{\beta}_j)$?

$$\hat{SE}(\hat{\beta}_j) = \sqrt{\widehat{V}(\hat{\beta}_j)} = \sqrt{\widehat{V}(\hat{\boldsymbol{\beta}})_{(j,j)}} = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{(j,j)}}$$

where $\mathbf{A}_{(j,j)}$ is the (j,j) element of matrix $\mathbf{A}$.

The t-Value for Multiple Linear Regression

- Consider testing a hypothesis about a single regression coefficient β_j:

$$H_0: \beta_j = c$$

- In the simple linear regression we used the t-value to test this kind of hypothesis.
- We can consider the same t-value about β_j for the multiple regression:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}(\hat{\beta}_j)}$$

- How do we compute $\hat{SE}(\hat{\beta}_j)$?

$$\hat{SE}(\hat{\beta}_j) = \sqrt{\widehat{V}(\hat{\beta}_j)} = \sqrt{\widehat{V}(\hat{\boldsymbol{\beta}})_{(j,j)}} = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}_{(j,j)}}$$

where $\mathbf{A}_{(j,j)}$ is the (j, j) element of matrix $\mathbf{A}$.
That is, take the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ and square root the diagonal element corresponding to j.

Hypothesis Testing in R

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
```

We can pull out the variance-covariance matrix $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ in R from the `lm()` object:

Hypothesis Testing in R

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
```

We can pull out the variance-covariance matrix $\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$ in R from the lm() object:

R Code

```
> V <- vcov(fit)
> V
              (Intercept)           fem          educ           age
(Intercept)  2.642311e-03 -3.455498e-04 -5.270913e-04 -3.357119e-05
fem         -3.455498e-04  5.623170e-04  2.249973e-05  8.285291e-07
educ        -5.270913e-04  2.249973e-05  1.922354e-04  3.411049e-06
age         -3.357119e-05  8.285291e-07  3.411049e-06  6.914098e-07

> sqrt(diag(V))
 (Intercept)          fem         educ          age
0.0514034097 0.0237132251 0.0138648980 0.0008315105
```

Sampling Distribution of the t-Value

Sampling Distribution of the t-Value

The t-values in multiple regressions essentially have the same statistical properties as the simple regression case. That is,

Sampling Distribution of the t-Value

The t-values in multiple regressions essentially have the same statistical properties as the simple regression case. That is,

Theorem (Small-Sample Distribution of the t-Value)

Under Assumptions 1–6, for any sample size n the t-value has the t distribution with $(n-k-1)$ degrees of freedom:

$$T = \frac{\widehat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

Sampling Distribution of the t-Value

The t-values in multiple regressions essentially have the same statistical properties as the simple regression case. That is,

Theorem (Small-Sample Distribution of the t-Value)

Under Assumptions 1–6, for any sample size n the t-value has the t distribution with $(n-k-1)$ degrees of freedom:

$$T = \frac{\widehat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

Theorem (Large-Sample Distribution of the t-Value)

Under Assumptions 1–5, as $n \to \infty$ the distribution of the t-value approaches the standard normal distribution:

$$T = \frac{\widehat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \overset{a.}{\sim} \mathcal{N}(0,1) \quad as \quad n \to \infty$$

Sampling Distribution of the t-Value

The t-values in multiple regressions essentially have the same statistical properties as the simple regression case. That is,

Theorem (Small-Sample Distribution of the t-Value)

Under Assumptions 1–6, for any sample size n the t-value has the t distribution with $(n-k-1)$ degrees of freedom:

$$T = \frac{\widehat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

Theorem (Large-Sample Distribution of the t-Value)

Under Assumptions 1–5, as $n \to \infty$ the distribution of the t-value approaches the standard normal distribution:

$$T = \frac{\widehat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \overset{a.}{\sim} \mathcal{N}(0,1) \quad \textit{as} \quad n \to \infty$$

- $t_{n-k-1} \to \mathcal{N}(0,1)$ as $n \to \infty$, so the difference disappears when n large.

Sampling Distribution of the t-Value

The t-values in multiple regressions essentially have the same statistical properties as the simple regression case. That is,

Theorem (Small-Sample Distribution of the t-Value)

Under Assumptions 1–6, for any sample size n the t-value has the t distribution with $(n-k-1)$ degrees of freedom:

$$T = \frac{\widehat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

Theorem (Large-Sample Distribution of the t-Value)

Under Assumptions 1–5, as $n \to \infty$ the distribution of the t-value approaches the standard normal distribution:

$$T = \frac{\widehat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \overset{a.}{\sim} \mathcal{N}(0,1) \quad as \quad n \to \infty$$

- $t_{n-k-1} \to \mathcal{N}(0,1)$ as $n \to \infty$, so the difference disappears when n large.
- In practice people often just use t_{n-k-1} to be on the conservative side.

Using the t-Value as a Test Statistic

Using the t-Value as a Test Statistic

The procedure for testing this null hypothesis ($\beta_j = c$) is identical to the simple regression case, except that our reference distribution is t_{n-k-1} instead of t_{n-2}.

Using the t-Value as a Test Statistic

The procedure for testing this null hypothesis ($\beta_j = c$) is identical to the simple regression case, except that our reference distribution is t_{n-k-1} instead of t_{n-2}.

1. Compute the t-value as $T = (\hat{\beta}_j - c)/\hat{SE}[\hat{\beta}_j]$

Using the t-Value as a Test Statistic

The procedure for testing this null hypothesis ($\beta_j = c$) is identical to the simple regression case, except that our reference distribution is t_{n-k-1} instead of t_{n-2}.

1. Compute the t-value as $T = (\hat{\beta}_j - c)/\hat{SE}[\hat{\beta}_j]$

2. Compare the value to the critical value $t_{\alpha/2}$ for the α level test, which under the null hypothesis satisfies

$$P\left(-t_{\alpha/2} \leq T \leq t_{\alpha/2}\right) = 1 - \alpha$$

Using the t-Value as a Test Statistic

The procedure for testing this null hypothesis ($\beta_j = c$) is identical to the simple regression case, except that our reference distribution is t_{n-k-1} instead of t_{n-2}.

1. Compute the t-value as $T = (\hat{\beta}_j - c)/\hat{SE}[\hat{\beta}_j]$
2. Compare the value to the critical value $t_{\alpha/2}$ for the α level test, which under the null hypothesis satisfies

$$P\left(-t_{\alpha/2} \leq T \leq t_{\alpha/2}\right) = 1 - \alpha$$

3. Decide whether the realized value of T in our data is unusual given the known distribution of the test statistic.

Using the t-Value as a Test Statistic

The procedure for testing this null hypothesis ($\beta_j = c$) is identical to the simple regression case, except that our reference distribution is t_{n-k-1} instead of t_{n-2}.

1. Compute the t-value as $T = (\hat{\beta}_j - c)/\hat{SE}[\hat{\beta}_j]$
2. Compare the value to the critical value $t_{\alpha/2}$ for the α level test, which under the null hypothesis satisfies

$$P\left(-t_{\alpha/2} \leq T \leq t_{\alpha/2}\right) = 1 - \alpha$$

3. Decide whether the realized value of T in our data is unusual given the known distribution of the test statistic.
4. Finally, either declare that we reject H_0 or not, or report the p-value.

