Week 7: Multiple Regression

Brandon Stewart[1]

Princeton

October 12–16, 2020

[1] These slides are heavily influenced by Matt Blackwell, Adam Glynn, Justin Grimmer, Jens Hainmueller and Erin Hartman.

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
  - matrix form of linear regression
  - inference and hypothesis tests

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

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - matrix form of linear regression
  - inference and hypothesis tests
- Next Week
  - diagnostics

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

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - matrix form of linear regression
  - inference and hypothesis tests
- Next Week
  - diagnostics
- Long Run
  - probability → inference → regression → causal inference

1. Matrix Form of Regression
   - Estimation
   - Fun With(out) Weights
2. OLS Classical Inference in Matrix Form
   - Unbiasedness
   - Classical Standard Errors
3. Agnostic Inference
4. Standard Hypothesis Tests
   - t-Tests
   - Adjusted R^2
   - F Tests for Joint Significance

1. Matrix Form of Regression
   - Estimation
   - Fun With(out) Weights

2. OLS Classical Inference in Matrix Form
   - Unbiasedness
   - Classical Standard Errors

3. Agnostic Inference

4. Standard Hypothesis Tests
   - *t*-Tests
   - Adjusted R^2
   - F Tests for Joint Significance

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 $\mathbf{X}$ and β be the following:

Grouping Things into Matrices

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

$$\underset{(4\times3)}{\mathbf{X}} = \begin{bmatrix} 1 & x_1 & z_1 \\ 1 & x_2 & z_2 \\ 1 & x_3 & z_3 \\ 1 & x_4 & z_4 \end{bmatrix} \quad \underset{(3\times1)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

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 $\mathbf{X}$:

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

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:

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

- 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:

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

- Rearranging:

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

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

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:

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:

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

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:

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

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

- Recall:

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

Back to OLS

- Recall:

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

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

Back to OLS

- Recall:

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

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

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

Back to OLS

- Recall:
$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{y}$$

- Let's assume, for now, that the inverse of $\mathbf{X}'\mathbf{X}$ exists
- Then we can write the OLS estimator as the following:
$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- See Aronow and Miller Theorem 4.1.4 for proof.

Back to OLS

- Recall:

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

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

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

- See Aronow and Miller Theorem 4.1.4 for proof.
- "ex prime ex inverse ex prime y"

Back to OLS

- Recall:

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

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

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

- See Aronow and Miller Theorem 4.1.4 for proof.
- "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

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

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 approximately 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 approximately composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is approximately 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 approximately composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is approximately composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

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

i.e. analogous to the simple linear regression case!

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 approximately composed of the covariances between the columns of $\mathbf{X}$ and $\mathbf{y}$
- "Denominator" $\mathbf{X}'\mathbf{X}$ is approximately composed of the sample variances and covariances of variables within $\mathbf{X}$
- Thus, we have something like:

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

i.e. analogous to the simple linear regression case!

Disclaimer: the final equation is exactly true for all non-intercept coefficients if you remove the intercept from $\mathbf{X}$ such that $\hat{\boldsymbol{\beta}}_{-0} = \text{Var}(\mathbf{X}_{-0})^{-1}\text{Cov}(\mathbf{X}_{-0}, \mathbf{y})$. The numerator and denominator are the variances and covariances if $\mathbf{X}$ and $\mathbf{y}$ are demeaned and normalized by the sample size minus 1.

Fun Without Weights

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 resistances to using linear models to make important social decisions and presents arguments that could weaken these resistances.*

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 of faculty ratings of graduate students. All gradu-

Improper Linear Models

- If you have to diagnose a disease are you better off with an expert or a statistical model?

Improper Linear Models

- If you have to diagnose a disease are you better off with an expert or a statistical model?
- 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

- If you have to diagnose a disease are you better off with an expert or a statistical model?
- 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.
- Proper linear model is one where predictor variables are given optimized weights in some way (for example through regression).

Improper Linear Models

- If you have to diagnose a disease are you better off with an expert or a statistical model?
- 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.
- Proper linear model is one where predictor variables are given optimized weights in some way (for example through regression).
- Dawes argues that even improper linear models (those where weights are set by hand or set to be equal), outperform clinical intuition.

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

Common pattern: c2, c3, c4, c5, c6 > c1

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")

Thoughts on the Argument

- Particularly in prediction, looking for the true or right model can be quixotic.

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.

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 it is hard to learn much from the data).

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 it is hard to learn much from the data).
- Dawes (1979) is an intellectual basis to support algorithmic decision making. Roughly, if simple models are better than experts, than with lots of data, complicated model could be much better than experts.

We Covered

We Covered

- Matrix notation for OLS
- Estimation mechanics

We Covered

- Matrix notation for OLS
- Estimation mechanics

Next Time: Classical Inference and Properties

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

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - matrix form of linear regression
  - inference and hypothesis tests
- Next Week
  - diagnostics
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. Matrix Form of Regression
   - Estimation
   - Fun With(out) Weights
2. OLS Classical Inference in Matrix Form
   - Unbiasedness
   - Classical Standard Errors
3. Agnostic Inference
4. Standard Hypothesis Tests
   - t-Tests
   - Adjusted R^2
   - F Tests for Joint Significance

OLS Assumptions in Matrix Form

OLS Assumptions in Matrix Form

1. Linearity: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
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: $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

OLS Assumptions in Matrix Form

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: $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

Assumption 3: No Perfect Collinearity

Assumption 3: No Perfect Collinearity

Definition (Rank)

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

Assumption 3: No Perfect Collinearity

Definition (Rank)

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

- If $\mathbf{X}$ has rank $k + 1$, then all of its columns are linearly independent

Assumption 3: No Perfect Collinearity

Definition (Rank)

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

- If **X** has rank $k+1$, then all of its columns are linearly independent
- ...If all of the columns are linearly independent, then the assumption of no perfect collinearity hold.