Confidence Intervals

To construct confidence intervals, there is again no difference compared to the case of $k = 1$, except that we need to use t_{n-k-1} instead of t_{n-2}

Confidence Intervals

To construct confidence intervals, there is again no difference compared to the case of $k = 1$, except that we need to use t_{n-k-1} instead of t_{n-2}

Since we know the sampling distribution for our t-value:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

Confidence Intervals

To construct confidence intervals, there is again no difference compared to the case of $k = 1$, except that we need to use t_{n-k-1} instead of t_{n-2}

Since we know the sampling distribution for our t-value:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

So we also know the probability that the value of our test statistics falls into a given interval:

$$P\left(-t_{\alpha/2} \leq \frac{\widehat{\beta}_j - \beta_j}{\hat{SE}[\hat{\beta}_j]} \leq t_{\alpha/2}\right) = 1 - \alpha$$

Confidence Intervals

To construct confidence intervals, there is again no difference compared to the case of $k = 1$, except that we need to use t_{n-k-1} instead of t_{n-2}

Since we know the sampling distribution for our t-value:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

So we also know the probability that the value of our test statistics falls into a given interval:

$$P\left(-t_{\alpha/2} \leq \frac{\widehat{\beta}_j - \beta_j}{\hat{SE}[\hat{\beta}_j]} \leq t_{\alpha/2}\right) = 1 - \alpha$$

We rearrange:

$$\left[\widehat{\beta}_j - t_{\alpha/2}\hat{SE}[\hat{\beta}_j]\,,\, \widehat{\beta}_j + t_{\alpha/2}\hat{SE}[\hat{\beta}_j]\right]$$

Confidence Intervals

To construct confidence intervals, there is again no difference compared to the case of $k = 1$, except that we need to use t_{n-k-1} instead of t_{n-2}

Since we know the sampling distribution for our t-value:

$$T = \frac{\hat{\beta}_j - c}{\hat{SE}[\hat{\beta}_j]} \sim t_{n-k-1}$$

So we also know the probability that the value of our test statistics falls into a given interval:

$$P\left(-t_{\alpha/2} \leq \frac{\widehat{\beta}_j - \beta_j}{\hat{SE}[\hat{\beta}_j]} \leq t_{\alpha/2}\right) = 1 - \alpha$$

We rearrange:

$$\left[\widehat{\beta}_j - t_{\alpha/2}\hat{SE}[\hat{\beta}_j]\, ,\, \widehat{\beta}_j + t_{\alpha/2}\hat{SE}[\hat{\beta}_j]\right]$$

and thus can construct the confidence intervals as usual using:

$$\hat{\beta}_j \pm t_{\alpha/2} \cdot \hat{SE}[\hat{\beta}_j]$$

Confidence Intervals in R

R Code

```
> fit <- lm(vote1 ~ fem + educ + age, data = d)
> summary(fit)
~~~~~
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4042284  0.0514034   7.864 6.57e-15 ***
fem          0.1360034  0.0237132   5.735 1.15e-08 ***
educ        -0.0607604  0.0138649  -4.382 1.25e-05 ***
age          0.0037786  0.0008315   4.544 5.90e-06 ***
---
```

R Code

```
> confint(fit)
                   2.5 %      97.5 %
(Intercept)  0.303407780  0.50504909
fem          0.089493169  0.18251357
educ        -0.087954435 -0.03356629
age          0.002147755  0.00540954
```

Testing Hypothesis About a Linear Combination of β_j

Testing Hypothesis About a Linear Combination of β_j

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica      -2552.8     1204.5  -2.119   0.0372 *
RegionAsia          148.9     1149.8   0.129   0.8973
RegionLatAmerica   -271.3     1007.0  -0.269   0.7883
RegionOecd         9671.3     1007.0   9.604 5.74e-15 ***
```

Testing Hypothesis About a Linear Combination of β_j

R Code
```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica        -2552.8     1204.5  -2.119   0.0372 *
RegionAsia            148.9     1149.8   0.129   0.8973
RegionLatAmerica     -271.3     1007.0  -0.269   0.7883
RegionOecd           9671.3     1007.0   9.604 5.74e-15 ***
```

- $\hat{\beta}_{Asia}$ and $\hat{\beta}_{LAm}$ are close. So we may want to test the null hypothesis:

$$H_0: \ \beta_{LAm} = \beta_{Asia} \ \Leftrightarrow \ \beta_{LAm} - \beta_{Asia} = 0$$

Testing Hypothesis About a Linear Combination of β_j

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica       -2552.8     1204.5  -2.119   0.0372 *
RegionAsia           148.9     1149.8   0.129   0.8973
RegionLatAmerica    -271.3     1007.0  -0.269   0.7883
RegionOecd          9671.3     1007.0   9.604 5.74e-15 ***
```

- $\hat{\beta}_{Asia}$ and $\hat{\beta}_{LAm}$ are close. So we may want to test the null hypothesis:

$$H_0: \ \beta_{LAm} = \beta_{Asia} \ \Leftrightarrow \ \beta_{LAm} - \beta_{Asia} = 0$$

against the alternative of

$$H_1: \ \beta_{LAm} \neq \beta_{Asia} \ \Leftrightarrow \ \beta_{LAm} - \beta_{Asia} \neq 0$$

Testing Hypothesis About a Linear Combination of β_j

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica       -2552.8     1204.5  -2.119   0.0372 *
RegionAsia           148.9     1149.8   0.129   0.8973
RegionLatAmerica    -271.3     1007.0  -0.269   0.7883
RegionOecd          9671.3     1007.0   9.604 5.74e-15 ***
```

- $\hat{\beta}_{Asia}$ and $\hat{\beta}_{LAm}$ are close. So we may want to test the null hypothesis:

$$H_0: \ \beta_{LAm} = \beta_{Asia} \ \Leftrightarrow \ \beta_{LAm} - \beta_{Asia} = 0$$

  against the alternative of

$$H_1: \ \beta_{LAm} \neq \beta_{Asia} \ \Leftrightarrow \ \beta_{LAm} - \beta_{Asia} \neq 0$$

- What would be an appropriate test statistic for this hypothesis?

Testing Hypothesis About a Linear Combination of β_j

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica       -2552.8     1204.5  -2.119   0.0372 *
RegionAsia           148.9     1149.8   0.129   0.8973
RegionLatAmerica    -271.3     1007.0  -0.269   0.7883
RegionOecd          9671.3     1007.0   9.604 5.74e-15 ***
```

- Let's consider a t-value:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})}$$

We will reject H_0 if T is sufficiently different from zero.

Testing Hypothesis About a Linear Combination of β_j

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> summary(fit)

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)        4452.7      783.4   5.684 2.07e-07 ***
RegionAfrica      -2552.8     1204.5  -2.119   0.0372 *
RegionAsia          148.9     1149.8   0.129   0.8973
RegionLatAmerica   -271.3     1007.0  -0.269   0.7883
RegionOecd         9671.3     1007.0   9.604 5.74e-15 ***
```

- Let's consider a t-value:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})}$$

  We will reject H_0 if T is sufficiently different from zero.

- Note that unlike the test of a single hypothesis, both $\hat{\beta}_{LAm}$ and $\hat{\beta}_{Asia}$ are random variables, hence the denominator.

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \;\sim\; t_{n-k-1}$$

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$?

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$? No!

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$? No!
- Is it $\hat{SE}(\hat{\beta}_{LAm}) + \hat{SE}(\hat{\beta}_{Asia})$?

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$? No!
- Is it $\hat{SE}(\hat{\beta}_{LAm}) + \hat{SE}(\hat{\beta}_{Asia})$? No!

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$? No!
- Is it $\hat{SE}(\hat{\beta}_{LAm}) + \hat{SE}(\hat{\beta}_{Asia})$? No!
- Recall the following property of the variance:

$$V(X \pm Y) = V(X) + V(Y) \pm 2\text{Cov}(X, Y)$$

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$? No!
- Is it $\hat{SE}(\hat{\beta}_{LAm}) + \hat{SE}(\hat{\beta}_{Asia})$? No!
- Recall the following property of the variance:

$$V(X \pm Y) = V(X) + V(Y) \pm 2\text{Cov}(X, Y)$$

Therefore, the standard error for a linear combination of coefficients is:

$$\hat{SE}(\widehat{\beta}_1 \pm \widehat{\beta}_2) = \sqrt{\widehat{V}(\widehat{\beta}_1) + \widehat{V}(\widehat{\beta}_2) \pm 2\widehat{\text{Cov}}[\widehat{\beta}_1, \widehat{\beta}_2]}$$

which we can calculate from the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

Testing Hypothesis About A Linear Combination of β_j

- Our test statistic:

$$T = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \sim t_{n-k-1}$$

- How do you find $\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})$?
- Is it $\hat{SE}(\hat{\beta}_{LAm}) - \hat{SE}(\hat{\beta}_{Asia})$? No!
- Is it $\hat{SE}(\hat{\beta}_{LAm}) + \hat{SE}(\hat{\beta}_{Asia})$? No!
- Recall the following property of the variance:

$$V(X \pm Y) = V(X) + V(Y) \pm 2\text{Cov}(X, Y)$$

Therefore, the standard error for a linear combination of coefficients is:

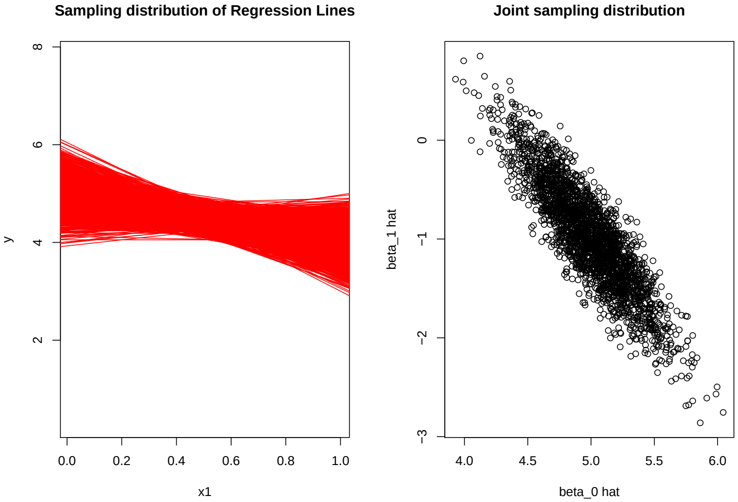
$$\hat{SE}(\widehat{\beta}_1 \pm \widehat{\beta}_2) = \sqrt{\widehat{V}(\widehat{\beta}_1) + \widehat{V}(\widehat{\beta}_2) \pm 2\widehat{\text{Cov}}[\widehat{\beta}_1, \widehat{\beta}_2]}$$

which we can calculate from the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

- Since the estimates of the coefficients are correlated, we need the covariance term.

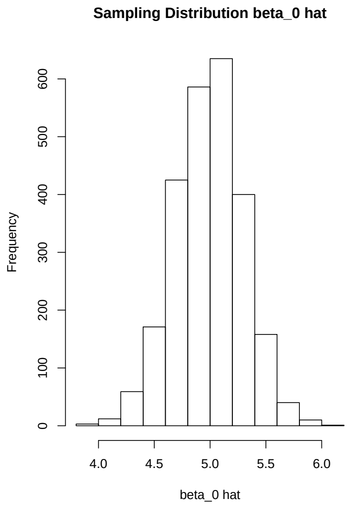
Joint Normality: Simulation

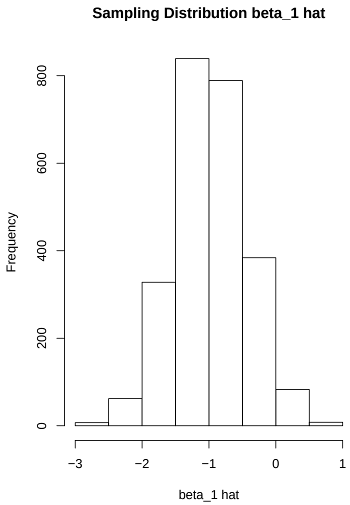
$Y = \beta_0 + \beta_1 X_1 + u$ with $u \sim N(0, \sigma_u^2 = 4)$ and $\beta_0 = 5$, $\beta_1 = -1$, and $n = 100$:



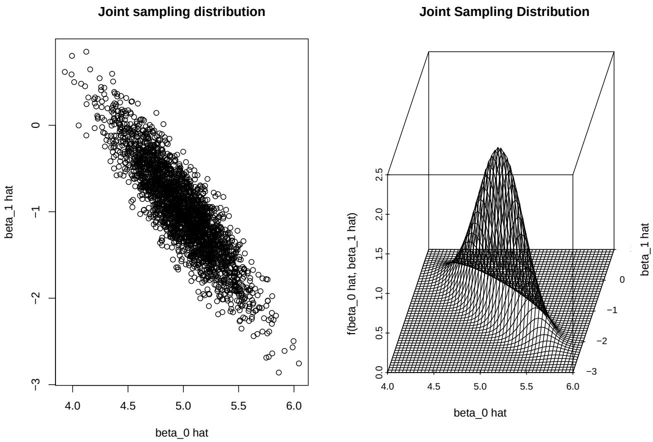
Marginal Sampling Distribution

$Y = \beta_0 + \beta_1 X_1 + u$ with $u \sim N(0, \sigma_u^2 = 4)$ and $\beta_0 = 5$, $\beta_1 = -1$, and $n = 100$:





Joint Sampling Distribution



The variance-covariance matrix of the estimators is:

	$\widehat{\beta}_0$	$\widehat{\beta}_1$
$\widehat{\beta}_0$	.08	$-.11$
$\widehat{\beta}_1$	$-.11$	.24

Example: GDP per capita on Regions

Example: GDP per capita on Regions

R Code

```
> fit <- lm(REALGDPCAP ~ Region, data = D)
> V <- vcov(fit)
> V
               (Intercept) RegionAfrica RegionAsia RegionLatAmerica
(Intercept)       613769.9    -613769.9  -613769.9        -613769.9
RegionAfrica     -613769.9    1450728.8   613769.9         613769.9
RegionAsia       -613769.9     613769.9  1321965.9         613769.9
RegionLatAmerica -613769.9     613769.9   613769.9        1014054.6
RegionOecd       -613769.9     613769.9   613769.9         613769.9
                RegionOecd
(Intercept)      -613769.9
RegionAfrica      613769.9
RegionAsia        613769.9
RegionLatAmerica  613769.9
RegionOecd       1014054.6
```

Example: GDP per capita on Regions

We can then compute the test statistic for the hypothesis of interest:

Example: GDP per capita on Regions

We can then compute the test statistic for the hypothesis of interest:

R Code

```
> se <- sqrt(V[4,4] + V[3,3] - 2*V[3,4])
> se
[1] 1052.844
>
> tstat <- (coef(fit)[4] - coef(fit)[3])/se
> tstat
RegionLatAmerica
      -0.3990977
```

$$t = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \quad \text{where}$$

$$\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}) = \sqrt{\widehat{V}(\widehat{\beta}_{LAm}) + \widehat{V}(\widehat{\beta}_{Asia}) - 2\widehat{\text{Cov}}[\widehat{\beta}_{LAm}, \widehat{\beta}_{Asia}]}$$

Plugging in we get $t \simeq -0.40$. So what do we conclude?

Example: GDP per capita on Regions

We can then compute the test statistic for the hypothesis of interest:

R Code

```
> se <- sqrt(V[4,4] + V[3,3] - 2*V[3,4])
> se
[1] 1052.844
>
> tstat <- (coef(fit)[4] - coef(fit)[3])/se
> tstat
RegionLatAmerica
      -0.3990977
```

$$t = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \quad \text{where}$$

$$\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}) = \sqrt{\widehat{V}(\widehat{\beta}_{LAm}) + \widehat{V}(\widehat{\beta}_{Asia}) - 2\widehat{\text{Cov}}[\widehat{\beta}_{LAm}, \widehat{\beta}_{Asia}]}$$

Plugging in we get $t \simeq -0.40$. So what do we conclude?

We cannot reject the null that the difference in average GDP resulted from chance.

F Test for Joint Significance of Coefficients

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)
- Suppose our regression model is:

$$Voted = \beta_0 + \gamma_1 FEMALE + \beta_1 EDUCATION +$$

$$\gamma_2 (FEMALE \cdot EDUCATION) + \beta_2 AGE + \gamma_3 (FEMALE \cdot AGE) + u$$

and we want to test

$$H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)
- Suppose our regression model is:

$$Voted = \beta_0 + \gamma_1 FEMALE + \beta_1 EDUCATION +$$

$$\gamma_2 (FEMALE \cdot EDUCATION) + \beta_2 AGE + \gamma_3 (FEMALE \cdot AGE) + u$$

  and we want to test

$$H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

- Substantively, what question are we asking?

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)
- Suppose our regression model is:

$$Voted = \beta_0 + \gamma_1 FEMALE + \beta_1 EDUCATION +$$

$$\gamma_2 (FEMALE \cdot EDUCATION) + \beta_2 AGE + \gamma_3 (FEMALE \cdot AGE) + u$$

  and we want to test

$$H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

- Substantively, what question are we asking?