Assumption 3: No Perfect Collinearity

Definition (Rank)

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

- If $\mathbf{X}$ has rank $k+1$, then all of its columns are linearly independent
- ...If all of the columns are linearly independent, then the assumption of no perfect collinearity hold.
- If $\mathbf{X}$ has rank $k+1$, then $(\mathbf{X}'\mathbf{X})$ is invertible (see linear algebra book for proof)

Assumption 3: No Perfect Collinearity

Definition (Rank)

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

- If $\mathbf{X}$ has rank $k+1$, then all of its columns are linearly independent
- ...If all of the columns are linearly independent, then the assumption of no perfect collinearity hold.
- If $\mathbf{X}$ has rank $k+1$, then $(\mathbf{X}'\mathbf{X})$ is invertible (see linear algebra book for proof)
- 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:

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

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

Is $\hat{\boldsymbol{\beta}}$ still unbiased under assumptions 1-4? Does $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$?

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

Is $\hat{\boldsymbol{\beta}}$ still unbiased under assumptions 1-4? Does $E[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{u})$$

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

$$\hat{\beta} = \mathbf{I}\beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

Unbiasedness of $\hat{\beta}$

Is $\hat{\beta}$ still unbiased under assumptions 1-4? Does $E[\hat{\beta}] = \beta$?

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

So, yes!

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

A shorter (but less helpful later) proof of unbiasedness,

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

A shorter (but less helpful later) proof of unbiasedness,

$$E[E[\hat{\beta}|\mathbf{X}]] = E[E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}|\mathbf{X}]] \text{ (definition of the estimator)}$$

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

A shorter (but less helpful later) proof of unbiasedness,

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

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

A shorter (but less helpful later) proof of unbiasedness,

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

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

A shorter (but less helpful later) proof of unbiasedness,

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

Now we know the sampling distribution is centered on β we want to derive the variance of the sampling distribution conditional on X.

Assumption 5: Homoskedasticity

Assumption 5: Homoskedasticity

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

Assumption 5: Homoskedasticity

- The stated homoskedasticity assumption is: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- To really understand this we need to know what $\text{var}(\mathbf{u}|\mathbf{X})$ is in full generality.

Assumption 5: Homoskedasticity

- The stated homoskedasticity assumption is: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- To really understand this we need to know what $\text{var}(\mathbf{u}|\mathbf{X})$ is in full generality.
- The variance of a vector is actually a matrix:

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

Assumption 5: Homoskedasticity

- The stated homoskedasticity assumption is: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
- To really understand this we need to know what $\text{var}(\mathbf{u}|\mathbf{X})$ is in full generality.
- 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 always symmetric since $\text{cov}(u_i, u_j) = \text{cov}(u_j, u_i)$ by definition.

Assumption 5: The Meaning of Homoskedasticity

- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?

Assumption 5: The Meaning of Homoskedasticity

- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.

Assumption 5: The Meaning of Homoskedasticity

- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.
- Visually:

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

Assumption 5: The Meaning of Homoskedasticity

- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.
- 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:

Assumption 5: The Meaning of Homoskedasticity

- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.
- 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)

Assumption 5: The Meaning of Homoskedasticity

- What does $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$ mean?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.
- 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)

Rule: Variance of Linear Function 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.

Rule: Variance of Linear Function 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.

We will need an analogous rule for linear functions of random vectors.

Rule: Variance of Linear Function 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.

We will need an analogous rule for linear functions of random vectors.

Definition (Variance of Linear Transformation of Random Vector)

Let $f(\mathbf{u}) = \mathbf{A}\mathbf{u} + \mathbf{B}$ 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:

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

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 $\hat{\beta}$

$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}$ 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:

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

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:

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

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:

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

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:

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

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:

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

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 $\hat{\beta}$

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

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

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:

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

Sampling Variance for $\hat{\beta}$

Under assumptions 1-5, the variance-covariance matrix of the OLS estimators is given by:

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

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

Recall that standard errors are the square root of the diagonals of this matrix.

Overview of Inference in the General Setting

- Under assumption 1-5 in large samples:

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

Overview of Inference in the General Setting

- Under assumption 1-5 in large samples:

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

- In small samples, under assumptions 1-6,

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{SE}[\widehat{\beta}_j]} \sim t_{n-(k+1)}$$

Overview of Inference in the General Setting

- Under assumption 1-5 in large samples:

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

- In small samples, under assumptions 1-6,

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{SE}[\widehat{\beta}_j]} \sim t_{n-(k+1)}$$

- Estimated standard errors are:

$$\widehat{SE}[\hat{\beta}_j] = \sqrt{\widehat{\text{var}}[\widehat{\boldsymbol{\beta}}]_{jj}}$$

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

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

Overview of Inference in the General Setting

- Under assumption 1-5 in large samples:

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

- In small samples, under assumptions 1-6,

$$\frac{\widehat{\beta}_j - \beta_j}{\widehat{SE}[\widehat{\beta}_j]} \sim t_{n-(k+1)}$$

- Estimated standard errors are:

$$\widehat{SE}[\hat{\beta}_j] = \sqrt{\widehat{\text{var}}[\widehat{\boldsymbol{\beta}}]_{jj}}$$

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

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

- Thus, confidence intervals and hypothesis tests proceed in essentially the same way.

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* $\sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$*:*

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

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

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 $V[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$*
- *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 is a joint distribution because it involves multiple random variables.

Implications of the Variance-Covariance Matrix

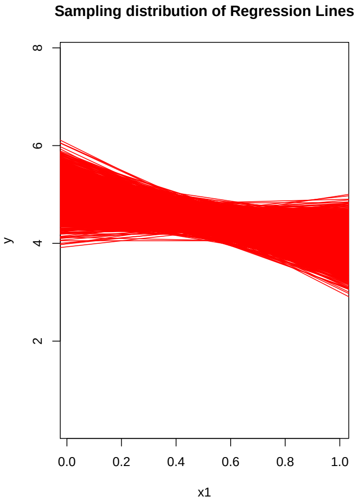
- Note that the sampling distribution is a joint distribution because it involves multiple random variables.
- This is because the sampling distribution of the terms in $\hat{\boldsymbol{\beta}}$ are correlated.