  $\rightarrow$ Do females and males vote systematically differently from each other? (Under the null, there is no difference in either the intercept or slopes between females and males).

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)
- Suppose our regression model is:

$$Voted = \beta_0 + \gamma_1 FEMALE + \beta_1 EDUCATION +$$

$$\gamma_2 (FEMALE \cdot EDUCATION) + \beta_2 AGE + \gamma_3 (FEMALE \cdot AGE) + u$$

  and we want to test

$$H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

- Substantively, what question are we asking?

  $\rightarrow$ Do females and males vote systematically differently from each other? (Under the null, there is no difference in either the intercept or slopes between females and males).
- This is an example of a joint hypothesis test involving three restrictions: $\gamma_1 = 0$, $\gamma_2 = 0$, and $\gamma_3 = 0$.

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)
- Suppose our regression model is:

$$Voted = \beta_0 + \gamma_1 FEMALE + \beta_1 EDUCATION +$$

$$\gamma_2(FEMALE \cdot EDUCATION) + \beta_2 AGE + \gamma_3(FEMALE \cdot AGE) + u$$

  and we want to test

$$H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

- Substantively, what question are we asking?

  $\rightarrow$ Do females and males vote systematically differently from each other? (Under the null, there is no difference in either the intercept or slopes between females and males).
- This is an example of a joint hypothesis test involving three restrictions: $\gamma_1 = 0$, $\gamma_2 = 0$, and $\gamma_3 = 0$.
- If all the interaction terms and the group lower order term are close to zero, then we fail to reject the null hypothesis of no gender difference.

F Test for Joint Significance of Coefficients

- In research we often want to test a joint hypothesis which involves multiple linear restrictions (e.g. $\beta_1 = \beta_2 = \beta_3 = 0$)
- Suppose our regression model is:

$$Voted = \beta_0 + \gamma_1 FEMALE + \beta_1 EDUCATION +$$

$$\gamma_2 (FEMALE \cdot EDUCATION) + \beta_2 AGE + \gamma_3 (FEMALE \cdot AGE) + u$$

  and we want to test

$$H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

- Substantively, what question are we asking?

  $\rightarrow$ Do females and males vote systematically differently from each other? (Under the null, there is no difference in either the intercept or slopes between females and males).
- This is an example of a joint hypothesis test involving three restrictions: $\gamma_1 = 0$, $\gamma_2 = 0$, and $\gamma_3 = 0$.
- If all the interaction terms and the group lower order term are close to zero, then we fail to reject the null hypothesis of no gender difference.
- F tests allows us to to test joint hypothesis

The χ^2 Distribution

The χ^2 Distribution

- To test more than one hypothesis jointly we need to introduce some new probability distributions.

The χ^2 Distribution

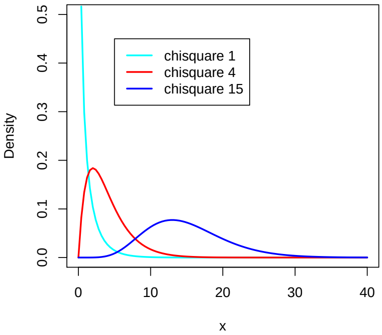
- To test more than one hypothesis jointly we need to introduce some new probability distributions.
- Suppose $Z_1, ..., Z_n$ are n i.i.d. random variables following $\mathcal{N}(0, 1)$.

The χ^2 Distribution

- To test more than one hypothesis jointly we need to introduce some new probability distributions.
- Suppose $Z_1, ..., Z_n$ are n i.i.d. random variables following $\mathcal{N}(0, 1)$.
- Then, the sum of their squares, $X = \sum_{i=1}^{n} Z_i^2$, is distributed according to the χ^2 distribution with n degrees of freedom, $X \sim \chi_n^2$.

The χ^2 Distribution

- To test more than one hypothesis jointly we need to introduce some new probability distributions.
- Suppose $Z_1, ..., Z_n$ are n i.i.d. random variables following $\mathcal{N}(0, 1)$.
- Then, the sum of their squares, $X = \sum_{i=1}^{n} Z_i^2$, is distributed according to the χ^2 distribution with n degrees of freedom, $X \sim \chi^2_n$.



Properties: $X > 0$, $E[X] = n$ and $V[X] = 2n$. In R: `dchisq()`, `pchisq()`, `rchisq()`

The F distribution

The F distribution

The F distribution arises as a ratio of two independent chi-squared distributed random variables:

$$F = \frac{X_1/df_1}{X_2/df_2} \sim \mathcal{F}_{df_1, df_2}$$

where $X_1 \sim \chi^2_{df_1}$, $X_2 \sim \chi^2_{df_2}$, and $X_1 \perp\!\!\!\perp X_2$.

The F distribution

The F distribution arises as a ratio of two independent chi-squared distributed random variables:

$$F = \frac{X_1/df_1}{X_2/df_2} \sim \mathcal{F}_{df_1, df_2}$$

where $X_1 \sim \chi^2_{df_1}$, $X_2 \sim \chi^2_{df_2}$, and $X_1 \perp\!\!\!\perp X_2$.

df_1 and df_2 are called the numerator degrees of freedom and the denominator degrees of freedom.

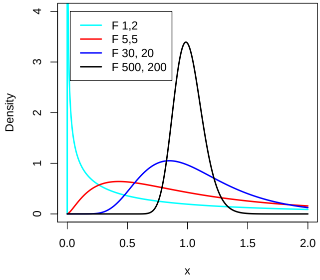
The F distribution

The F distribution arises as a ratio of two independent chi-squared distributed random variables:

$$F = \frac{X_1/df_1}{X_2/df_2} \sim \mathcal{F}_{df_1, df_2}$$

where $X_1 \sim \chi^2_{df_1}$, $X_2 \sim \chi^2_{df_2}$, and $X_1 \perp\!\!\!\perp X_2$.

df_1 and df_2 are called the numerator degrees of freedom and the denominator degrees of freedom.



In R: `df()`, `pf()`, `rf()`

F Test against $H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$

F Test against $H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

F Test against $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

   $$Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2 (FEM * EDUC) + \beta_2 AGE + \gamma_3 (FEM * AGE) + u$$

F Test against $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

   $$Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2 (FEM * EDUC) + \beta_2 AGE + \gamma_3 (FEM * AGE) + u$$

2. Fit the Restricted Model (R) which *does* impose H_0:

   $$Vote = \beta_0 + \beta_1 EDUC + \beta_2 AGE + u$$

F Test against $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

   $$Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2 (FEM * EDUC) + \beta_2 AGE + \gamma_3 (FEM * AGE) + u$$

2. Fit the Restricted Model (R) which *does* impose H_0:

   $$Vote = \beta_0 + \beta_1 EDUC + \beta_2 AGE + u$$

3. From the two results, compute the F Statistic:

   $$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

   where SSR=sum of squared residuals, q=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

F Test against $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

   $Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2 (FEM * EDUC) + \beta_2 AGE + \gamma_3 (FEM * AGE) + u$

2. Fit the Restricted Model (R) which *does* impose H_0:

   $$Vote = \beta_0 + \beta_1 EDUC + \beta_2 AGE + u$$

3. From the two results, compute the F Statistic:

   $$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

   where SSR=sum of squared residuals, q=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

   Intuition:

   $$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

F Test against $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

   $$Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2 (FEM * EDUC) + \beta_2 AGE + \gamma_3 (FEM * AGE) + u$$

2. Fit the Restricted Model (R) which *does* impose H_0:

   $$Vote = \beta_0 + \beta_1 EDUC + \beta_2 AGE + u$$

3. From the two results, compute the F Statistic:

   $$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

   where SSR=sum of squared residuals, q=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

   Intuition:

   $$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

The F statistics have the following sampling distributions:

F Test against $H_0: \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

   $$Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2 (FEM * EDUC) + \beta_2 AGE + \gamma_3 (FEM * AGE) + u$$

2. Fit the Restricted Model (R) which *does* impose H_0:

   $$Vote = \beta_0 + \beta_1 EDUC + \beta_2 AGE + u$$

3. From the two results, compute the F Statistic:

   $$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

   where SSR=sum of squared residuals, q=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

   Intuition:

   $$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

The F statistics have the following sampling distributions:

- Under Assumptions 1–6, $F_0 \sim \mathcal{F}_{q,n-k-1}$ regardless of the sample size.

F Test against $H_0: \ \gamma_1 = \gamma_2 = \gamma_3 = 0.$

The F statistic can be calculated by the following procedure:

1. Fit the Unrestricted Model (UR) which *does not* impose H_0:

$$Vote = \beta_0 + \gamma_1 FEM + \beta_1 EDUC + \gamma_2 (FEM * EDUC) + \beta_2 AGE + \gamma_3 (FEM * AGE) + u$$

2. Fit the Restricted Model (R) which *does* impose H_0:

$$Vote = \beta_0 + \beta_1 EDUC + \beta_2 AGE + u$$

3. From the two results, compute the F Statistic:

$$F_0 = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

where SSR=sum of squared residuals, q=number of restrictions, k=number of predictors in the unrestricted model, and n= # of observations.

Intuition:

$$\frac{\text{increase in prediction error}}{\text{original prediction error}}$$

The F statistics have the following sampling distributions:

- Under Assumptions 1–6, $F_0 \sim \mathcal{F}_{q,n-k-1}$ regardless of the sample size.
- Under Assumptions 1–5, $qF_0 \overset{a.}{\sim} \chi^2_q$ as $n \to \infty$ (see next section).

Unrestricted Model (UR)

R Code

```
> fit.UR <- lm(vote1 ~ fem + educ + age + fem:age + fem:educ, data = Chile
> summary(fit.UR)
~~~~~
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.293130   0.069242   4.233 2.42e-05 ***
fem          0.368975   0.098883   3.731 0.000197 ***
educ        -0.038571   0.019578  -1.970 0.048988 *
age          0.005482   0.001114   4.921 9.44e-07 ***
fem:age     -0.003779   0.001673  -2.259 0.024010 *
fem:educ    -0.044484   0.027697  -1.606 0.108431
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.487 on 1697 degrees of freedom
Multiple R-squared: 0.05451,        Adjusted R-squared: 0.05172
F-statistic: 19.57 on 5 and 1697 DF,  p-value: < 2.2e-16
```

Restricted Model (R)

R Code

```
> fit.R <- lm(vote1 ~ educ + age, data = Chile)
> summary(fit.R)
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.4878039  0.0497550   9.804  < 2e-16 ***
educ        -0.0662022  0.0139615  -4.742 2.30e-06 ***
age          0.0035783  0.0008385   4.267 2.09e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.4921 on 1700 degrees of freedom
Multiple R-squared:  0.03275,       Adjusted R-squared:  0.03161
F-statistic: 28.78 on 2 and 1700 DF,  p-value: 5.097e-13
```

F Test in R

R Code
```
> SSR.UR <- sum(resid(fit.UR)^2)  # = 402
> SSR.R <- sum(resid(fit.R)^2)    # = 411

> DFdenom <- df.residual(fit.UR)  # = 1703
> DFnum <- 3

> F <- ((SSR.R - SSR.UR)/DFnum) / (SSR.UR/DFdenom)
> F
[1] 13.01581

> qf(0.99, DFnum, DFdenom)
[1] 3.793171
```

Given above, what do we conclude?

F Test in R

R Code

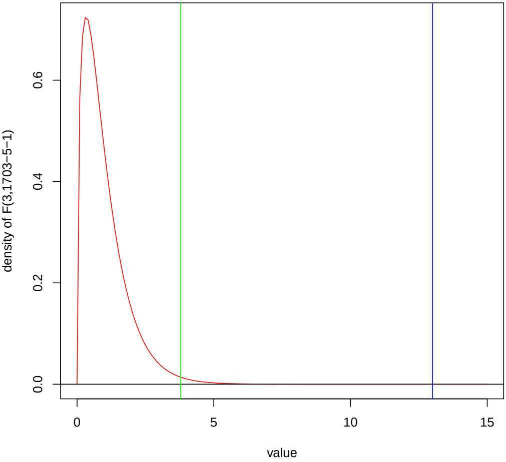
```
> SSR.UR <- sum(resid(fit.UR)^2)  # = 402
> SSR.R <- sum(resid(fit.R)^2)    # = 411

> DFdenom <- df.residual(fit.UR)  # = 1703
> DFnum <- 3

> F <- ((SSR.R - SSR.UR)/DFnum) / (SSR.UR/DFdenom)
> F
[1] 13.01581

> qf(0.99, DFnum, DFdenom)
[1] 3.793171
```

Given above, what do we conclude?

$F_0 = 13$ is greater than the critical value for a .01 level test. So we *reject* the null hypothesis.

Null Distribution, Critical Value, and Test Statistic

Note that the F statistic is always positive, so we only look at the right tail of the reference F (or χ^2 in a large sample) distribution.



F Test Examples I

F Test Examples I

The F test can be used to test various joint hypotheses which involve multiple linear restrictions.

F Test Examples I

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

F Test Examples I

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

We may want to test:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

F Test Examples I

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

We may want to test:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

- What question are we asking?

F Test Examples I

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

We may want to test:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

- What question are we asking?

  $\rightarrow$ Does any of the X variables help to predict Y?

F Test Examples I

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

We may want to test:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

- What question are we asking?

  $\rightarrow$ Does any of the X variables help to predict Y?

- This is called the omnibus test and is routinely reported by statistical software.

Omnibus Test in R

R Code

```
> summary(fit.UR)
~~~~~

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.293130   0.069242   4.233 2.42e-05 ***
fem          0.368975   0.098883   3.731 0.000197 ***
educ        -0.038571   0.019578  -1.970 0.048988 *
age          0.005482   0.001114   4.921 9.44e-07 ***
fem:age     -0.003779   0.001673  -2.259 0.024010 *
fem:educ    -0.044484   0.027697  -1.606 0.108431
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.487 on 1697 degrees of freedom
Multiple R-squared: 0.05451,        Adjusted R-squared: 0.05172
F-statistic: 19.57 on 5 and 1697 DF,  p-value: < 2.2e-16
```

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- What question are we asking?

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- What question are we asking?

  $\rightarrow$ Are the effects of X_1, X_2 and X_3 different from each other?

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- What question are we asking?
  $\rightarrow$ Are the effects of X_1, X_2 and X_3 different from each other?
- How many restrictions?

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- What question are we asking?

  $\rightarrow$ Are the effects of X_1, X_2 and X_3 different from each other?

- How many restrictions?

  $\rightarrow$ Two ($\beta_1 - \beta_2 = 0$ and $\beta_2 - \beta_3 = 0$)

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- What question are we asking?
  $\rightarrow$ Are the effects of X_1, X_2 and X_3 different from each other?
- How many restrictions?
  $\rightarrow$ Two ($\beta_1 - \beta_2 = 0$ and $\beta_2 - \beta_3 = 0$)
- How do we fit the restricted model?

F Test Examples II

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + u$$

Next, let's consider:

$$H_0 : \beta_1 = \beta_2 = \beta_3$$

- What question are we asking?
  $\rightarrow$ Are the effects of X_1, X_2 and X_3 different from each other?
- How many restrictions?
  $\rightarrow$ Two ($\beta_1 - \beta_2 = 0$ and $\beta_2 - \beta_3 = 0$)
- How do we fit the restricted model?
  $\rightarrow$ The null hypothesis implies that the model can be written as:
  $$Y = \beta_0 + \beta_1(X_1 + X_2 + X_3) + ... + \beta_k X_k + u$$
  So we create a new variable $X^* = X_1 + X_2 + X_3$ and fit:
  $$Y = \beta_0 + \beta_1 X^* + ... + \beta_k X_k + u$$

Testing Equality of Coefficients in R

R Code

```
> fit.UR2 <- lm(REALGDPCAP ~ Asia + LatAmerica + Transit + Oecd, data = D)
> summary(fit.UR2)
~~~~~

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1899.9      914.9   2.077   0.0410 *
Asia          2701.7     1243.0   2.173   0.0327 *
LatAmerica    2281.5     1112.3   2.051   0.0435 *
Transit       2552.8     1204.5   2.119   0.0372 *
Oecd         12224.2     1112.3  10.990   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3034 on 80 degrees of freedom
Multiple R-squared: 0.7096,        Adjusted R-squared: 0.6951
F-statistic: 48.88 on 4 and 80 DF,  p-value: < 2.2e-16
```

Are the coefficients on *Asia*, *LatAmerica* and *Transit* statistically significantly different?

Testing Equality of Coefficients in R

R Code