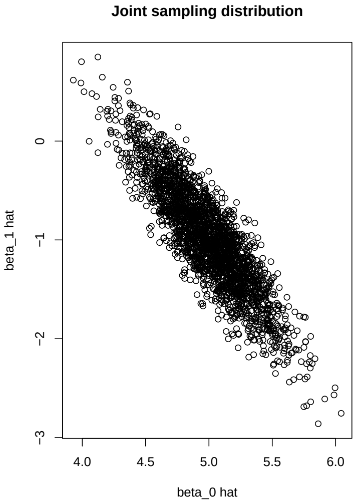
Implications of the Variance-Covariance Matrix

- Note that the sampling distribution is a joint distribution because it involves multiple random variables.
- This is because the sampling distribution of the terms in $\hat{\boldsymbol{\beta}}$ are correlated.
- In a practical sense, this means that our uncertainty about coefficients is correlated across variables.

Multivariate Normal: Simulation

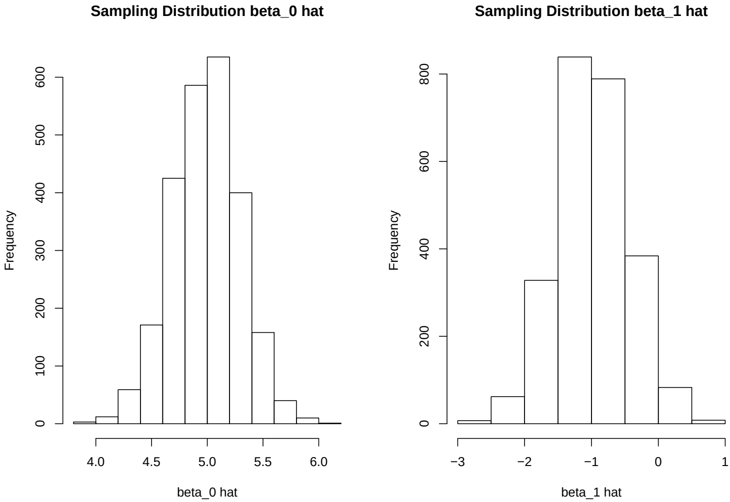
$Y = \beta_0 + \beta_1 X_1 + u$ with $u \sim N(0, \sigma_u^2 = 4)$ and $\beta_0 = 5$, $\beta_1 = -1$, and $n = 100$:





Marginals of Multivariate Normal RVs are Normal

$Y = \beta_0 + \beta_1 X_1 + u$ with $u \sim N(0, \sigma_u^2 = 4)$ and $\beta_0 = 5$, $\beta_1 = -1$, and $n = 100$:



Matrix Notation Overview

	Old notation (for univariate regression)	Matrix notation
Linear model	$Y_i = \beta_0 + \beta_1 X_i + u$	$y = X\beta + u$
Coefficient	$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$	$\hat{\beta} = (X'X)^{-1} X' y$
Homoskedasticity assumption	$Var[u \mid X] = \sigma_u^2$	$Var[u \mid X] = \sigma_u^2 I_n$
Variance of coefficient	$\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$	$\sigma_u^2 (X'X)^{-1}$
Error variance	$\frac{\sum_{i=1}^n \hat{u}_i^2}{n-2}$	$\hat{\sigma}_u^2 = \frac{\hat{u}'\hat{u}}{n-k-1}$, $\hat{u}'\hat{u} = (y - X\hat{\beta})'(y - X\hat{\beta})$
SS_{tot}	$\sum_{i=1}^n (Y_i - \bar{Y}_i)^2$	$(y - \bar{y})'(y - \bar{y})$
SS_{res}	$\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$	$\hat{u}'\hat{u}$

We Covered

We Covered

- Unbiasedness
- Classical Standard Errors

We Covered

- Unbiasedness
- Classical Standard Errors

Next Time: Agnostic Inference

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

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - matrix form of linear regression
  - inference and hypothesis tests
- Next Week
  - diagnostics
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. Matrix Form of Regression
   - Estimation
   - Fun With(out) Weights
2. OLS Classical Inference in Matrix Form
   - Unbiasedness
   - Classical Standard Errors
3. Agnostic Inference
4. Standard Hypothesis Tests
   - t-Tests
   - Adjusted R^2
   - F Tests for Joint Significance

Agnostic Perspective on the OLS estimator

- We know the population value of β is:

$$\beta = E\left[\mathbf{X}'\mathbf{X}\right]^{-1} E\left[\mathbf{X}'\mathbf{y}\right]$$

Agnostic Perspective on the OLS estimator

- We know the population value of β is:

$$\beta = E\left[\mathbf{X}'\mathbf{X}\right]^{-1} E\left[\mathbf{X}'\mathbf{y}\right]$$

- How do we get an estimator of this?

Agnostic Perspective on the OLS estimator

- We know the population value of β is:

$$\beta = E\left[\mathbf{X}'\mathbf{X}\right]^{-1} E\left[\mathbf{X}'\mathbf{y}\right]$$

- How do we get an estimator of this?
- Plug-in principle $\rightsquigarrow$ replace population expectation with sample versions:

$$\hat{\beta} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y}$$

Asymptotic OLS inference

- With this representation, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

Asymptotic OLS inference

- With this representation, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

- Core idea: $\mathbf{X}'\mathbf{u}$ is the sum of r.v.s so the CLT applies.

Asymptotic OLS inference

- With this representation, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

- Core idea: $\mathbf{X}'\mathbf{u}$ is the sum of r.v.s so the CLT applies.
- That, plus some asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Omega)$$

Asymptotic OLS inference

- With this representation, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

- Core idea: $\mathbf{X}'\mathbf{u}$ is the sum of r.v.s so the CLT applies.
- That, plus some asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Omega)$$

- The covariance matrix, Ω is given as:

$$\Omega = E[\mathbf{X}'\mathbf{X}]^{-1}E[\mathbf{X}'\text{Diag}(\mathbf{u^2})\mathbf{X}]E[\mathbf{X}'\mathbf{X}]^{-1}$$

Asymptotic OLS inference

- With this representation, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

- Core idea: $\mathbf{X}'\mathbf{u}$ is the sum of r.v.s so the CLT applies.
- That, plus some asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Omega)$$

- The covariance matrix, Ω is given as:

$$\Omega = E[\mathbf{X}'\mathbf{X}]^{-1}E[\mathbf{X}'\text{Diag}(\mathbf{u}^2)\mathbf{X}]E[\mathbf{X}'\mathbf{X}]^{-1}$$

- We will again be able to replace $\mathbf{u}$ with its empirical counterpart (the residuals) $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$, and $\mathbf{X}$ with its sample counterpart.