```
> D$Xstar <- D$Asia + D$LatAmerica + D$Transit
> fit.R2 <- lm(REALGDPCAP ~ Xstar + Oecd, data = D)

> SSR.UR2 <- sum(resid(fit.UR2)^2)
> SSR.R2 <- sum(resid(fit.R2)^2)

> DFdenom <- df.residual(fit.UR2)

> F <- ((SSR.R2 - SSR.UR2)/2) / (SSR.UR2/DFdenom)
> F
[1] 0.08786129

> pf(F, 2, DFdenom, lower.tail = F)
[1] 0.9159762
```

So, what do we conclude?

Testing Equality of Coefficients in R

R Code

```
> D$Xstar <- D$Asia + D$LatAmerica + D$Transit
> fit.R2 <- lm(REALGDPCAP ~ Xstar + Oecd, data = D)

> SSR.UR2 <- sum(resid(fit.UR2)^2)
> SSR.R2 <- sum(resid(fit.R2)^2)

> DFdenom <- df.residual(fit.UR2)

> F <- ((SSR.R2 - SSR.UR2)/2) / (SSR.UR2/DFdenom)
> F
[1] 0.08786129

> pf(F, 2, DFdenom, lower.tail = F)
[1] 0.9159762
```

So, what do we conclude?
The three coefficients are statistically indistinguishable from each other, with the p-value of 0.916.

t Test vs. F Test

Consider the hypothesis test of

$$H_0: \ \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \ \beta_1 \neq \beta_2$$

What ways have we learned to conduct this test?

t Test vs. F Test

Consider the hypothesis test of

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

What ways have we learned to conduct this test?

- Option 1: Compute $T = (\hat{\beta}_1 - \hat{\beta}_2)/\hat{SE}(\hat{\beta}_1 - \hat{\beta}_2)$ and do the t test.

t Test vs. F Test

Consider the hypothesis test of

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

What ways have we learned to conduct this test?

- Option 1: Compute $T = (\hat{\beta}_1 - \hat{\beta}_2)/\hat{SE}(\hat{\beta}_1 - \hat{\beta}_2)$ and do the t test.
- Option 2: Create $X^* = X_1 + X_2$, fit the restricted model, compute $F = (SSR_R - SSR_{UR})/(SSR_R/(n-k-1))$ and do the F test.

t Test vs. F Test

Consider the hypothesis test of

$$H_0: \ \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \ \beta_1 \neq \beta_2$$

What ways have we learned to conduct this test?

- Option 1: Compute $T = (\hat{\beta}_1 - \hat{\beta}_2)/\hat{SE}(\hat{\beta}_1 - \hat{\beta}_2)$ and do the t test.
- Option 2: Create $X^* = X_1 + X_2$, fit the restricted model, compute $F = (SSR_R - SSR_{UR})/(SSR_R/(n-k-1))$ and do the F test.

It turns out these two tests give identical results. This is because

$$X \sim t_{n-k-1} \quad \Longleftrightarrow \quad X^2 \sim \mathcal{F}_{1,n-k-1}$$

t Test vs. F Test

Consider the hypothesis test of

$$H_0: \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \beta_1 \neq \beta_2$$

What ways have we learned to conduct this test?

- Option 1: Compute $T = (\hat{\beta}_1 - \hat{\beta}_2)/\hat{SE}(\hat{\beta}_1 - \hat{\beta}_2)$ and do the t test.
- Option 2: Create $X^* = X_1 + X_2$, fit the restricted model, compute $F = (SSR_R - SSR_{UR})/(SSR_R/(n-k-1))$ and do the F test.

It turns out these two tests give identical results. This is because

$$X \sim t_{n-k-1} \iff X^2 \sim \mathcal{F}_{1,n-k-1}$$

- So, for testing a single hypothesis it does not matter whether one does a t test or an F test.

t Test vs. F Test

Consider the hypothesis test of

$$H_0: \ \beta_1 = \beta_2 \quad \text{vs.} \quad H_1: \ \beta_1 \neq \beta_2$$

What ways have we learned to conduct this test?

- Option 1: Compute $T = (\hat{\beta}_1 - \hat{\beta}_2)/\hat{SE}(\hat{\beta}_1 - \hat{\beta}_2)$ and do the t test.
- Option 2: Create $X^* = X_1 + X_2$, fit the restricted model, compute $F = (SSR_R - SSR_{UR})/(SSR_R/(n-k-1))$ and do the F test.

It turns out these two tests give identical results. This is because

$$X \sim t_{n-k-1} \iff X^2 \sim \mathcal{F}_{1,n-k-1}$$

- So, for testing a single hypothesis it does not matter whether one does a t test or an F test.
- Usually, the t test is used for single hypotheses and the F test is used for joint hypotheses.

Some More Notes on F Tests

- The F-value can also be calculated from R^2:

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n - k - 1)}$$

Some More Notes on F Tests

- The F-value can also be calculated from R^2:

$$F = \frac{(R^2_{UR} - R^2_R)/q}{(1 - R^2_{UR})/(n-k-1)}$$

- F tests only work for testing nested models, i.e. the restricted model must be a special case of the unrestricted model.

  For example F tests cannot be used to test

$$Y = \beta_0 + \beta_1 X_1 \qquad\qquad + \beta_3 X_3 + u$$

  against

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \qquad\quad + u$$

Some More Notes on F Tests

- Joint significance does not necessarily imply the significance of individual coefficients, or vice versa:

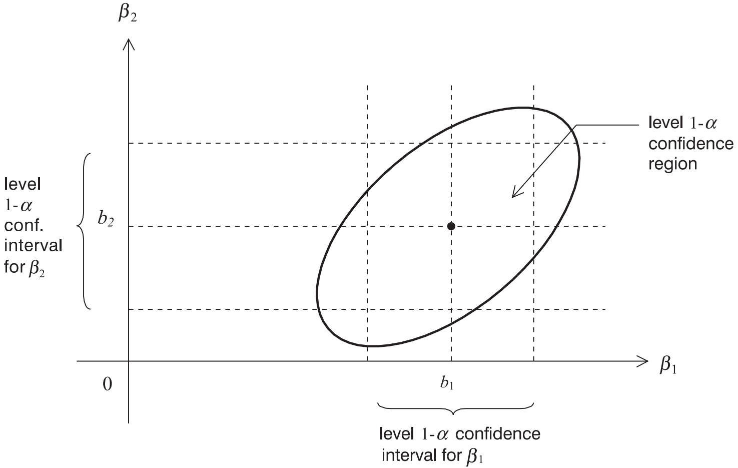


Figure 1.5: t- versus F-Tests

Limitation of the F Formula

Consider the following null hypothesis:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 3$$

or

$$H_0: \ \beta_1 = 2\beta_2 = 0.5\beta_3 + 1$$

Limitation of the F Formula

Consider the following null hypothesis:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 3$$

or

$$H_0: \ \beta_1 = 2\beta_2 = 0.5\beta_3 + 1$$

Can we test them using the F test?

Limitation of the F Formula

Consider the following null hypothesis:

$$H_0 : \ \beta_1 = \beta_2 = \beta_3 = 3$$

or

$$H_0 : \ \beta_1 = 2\beta_2 = 0.5\beta_3 + 1$$

Can we test them using the F test?
To compute the F value, we need to fit the restricted model. How?

Limitation of the F Formula

Consider the following null hypothesis:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 3$$

or

$$H_0: \ \beta_1 = 2\beta_2 = 0.5\beta_3 + 1$$

Can we test them using the F test?
To compute the F value, we need to fit the restricted model. How?

- Some restrictions are difficult to impose when fitting the model.

Limitation of the F Formula

Consider the following null hypothesis:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 3$$

or

$$H_0: \ \beta_1 = 2\beta_2 = 0.5\beta_3 + 1$$

Can we test them using the F test?
To compute the F value, we need to fit the restricted model. How?

- Some restrictions are difficult to impose when fitting the model.
- Even when we can, the procedure will be ad hoc and require some creativity.

Limitation of the F Formula

Consider the following null hypothesis:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = 3$$

or

$$H_0: \ \beta_1 = 2\beta_2 = 0.5\beta_3 + 1$$

Can we test them using the F test?
To compute the F value, we need to fit the restricted model. How?

- Some restrictions are difficult to impose when fitting the model.
- Even when we can, the procedure will be ad hoc and require some creativity.
- Is there a general solution?

General Procedure for Testing Linear Hypotheses

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

where

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

  where

  - $\mathbf{R}$ is a $q \times (k+1)$ matrix of prespecified coefficients on $\boldsymbol{\beta}$ (hypothesis matrix)

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

  where

  - $\mathbf{R}$ is a $q \times (k+1)$ matrix of prespecified coefficients on $\boldsymbol{\beta}$ (hypothesis matrix)
  - $\boldsymbol{\beta} = [\beta_0 \; \beta_1 \; \cdots \; \beta_k]'$

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

where

  - $\mathbf{R}$ is a $q \times (k+1)$ matrix of prespecified coefficients on $\boldsymbol{\beta}$ (hypothesis matrix)
  - $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \cdots \ \beta_k]'$
  - $\mathbf{r}$ is a $q \times 1$ vector of prespecified constants

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

  where
  - $\mathbf{R}$ is a $q \times (k+1)$ matrix of prespecified coefficients on $\boldsymbol{\beta}$ (hypothesis matrix)
  - $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \cdots \ \beta_k]'$
  - $\mathbf{r}$ is a $q \times 1$ vector of prespecified constants

- Examples:

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

  where

  - $\mathbf{R}$ is a $q \times (k+1)$ matrix of prespecified coefficients on $\boldsymbol{\beta}$ (hypothesis matrix)
  - $\boldsymbol{\beta} = [\beta_0 \; \beta_1 \; \cdots \; \beta_k]'$
  - $\mathbf{r}$ is a $q \times 1$ vector of prespecified constants

- Examples:

$$\beta_1 = \beta_2 = \beta_3 = 3 \Leftrightarrow \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

General Procedure for Testing Linear Hypotheses

- Notice that any set of q linear hypotheses can be written as

$$\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$$

  where
  - $\mathbf{R}$ is a $q \times (k+1)$ matrix of prespecified coefficients on $\boldsymbol{\beta}$ (hypothesis matrix)
  - $\boldsymbol{\beta} = [\beta_0 \ \beta_1 \ \cdots \ \beta_k]'$
  - $\mathbf{r}$ is a $q \times 1$ vector of prespecified constants

- Examples:

$$\beta_1 = \beta_2 = \beta_3 = 3 \Leftrightarrow \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

$$\beta_1 = 2\beta_2 = 0.5\beta_3 + 1 \Leftrightarrow \begin{bmatrix} \beta_1 - 2\beta_2 \\ \beta_1 - 0.5\beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \Leftrightarrow \begin{bmatrix} 0 & 1 & -2 & 0 \\ 0 & 1 & 0 & -0.5 \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.
- It turns out that the following Wald statistic can be used:

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \cdot \left[\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \cdot \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)$$

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.
- It turns out that the following Wald statistic can be used:

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \cdot \left[\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \cdot \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)$$

- Looks complicated? Let's figure out why this makes sense:

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.
- It turns out that the following Wald statistic can be used:

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \cdot \left[\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \cdot \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)$$

- Looks complicated? Let's figure out why this makes sense:
  - The first and last components give the sum of squares of the components of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This summarizes its deviation from zero.

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.
- It turns out that the following Wald statistic can be used:

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \cdot \left[\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \cdot \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)$$

- Looks complicated? Let's figure out why this makes sense:
  - The first and last components give the sum of squares of the components of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This summarizes its deviation from zero.
  - The middle component is the variance of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This standardizes the sum of squares to have variance one.

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.
- It turns out that the following Wald statistic can be used:

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \cdot \left[\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \cdot \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)$$

- Looks complicated? Let's figure out why this makes sense:
  - The first and last components give the sum of squares of the components of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This summarizes its deviation from zero.
  - The middle component is the variance of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This standardizes the sum of squares to have variance one.
- We know $\hat{\boldsymbol{\beta}}$ is approximately normal $\Rightarrow \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should also be normal $\Longrightarrow W$ should therefore be ...

Wald Statistic

- Let's consider testing $H_0 : \mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, a set of q linear restrictions.
- If H_0 is true, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should be zero except for sampling variability.
- To formally evaluate the statistical significance of the deviation from zero, we must transform $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ to a statistic that can be compared to a reference distribution.
- It turns out that the following Wald statistic can be used:

$$W = \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)' \cdot \left[\hat{\sigma}^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right]^{-1} \cdot \left(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}\right)$$

- Looks complicated? Let's figure out why this makes sense:
  - The first and last components give the sum of squares of the components of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This summarizes its deviation from zero.
  - The middle component is the variance of $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. This standardizes the sum of squares to have variance one.
- We know $\hat{\boldsymbol{\beta}}$ is approximately normal $\Rightarrow \mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$ should also be normal $\Longrightarrow W$ should therefore be ... χ^2 distributed!

Sampling Distribution of the Wald Statistic

Sampling Distribution of the Wald Statistic

Theorem (Large-Sample Distribution of the Wald Statistic)

Under Assumptions 1–5, as $n \to \infty$ the distribution of the Wald statistic approaches the chi square distribution with q degrees of freedom:

$$W \overset{d}{\to} \chi^2_q \quad \text{as} \quad n \to \infty$$

Sampling Distribution of the Wald Statistic

Theorem (Large-Sample Distribution of the Wald Statistic)

Under Assumptions 1–5, as $n \to \infty$ the distribution of the Wald statistic approaches the chi square distribution with q degrees of freedom:

$$W \xrightarrow{d} \chi^2_q \quad \text{as} \quad n \to \infty$$

Theorem (Small-Sample Distribution of the Wald Statistic)

Under Assumptions 1–6, for any sample size n the Wald statistic divided by q has the F distribution with $(q, n-k-1)$ degrees of freedom:

$$W/q \sim \mathcal{F}_{q,n-k-1}$$

Sampling Distribution of the Wald Statistic

Theorem (Large-Sample Distribution of the Wald Statistic)

Under Assumptions 1–5, as $n \to \infty$ the distribution of the Wald statistic approaches the chi square distribution with q degrees of freedom:

$$W \xrightarrow{d} \chi^2_q \quad \text{as} \quad n \to \infty$$

Theorem (Small-Sample Distribution of the Wald Statistic)

Under Assumptions 1–6, for any sample size n the Wald statistic divided by q has the F distribution with $(q, n-k-1)$ degrees of freedom:

$$W/q \sim \mathcal{F}_{q,n-k-1}$$

- $q\mathcal{F}_{q,n-k-1} \xrightarrow{d} \chi^2_q$ as $n \to \infty$, so the difference disappears when n large.