Asymptotic OLS inference

- With this representation, we can write the OLS estimator as follows:

$$\hat{\beta} = \beta + \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{u}$$

- Core idea: $\mathbf{X}'\mathbf{u}$ is the sum of r.v.s so the CLT applies.
- That, plus some asymptotic theory allows us to say:

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Omega)$$

- The covariance matrix, Ω is given as:

$$\Omega = E[\mathbf{X}'\mathbf{X}]^{-1}E[\mathbf{X}'\text{Diag}(\mathbf{u^2})\mathbf{X}]E[\mathbf{X}'\mathbf{X}]^{-1}$$

- We will again be able to replace $\mathbf{u}$ with its empirical counterpart (the residuals) $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\beta}$, and $\mathbf{X}$ with its sample counterpart.
- No need for assumptions A1 (linearity), A4 (conditional mean zero errors) or A5 (homoskedasticity) needed! Just IID (A2), no perfect collinearity (A3) and asymptotics.

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \mathbf{\Sigma}$

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \mathbf{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \mathbf{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- With homoskedasticity, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, we simplified

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \boldsymbol{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- With homoskedasticity, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, we simplified

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \boldsymbol{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- With homoskedasticity, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, we simplified

$$\begin{aligned}\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{ (by homoskedasticity)}\end{aligned}$$

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \boldsymbol{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- With homoskedasticity, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, we simplified

$$\begin{aligned}\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{ (by homoskedasticity)} \\ &= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\end{aligned}$$

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \mathbf{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- With homoskedasticity, $\mathbf{\Sigma} = \sigma^2\mathbf{I}$, we simplified

$$\begin{aligned}
\text{Var}[\hat{\beta}|\mathbf{X}] &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{ (by homoskedasticity)} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}$$

Stepping Back: The Classical Approach Homoskedasticity

- Remember what we did before:

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

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \boldsymbol{\Sigma}$
- Recall before we used Assumptions 1-4 to show:

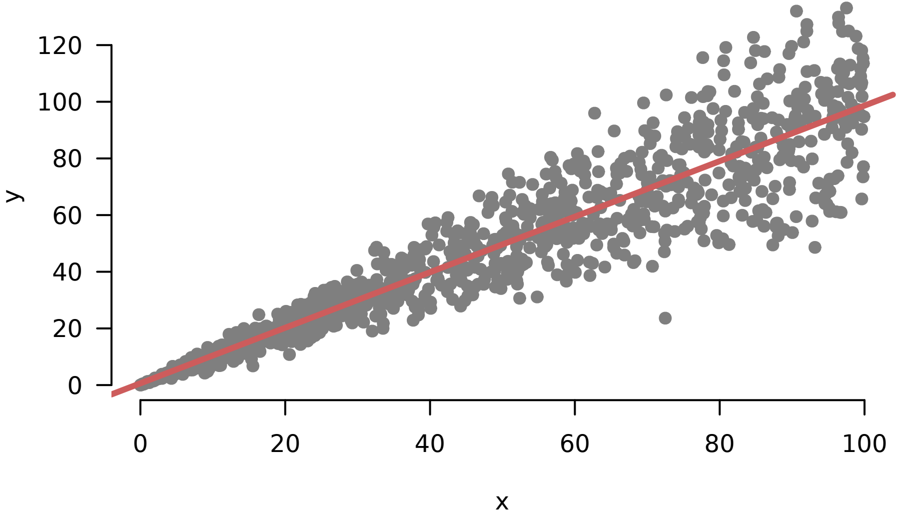
$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- With homoskedasticity, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, we simplified

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \text{ (by homoskedasticity)} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}
\end{aligned}$$

- Replace σ^2 with estimate $\widehat{\sigma}^2$ will give us our estimate of the covariance matrix

What Does This Rule Out?



Non-constant Error Variance

- Homoskedastic:

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

Non-constant Error Variance

- Homoskedastic:

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

- Heteroskedastic:

$$V[\mathbf{u}|\mathbf{X}] \;=\; \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Non-constant Error Variance

- Homoskedastic:

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

- Heteroskedastic:

$$V[\mathbf{u}|\mathbf{X}] \;=\; \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

- Independent, not identical

Non-constant Error Variance

- Homoskedastic:

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

- Heteroskedastic:

$$V[\mathbf{u}|\mathbf{X}] \;=\; \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

- Independent, not identical
- $\text{Cov}(u_i, u_j|\mathbf{X}) = 0$

Non-constant Error Variance

- Homoskedastic:

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

- Heteroskedastic:

$$V[\mathbf{u}|\mathbf{X}] \;=\; \begin{bmatrix} \sigma_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix}$$

- Independent, not identical
- $\text{Cov}(u_i, u_j|\mathbf{X}) = 0$
- $\text{Var}(u_i|\mathbf{X}) = \sigma_i^2$

Consequences of Heteroskedasticity Under Classical SEs

Consequences of Heteroskedasticity Under Classical SEs

- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}}$$

Consequences of Heteroskedasticity Under Classical SEs

- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}}$$

- α-level tests, the probability of Type I error $\neq \alpha$

Consequences of Heteroskedasticity Under Classical SEs

- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}}$$

- α-level tests, the probability of Type I error $\neq \alpha$
- Coverage of $1 - \alpha$ CIs $\neq 1 - \alpha$

Consequences of Heteroskedasticity Under Classical SEs

- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}}$$

- α-level tests, the probability of Type I error $\neq \alpha$
- Coverage of $1 - \alpha$ CIs $\neq 1 - \alpha$
- OLS is not BLUE

Consequences of Heteroskedasticity Under Classical SEs

- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}}$$

- α-level tests, the probability of Type I error $\neq \alpha$
- Coverage of $1 - \alpha$ CIs $\neq 1 - \alpha$
- OLS is not BLUE
- However:

Consequences of Heteroskedasticity Under Classical SEs

- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}}$$

- α-level tests, the probability of Type I error $\neq \alpha$
- Coverage of $1 - \alpha$ CIs $\neq 1 - \alpha$
- OLS is not BLUE
- However:
  - $\widehat{\boldsymbol{\beta}}$ still unbiased and consistent for $\boldsymbol{\beta}$

Consequences of Heteroskedasticity Under Classical SEs

- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}}$$

- α-level tests, the probability of Type I error $\neq \alpha$
- Coverage of $1 - \alpha$ CIs $\neq 1 - \alpha$
- OLS is not BLUE
- However:
  - $\widehat{\boldsymbol{\beta}}$ still unbiased and consistent for $\boldsymbol{\beta}$
  - degree of the problem depends on how serious the heteroskedasticity is

Heteroskedasticity Consistent Estimator

Heteroskedasticity Consistent Estimator

- Under non-constant error variance:

$$\text{Var}[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

Heteroskedasticity Consistent Estimator

- Under non-constant error variance:

$$\text{Var}[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

- When $\mathbf{\Sigma} \neq \sigma^2 \mathbf{I}$, we are stuck with this expression:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{\Sigma}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Heteroskedasticity Consistent Estimator

- Under non-constant error variance:

$$\text{Var}[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

- When $\mathbf{\Sigma} \neq \sigma^2 \mathbf{I}$, we are stuck with this expression:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{\Sigma}\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- Idea: If we can consistently estimate the components of $\mathbf{\Sigma}$, we could directly use this expression by replacing $\mathbf{\Sigma}$ with its estimate, $\hat{\mathbf{\Sigma}}$.

White's Heteroskedasticity Consistent Estimator

Suppose we have heteroskedasticity of unknown form (but zero covariance):

$$V[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

White's Heteroskedasticity Consistent Estimator

Suppose we have heteroskedasticity of unknown form (but zero covariance):

$$V[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

then $V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and White (1980) shows that

White's Heteroskedasticity Consistent Estimator

Suppose we have heteroskedasticity of unknown form (but zero covariance):

$$V[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

then $V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and White (1980) shows that

$$\widehat{V[\hat{\beta}|\mathbf{X}]} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

is a consistent estimator of $V[\hat{\beta}|\mathbf{X}]$ under any form of heteroskedasticity consistent with $V[\mathbf{u}]$ above.

White's Heteroskedasticity Consistent Estimator

Suppose we have heteroskedasticity of unknown form (but zero covariance):

$$V[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

then $V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and White (1980) shows that

$$\widehat{V[\hat{\beta}|\mathbf{X}]} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

is a consistent estimator of $V[\hat{\beta}|\mathbf{X}]$ under any form of heteroskedasticity consistent with $V[\mathbf{u}]$ above.

The estimate based on the above is called the heteroskedasticity consistent (HC) or robust standard errors. This also coincides with the agnostic standard errors!

Intuition for Robust Standard Errors

Core intuition: while $\widehat{\mathbf{\Sigma}}$ is an $n \times n$ matrix, $\mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix.

Intuition for Robust Standard Errors

Core intuition: while $\widehat{\mathbf{\Sigma}}$ is an $n \times n$ matrix, $\mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix. So there is hope of estimating it consistently as sample size grows *even when every true error variance is unique*.

Intuition for Robust Standard Errors

Core intuition: while $\widehat{\mathbf{\Sigma}}$ is an $n \times n$ matrix, $\mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix. So there is hope of estimating it consistently as sample size grows *even when every true error variance is unique*.

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

Intuition for Robust Standard Errors

Core intuition: while $\widehat{\mathbf{\Sigma}}$ is an $n \times n$ matrix, $\mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X}$ is a $(k+1) \times (k+1)$ matrix. So there is hope of estimating it consistently as sample size grows *even when every true error variance is unique*.

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{\Sigma}\mathbf{X} = \begin{bmatrix} \sum_i x_{i,1} x_{i,1} \hat{u}_i^2 & \sum_i x_{i,1} x_{i,2} \hat{u}_i^2 & \dots & \sum_i x_{i,1} x_{i,k+1} \hat{u}_i^2 \\ \sum_i x_{i,2} x_{i,1} \hat{u}_i^2 & \sum_i x_{i,2} x_{i,2} \hat{u}_i^2 & \dots & \sum_i x_{i,2} x_{i,k+1} \hat{u}_i^2 \\ & & \vdots & \\ \sum_i x_{i,k+1} x_{i,1} \hat{u}_i^2 & \sum_i x_{i,k+1} x_{i,2} \hat{u}_i^2 & \dots & \sum_i x_{i,k+1} x_{i,k+1} \hat{u}_i^2 \end{bmatrix}$$

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

1. Fit the regression and obtain the residuals $\hat{\mathbf{u}}$

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

1. Fit the regression and obtain the residuals $\hat{\mathbf{u}}$
2. Construct the "meat" matrix $\widehat{\boldsymbol{\Sigma}}$ with squared residuals in diagonal:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

1. Fit the regression and obtain the residuals $\hat{\mathbf{u}}$
2. Construct the "meat" matrix $\widehat{\mathbf{\Sigma}}$ with squared residuals in diagonal:

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

3. Plug $\widehat{\mathbf{\Sigma}}$ into the sandwich formula to obtain the robust estimator of the variance-covariance matrix

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

1. Fit the regression and obtain the residuals $\hat{\mathbf{u}}$
2. Construct the "meat" matrix $\widehat{\mathbf{\Sigma}}$ with squared residuals in diagonal:

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

3. Plug $\widehat{\mathbf{\Sigma}}$ into the sandwich formula to obtain the robust estimator of the variance-covariance matrix

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

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

1. Fit the regression and obtain the residuals $\hat{\mathbf{u}}$
2. Construct the "meat" matrix $\widehat{\mathbf{\Sigma}}$ with squared residuals in diagonal:

$$\widehat{\mathbf{\Sigma}} = \begin{bmatrix} \hat{u}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{u}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{u}_n^2 \end{bmatrix}$$

3. Plug $\widehat{\mathbf{\Sigma}}$ into the sandwich formula to obtain the robust estimator of the variance-covariance matrix

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

- There are various small sample corrections to improve performance when sample size is small. The most common variant (sometimes labeled HC1) is:

$$V[\hat{\beta}|\mathbf{X}] = \frac{n}{n-k-1}\cdot\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{\Sigma}}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Notes on White's Robust Standard Errors

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\boldsymbol{\beta}}$.

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\boldsymbol{\beta}}$.
- Provides a plug-and-play estimate of $V[\widehat{\beta}]$ which can be used with SEs, confidence intervals etc.—does not provide $V[u]$.

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\boldsymbol{\beta}}$.
- Provides a plug-and-play estimate of $V[\widehat{\beta}]$ which can be used with SEs, confidence intervals etc.—does not provide $V[u]$.
- Consistent for $V[\widehat{\beta}]$ under any form of heteroskedasticity (i.e. where the covariances are 0).

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\boldsymbol{\beta}}$.
- Provides a plug-and-play estimate of $V[\widehat{\beta}]$ which can be used with SEs, confidence intervals etc.—does not provide $V[u]$.
- Consistent for $V[\widehat{\beta}]$ under any form of heteroskedasticity (i.e. where the covariances are 0).
- This is a large sample result, best with large n

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\boldsymbol{\beta}}$.
- Provides a plug-and-play estimate of $V[\widehat{\beta}]$ which can be used with SEs, confidence intervals etc.—does not provide $V[u]$.
- Consistent for $V[\widehat{\beta}]$ under any form of heteroskedasticity (i.e. where the covariances are 0).
- This is a large sample result, best with large n
- For small n, performance might be poor and the estimates are downward biased (correction factors exist but are often insufficient)

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\beta}$.
- Provides a plug-and-play estimate of $V[\widehat{\beta}]$ which can be used with SEs, confidence intervals etc.—does not provide $V[u]$.
- Consistent for $V[\widehat{\beta}]$ under any form of heteroskedasticity (i.e. where the covariances are 0).
- This is a large sample result, best with large n
- For small n, performance might be poor and the estimates are downward biased (correction factors exist but are often insufficient)
- As we saw, we can arrive at White's heteroskedasticity consistent standard errors using the plug-in principle and thus in some ways, these are the natural way of getting standard errors in the agnostic regression framework.

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\beta}$.
- Provides a plug-and-play estimate of $V[\widehat{\beta}]$ which can be used with SEs, confidence intervals etc.—does not provide $V[u]$.
- Consistent for $V[\widehat{\beta}]$ under any form of heteroskedasticity (i.e. where the covariances are 0).
- This is a large sample result, best with large n
- For small n, performance might be poor and the estimates are downward biased (correction factors exist but are often insufficient)
- As we saw, we can arrive at White's heteroskedasticity consistent standard errors using the plug-in principle and thus in some ways, these are the natural way of getting standard errors in the agnostic regression framework.
- Robust SEs converge to same point as the bootstrap.

Notes on White's Robust Standard Errors

- Doesn't change estimate $\widehat{\beta}$.
- Provides a plug-and-play estimate of $V[\widehat{\beta}]$ which can be used with SEs, confidence intervals etc.—does not provide $V[u]$.
- Consistent for $V[\widehat{\beta}]$ under any form of heteroskedasticity (i.e. where the covariances are 0).
- This is a large sample result, best with large n
- For small n, performance might be poor and the estimates are downward biased (correction factors exist but are often insufficient)
- As we saw, we can arrive at White's heteroskedasticity consistent standard errors using the plug-in principle and thus in some ways, these are the natural way of getting standard errors in the agnostic regression framework.
- Robust SEs converge to same point as the bootstrap.
- This is a general framework (more to come in Week 8).

We Covered

We Covered

- Agnostic approach to deriving the estimator (see more in the Aronow and Miller textbook if you are interested).

We Covered

- Agnostic approach to deriving the estimator (see more in the Aronow and Miller textbook if you are interested).
- Robust standard errors and how they flow naturally from the plugin principle.

We Covered

- Agnostic approach to deriving the estimator (see more in the Aronow and Miller textbook if you are interested).
- Robust standard errors and how they flow naturally from the plugin principle.

Next Time: Hypothesis Tests

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

- Last Week
  - regression with two variables
  - omitted variables, multicollinearity, interactions
- This Week
  - matrix form of linear regression
  - inference and hypothesis tests
- Next Week
  - diagnostics
- Long Run
  - probability $\rightarrow$ inference $\rightarrow$ regression $\rightarrow$ causal inference

1. Matrix Form of Regression
   - Estimation
   - Fun With(out) Weights
2. OLS Classical Inference in Matrix Form
   - Unbiasedness
   - Classical Standard Errors
3. Agnostic Inference
4. Standard Hypothesis Tests
   - *t*-Tests
   - Adjusted R^2
   - F Tests for Joint Significance

1. Matrix Form of Regression
   - Estimation
   - Fun With(out) Weights
2. OLS Classical Inference in Matrix Form
   - Unbiasedness
   - Classical Standard Errors
3. Agnostic Inference
4. Standard Hypothesis Tests
   - *t*-Tests
   - Adjusted R^2
   - *F* Tests for Joint Significance

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

Model the intended Pinochet vote as a linear function of gender, education, and age of respondents.

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.

Getting the Standard Errors

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:

Getting the Standard Errors

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
```

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 distribution of the test statistic under the null hypothesis.

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 distribution of the test statistic under the null hypothesis.
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.

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 $t \approx -0.40$. 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
```

$$t = \frac{\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}}{\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia})} \quad \text{where}$$

$$\hat{SE}(\widehat{\beta}_{LAm} - \widehat{\beta}_{Asia}) = \sqrt{\widehat{V}(\widehat{\beta}_{LAm}) + \widehat{V}(\widehat{\beta}_{Asia}) - 2\widehat{Cov}[\widehat{\beta}_{LAm}, \widehat{\beta}_{Asia}]}$$

Plugging in we get $t \approx -0.40$. So what do we conclude?