```
> pf(3.1, 2, 500,lower.tail=F) [1] 0.04591619

> pchisq(2*3.1, 2,lower.tail=F) [1] 0.0450492

> pf(3.1, 2, 50000,lower.tail=F) [1] 0.04505786
```

Testing General Linear Hypotheses in R

In R, the `linearHypothesis()` function in the car package does the Wald test for general linear hypotheses.

R Code

```
> fit.UR2 <- lm(REALGDPCAP ~ Asia + LatAmerica + Transit + Oecd, data = D)
> R <- matrix(c(0,1,-1,0,0, 0,1,0,-1,0), nrow = 2, byrow = T)
> r <- c(0,0)
> linearHypothesis(fit.UR2, R, r)
Linear hypothesis test

Hypothesis:
Asia - LatAmerica = 0
Asia - Transit = 0

Model 1: restricted model
Model 2: REALGDPCAP ~ Asia + LatAmerica + Transit + Oecd

  Res.Df       RSS Df Sum of Sq      F Pr(>F)
1     82 738141635
2     80 736523836  2   1617798 0.0879  0.916
```

Next Week (of Classes)

Next Week (of Classes)

- Linear Regression in the Social Sciences
- Reading:

Next Week (of Classes)

- Linear Regression in the Social Sciences
- Reading:
  - Healy and Moody (2014) "Data Visualization in Sociology" *Annual Review of Sociology*
  - Morgan and Winship (2015) Chapter 1: Causality and Empirical Research in the Social Sciences
  - Morgan and Winship (2015) Chapter 13.1: Objections to Adoption of the Counterfactual Approach
  - Optional: Morgan and Winship (2015) Chapter 2-3 (Potential Outcomes and Causal Graphs)
  - Optional: Hernán and Robins (2016) Chapter 1: A definition of a causal effect.

Fun Without Weights

The Robust Beauty of Improper Linear Models in Decision Making

ROBYN M. DAWES *University of Oregon*

ABSTRACT: *Proper linear models are those in which predictor variables are given weights in such a way that the resulting linear composite optimally predicts some criterion of interest; examples of proper linear models are standard regression analysis, discriminant function analysis, and ridge regression analysis. Research summarized in Paul Meehl's book on clinical versus statistical prediction—and a plethora of research stimulated in part by that book—all indicates that when a numerical criterion variable (e.g., graduate grade point average) is to be predicted from numerical predictor variables, proper linear models outperform clinical intuition. Improper linear models are those in which the weights of the predictor variables are obtained by some nonoptimal method; for example, they may be obtained on the basis of intuition, derived from simulating a clinical judge's predictions, or set to be equal. This article presents evidence that even such improper linear models are superior to clinical intuition when predicting a numerical criterion from numerical predictors. In fact, unit (i.e., equal) weighting is quite robust for making such predictions. The article discusses, in some detail, the application of unit weights to decide what bullet the Denver Police Department should use. Finally, the article considers commonly raised technical, psychological, and ethical*

A *proper linear model* is one in which the weights given to the predictor variables are chosen in such a way as to optimize the relationship between the prediction and the criterion. Simple regression analysis is the most common example of a proper linear model; the predictor variables are weighted in such a way as to maximize the correlation between the subsequent weighted composite and the actual criterion. Discriminant function analysis is another example of a proper linear model; weights are given to the predictor variables in such a way that the resulting linear composites maximize the discrepancy between two or more groups. Ridge regression analysis, another example (Darlington, 1978; Marquardt & Snee, 1975), attempts to assign weights in such a way that the linear composites correlate maximally with the criterion of interest in a new set of data.

Thus, there are many types of proper linear models and they have been used in a variety of contexts. One example (Dawes, 1971) was presented in this Journal; it involved the prediction

Improper Linear Models

- Proper linear model is one where predictor variables are given optimized weights in some way (for example through regression)

Improper Linear Models

- Proper linear model is one where predictor variables are given optimized weights in some way (for example through regression)
- Meehl (1954) *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence* argued that proper linear models outperform clinical intuition in many areas.

Improper Linear Models

- Proper linear model is one where predictor variables are given optimized weights in some way (for example through regression)
- Meehl (1954) *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence* argued that proper linear models outperform clinical intuition in many areas.
- Dawes argues that even improper linear models (those where weights are set by hand or set to be equal), outperform clinical intuition.

Improper Linear Models

- Proper linear model is one where predictor variables are given optimized weights in some way (for example through regression)
- Meehl (1954) *Clinical Versus Statistical Prediction: A Theoretical Analysis and Review of the Evidence* argued that proper linear models outperform clinical intuition in many areas.
- Dawes argues that even improper linear models (those where weights are set by hand or set to be equal), outperform clinical intuition.
- Equal weight models are argued to be quite robust for these predictions

Example: Graduate Admissions

- Faculty rated all students in the psych department at University of Oregon

Example: Graduate Admissions

- Faculty rated all students in the psych department at University of Oregon
- Ratings predicted from a proper linear model of student GRE scores, undergrad GPA and selectivity of student's undergraduate institution. Cross-validated correlation was .38

Example: Graduate Admissions

- Faculty rated all students in the psych department at University of Oregon
- Ratings predicted from a proper linear model of student GRE scores, undergrad GPA and selectivity of student's undergraduate institution. Cross-validated correlation was .38
- Correlation of faculty ratings with average rating of admissions committee was .19

Example: Graduate Admissions

- Faculty rated all students in the psych department at University of Oregon
- Ratings predicted from a proper linear model of student GRE scores, undergrad GPA and selectivity of student's undergraduate institution. Cross-validated correlation was .38
- Correlation of faculty ratings with average rating of admissions committee was .19
- Standardized and equally weighted improper linear model, correlated at .48

Other Examples

Other Examples

- Self-assessed measures of marital happiness: modeled with improper linear model of (`rate of lovemaking - rate of arguments`): correlation of .40

Other Examples

- Self-assessed measures of marital happiness: modeled with improper linear model of (`rate of lovemaking - rate of arguments`): correlation of .40
- Einhorn (1972) study of doctors coding biopsies of patients with Hodgkin's disease and then rated severity. Their rating of severity was essentially uncorrelated with survival times, but the variables they coded predicted outcomes using a regression model.

Other Examples

TABLE 1

Correlations Between Predictions and Criterion Values

Example	Average validity of judge	Average validity of judge model	Average validity of random model	Validity of equal weighting model	Cross-validity of regression analysis	Validity of optimal linear model
Prediction of neurosis vs. psychosis	.28	.31	.30	.34	.46	.46
Illinois students' predictions of GPA	.33	.50	.51	.60	.57	.69
Oregon students' predictions of GPA	.37	.43	.51	.60	.57	.69
Prediction of later faculty ratings at Oregon	.19	.25	.39	.48	.38	.54
Yntema & Torgerson's (1961) experiment	.84	.89	.84	.97	—	.97

Note. GPA = grade point average.

Column descriptions:

C1) average of human judges

C2) model based on human judges

C3) randomly chosen weights preserving signs

C4) equal weighting

C5) cross-validated weights

C6) unattainable optimal linear model

The Argument

- "People – especially the experts in a field – are much better at selecting and coding information than they are at integrating it." (573)

The Argument

- "People – especially the experts in a field – are much better at selecting and coding information than they are at integrating it." (573)
- The choice of variables is extremely important for prediction!

The Argument

- "People – especially the experts in a field – are much better at selecting and coding information than they are at integrating it." (573)
- The choice of variables is extremely important for prediction!
- This parallels a piece of folk wisdom in the machine learning literature that a better predictor will beat a better model every time.

The Argument

- "People – especially the experts in a field – are much better at selecting and coding information than they are at integrating it." (573)
- The choice of variables is extremely important for prediction!
- This parallels a piece of folk wisdom in the machine learning literature that a better predictor will beat a better model every time.
- People are good at picking out relevant information, but terrible at integrating it.

The Argument

- "People – especially the experts in a field – are much better at selecting and coding information than they are at integrating it." (573)
- The choice of variables is extremely important for prediction!
- This parallels a piece of folk wisdom in the machine learning literature that a better predictor will beat a better model every time.
- People are good at picking out relevant information, but terrible at integrating it.
- The difficulty arises in part because people in general lack a strong reference to the distribution of the predictors.

The Argument

- "People – especially the experts in a field – are much better at selecting and coding information than they are at integrating it." (573)
- The choice of variables is extremely important for prediction!
- This parallels a piece of folk wisdom in the machine learning literature that a better predictor will beat a better model every time.
- People are good at picking out relevant information, but terrible at integrating it.
- The difficulty arises in part because people in general lack a strong reference to the distribution of the predictors.
- Linear models are robust to deviations from the optimal weights (see also Waller 2008 on "Fungible Weights in Multiple Regression")

My Thoughts on the Argument

- Particularly in prediction, looking for the true or right model can be quixotic

My Thoughts on the Argument

- Particularly in prediction, looking for the true or right model can be quixotic
- The broader research project suggests that a big part of what quantitative models are doing predictively, is focusing human talent in the right place.

My Thoughts on the Argument

- Particularly in prediction, looking for the true or right model can be quixotic
- The broader research project suggests that a big part of what quantitative models are doing predictively, is focusing human talent in the right place.
- This all applies because predictors well chosen and the sample size is small (so the weight optimization isn't great)

My Thoughts on the Argument

- Particularly in prediction, looking for the true or right model can be quixotic
- The broader research project suggests that a big part of what quantitative models are doing predictively, is focusing human talent in the right place.
- This all applies because predictors well chosen and the sample size is small (so the weight optimization isn't great)
- It is a fascinating paper!