We cannot reject the null that the difference in average GDP resulted from chance.

Aside: Adjusted R^2

Aside: Adjusted R^2

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
```

Aside: Adjusted R^2

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit.

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit. Recall

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

where SS_{res} are the sum of squared residuals and the SS_{tot} are the sum of the squared deviations from the mean.

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit. Recall

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

where SS_{res} are the sum of squared residuals and the SS_{tot} are the sum of the squared deviations from the mean.

- Perhaps problematically, it can be shown that R^2 always stays constant or increases with more explanatory variables

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit. Recall

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

  where SS_{res} are the sum of squared residuals and the SS_{tot} are the sum of the squared deviations from the mean.

- Perhaps problematically, it can be shown that R^2 always stays constant or increases with more explanatory variables
- So, how do we penalize more complex models?

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit. Recall

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

  where SS_{res} are the sum of squared residuals and the SS_{tot} are the sum of the squared deviations from the mean.

- Perhaps problematically, it can be shown that R^2 always stays constant or increases with more explanatory variables
- So, how do we penalize more complex models? Adjusted R^2

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit. Recall

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

  where SS_{res} are the sum of squared residuals and the SS_{tot} are the sum of the squared deviations from the mean.

- Perhaps problematically, it can be shown that R^2 always stays constant or increases with more explanatory variables
- So, how do we penalize more complex models? Adjusted R^2
- This makes R^2 more 'comparable' across models with different numbers of variables, but the next section will show you an even better way to approach that problem in a testing framework.

Aside: Adjusted R^2

- R^2 often used to assess in-sample model fit. Recall

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}}{\text{SS}_{\text{tot}}}$$

  where SS_{res} are the sum of squared residuals and the SS_{tot} are the sum of the squared deviations from the mean.

- Perhaps problematically, it can be shown that R^2 always stays constant or increases with more explanatory variables
- So, how do we penalize more complex models? Adjusted R^2
- This makes R^2 more 'comparable' across models with different numbers of variables, but the next section will show you an even better way to approach that problem in a testing framework.
- Still since people report it, the next slide derives adjusted R^2 (but we are going to skip it),

Aside: Adjusted R^2

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}/n}{\text{SS}_{\text{tot}}/n}$$

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}/n}{\text{SS}_{\text{tot}}/n}$$
$$= 1 - \frac{\tilde{V}(\text{SS}_{\text{res}})}{\tilde{V}(\text{SS}_{\text{tot}})}$$

where $\tilde{V}$ is a biased estimator of the population variance.

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

$$
\begin{aligned}
R^2 &= 1 - \frac{\text{SS}_{\text{res}}/n}{\text{SS}_{\text{tot}}/n} \\
&= 1 - \frac{\tilde{V}(\text{SS}_{\text{res}})}{\tilde{V}(\text{SS}_{\text{tot}})}
\end{aligned}
$$

  where $\tilde{V}$ is a biased estimator of the population variance.

- What if we replace the biased estimator with the unbiased estimators

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}/n}{\text{SS}_{\text{tot}}/n} = 1 - \frac{\tilde{V}(\text{SS}_{\text{res}})}{\tilde{V}(\text{SS}_{\text{tot}})}$$

  where $\tilde{V}$ is a biased estimator of the population variance.

- What if we replace the biased estimator with the unbiased estimators

$$\hat{V}(\text{SS}_{\text{res}}) = \text{SS}_{\text{res}}/(n-k-1)$$
$$\hat{V}(\text{SS}_{\text{tot}}) = \text{SS}_{\text{tot}}/(n-1)$$

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

$$R^2 = 1 - \frac{\text{SS}_{\text{res}}/n}{\text{SS}_{\text{tot}}/n} = 1 - \frac{\tilde{V}(\text{SS}_{\text{res}})}{\tilde{V}(\text{SS}_{\text{tot}})}$$

where $\tilde{V}$ is a biased estimator of the population variance.

- What if we replace the biased estimator with the unbiased estimators

$$\hat{V}(\text{SS}_{\text{res}}) = \text{SS}_{\text{res}}/(n-k-1)$$
$$\hat{V}(\text{SS}_{\text{tot}}) = \text{SS}_{\text{tot}}/(n-1)$$

- Some algebra gets us to

$$R^2_{adj} = R^2 - \underbrace{(1-R^2)\frac{k-1}{n-k}}_{\text{model complexity penalty}}$$

Aside: Adjusted R^2

- Key idea: rewrite R^2 in terms of variances

$$R^2 = 1 - \frac{SS_{res}/n}{SS_{tot}/n}$$
$$= 1 - \frac{\tilde{V}(SS_{res})}{\tilde{V}(SS_{tot})}$$

where $\tilde{V}$ is a biased estimator of the population variance.

- What if we replace the biased estimator with the unbiased estimators

$$\hat{V}(SS_{res}) = SS_{res}/(n-k-1)$$
$$\hat{V}(SS_{tot}) = SS_{tot}/(n-1)$$

- Some algebra gets us to

$$R^2_{adj} = R^2 - \underbrace{(1-R^2)\frac{k-1}{n-k}}_{\text{model complexity penalty}}$$

- Adjusted R^2 will always be smaller than R^2 and can sometimes be negative!

Why Blow Through R^2

Why Blow Through R^2

- In-sample model fit is not a particularly good indicator of model fit on a new sample.

Why Blow Through R^2

- In-sample model fit is not a particularly good indicator of model fit on a new sample.
- Adjusted R^2 is solving a problem about increasingly complex models, but by the time you reach this problem, you should be using held-out data.

Why Blow Through R^2

- In-sample model fit is not a particularly good indicator of model fit on a new sample.
- Adjusted R^2 is solving a problem about increasingly complex models, but by the time you reach this problem, you should be using held-out data.
- Stay tuned for more in Week 8!

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

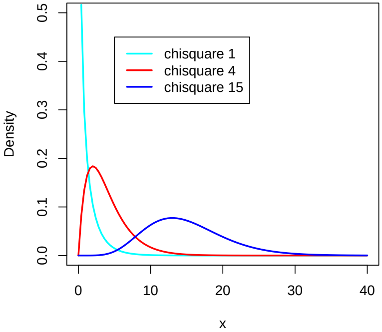
- To test more than one hypothesis jointly we need to introduce some new probability distributions.
- Suppose $Z_1, ..., Z_n$ are n i.i.d. random variables following $\mathcal{N}(0, 1)$.

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, $X \sim \chi_n^2$.



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.

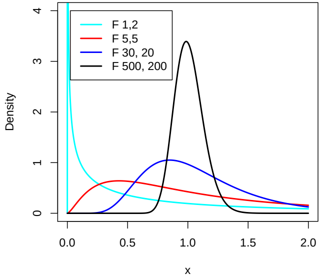
The F distribution

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

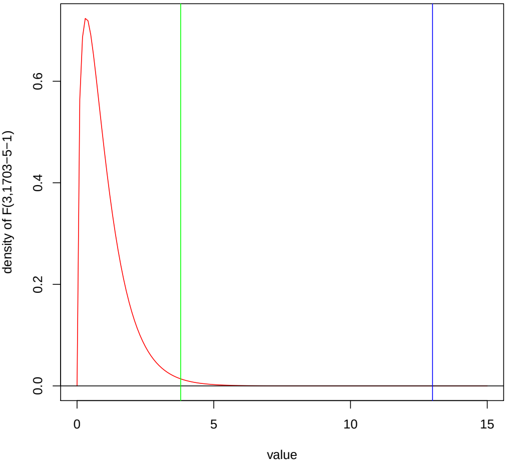
```
> SSR.UR <- sum(resid(fit.UR)^2)  # = 402
> SSR.R <- sum(resid(fit.R)^2)    # = 411

> 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:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0$$

- Have any of you used an F-test like this in your research?
- 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
```

Omnibus Test in R with Random Noise

R Code

```
> set.seed(08540)
> p <- 10; x <- matrix(rnorm(p*1000), nrow=1000)
> y <- rnorm(1000); summary(lm(y~x))
~~~~~

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.0115475  0.0320874  -0.360   0.7190
x1          -0.0019803  0.0333524  -0.059   0.9527
x2           0.0666275  0.0314087   2.121   0.0341 *
x3          -0.0008594  0.0321270  -0.027   0.9787
x4           0.0051185  0.0333678   0.153   0.8781
x5           0.0136656  0.0322592   0.424   0.6719
x6           0.0102115  0.0332045   0.308   0.7585
x7          -0.0103903  0.0307639  -0.338   0.7356
x8          -0.0401722  0.0318317  -1.262   0.2072
x9           0.0553019  0.0315548   1.753   0.0800 .
x10          0.0410906  0.0319742   1.285   0.1991
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.011 on 989 degrees of freedom
Multiple R-squared:  0.01129,       Adjusted R-squared:  0.001294
F-statistic: 1.129 on 10 and 989 DF,  p-value: 0.3364
```

F Test Examples II

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 coefficients 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 coefficients 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 coefficients 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 coefficients 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?
  → Are the coefficients X_1, X_2 and X_3 different from each other?
- How many restrictions?
  → Two ($\beta_1 - \beta_2 = 0$ and $\beta_2 - \beta_3 = 0$)
- How do we fit the restricted model?
  → 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 $F = (SSR_R - SSR_{UR}) \big/ (SSR_R/(n-k-1))$ 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

$$X \sim t_{n-k-1} \quad \Longleftrightarrow \quad X^2 \sim \mathcal{F}_{1,n-k-1}$$

- 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

$$X \ \sim \ t_{n-k-1} \quad \Longleftrightarrow \quad X^2 \ \sim \ \mathcal{F}_{1,n-k-1}$$

- 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

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \qquad + u$$

Some More Notes on F Tests

Joint significance does not necessarily imply the significance of individual coefficients, or vice versa:

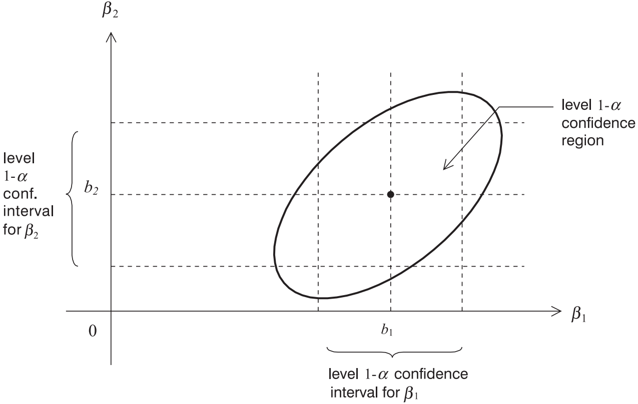


Figure 1.5: t- versus F-Tests

Image Credit: Hayashi (2011) *Econometrics*

Goal Check: Understand `lm()` Output

```
Call:
lm(formula = sr ~ pop15, data = LifeCycleSavings)

Residuals:
   Min     1Q Median     3Q    Max
-8.637 -2.374  0.349  2.022 11.155

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 17.49660    2.27972   7.675 6.85e-10 ***
pop15       -0.22302    0.06291  -3.545 0.000887 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.03 on 48 degrees of freedom
Multiple R-squared:  0.2075,Adjusted R-squared:  0.191
F-statistic: 12.57 on 1 and 48 DF,  p-value: 0.0008866
```

This Week in Review

You now have seen the full linear regression model!

This Week in Review

You now have seen the full linear regression model!

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

This Week in Review

You now have seen the full linear regression model!

- 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.

This Week in Review

You now have seen the full linear regression model!

- 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.
- You can also calculate robust standard errors which provide a plug and play replacement.

This Week in Review

You now have seen the full linear regression model!

- 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.
- You can also calculate robust standard errors which provide a plug and play replacement.
- Much of the hypothesis test infrastructure ports over nicely, plus there are new joint tests we can use.

This Week in Review

You now have seen the full linear regression model!

- 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.
- You can also calculate robust standard errors which provide a plug and play replacement.
- Much of the hypothesis test infrastructure ports over nicely, plus there are new joint tests we can use.

Next week: Troubleshooting the Linear Model!