

Week 7: Multiple Regression

Brandon Stewart¹

Princeton

October 24–26, 2022

¹Many of these slides adapted from Matt Blackwell, Adam Glynn, Justin Grimmer, Jens Hainmueller and Erin Hartman.

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 Notation of Regression
- 2 OLS Properties Under Gauss-Markov Assumptions
 - Unbiasedness
 - Classical Standard Errors
- 3 Asymptotics of OLS with Only IID Sampling
- 4 Standard Hypothesis Tests
 - t -Tests
 - Adjusted R^2
 - F Tests for Joint Significance

The Linear Model with New Notation

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

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

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

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

$$y_2 = \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad (\text{unit 2})$$

$$y_3 = \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad (\text{unit 3})$$

$$y_4 = \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad (\text{unit 4})$$

The Linear Model with New Notation

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

$$y_2 = \beta_0 + x_2\beta_1 + z_2\beta_2 + u_2 \quad (\text{unit 2})$$

$$y_3 = \beta_0 + x_3\beta_1 + z_3\beta_2 + u_3 \quad (\text{unit 3})$$

$$y_4 = \beta_0 + x_4\beta_1 + z_4\beta_2 + u_4 \quad (\text{unit 4})$$

- 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 $\mathbf{x}$, $\mathbf{z}$, and $\mathbf{u}$ vectors

Grouping Things into Matrices

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

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 a linear regression as:

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

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

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

Multiple Linear Regression in Matrix Form

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

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \quad \hat{\mathbf{y}} = \mathbf{X}\hat{\beta}$$

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

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

Residuals

- We can easily write the **residuals** in matrix form:

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

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

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

OLS Estimator in Matrix Form

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

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

- Rearranging:

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

- In order to isolate $\hat{\boldsymbol{\beta}}$, we need to move the $\mathbf{X}'\mathbf{X}$ term to the other side of the equals sign.
- Here we need the equivalent of a matrix 'division' operator.

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:

$$\mathbf{A}\mathbf{u} = \mathbf{b}$$

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

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

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

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

Back to OLS

- Recall:

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

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

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

- 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

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

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:

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

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

Disclaimer: the final equation is exactly true for all non-intercept coefficients if you remove the intercept from $\mathbf{X}$ such that $\hat{\beta}_{-0} = \text{Var}(\mathbf{X}_{-0})^{-1}\text{Cov}(\mathbf{X}_{-0}, \mathbf{y})$. 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.

Classical 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: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

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 essentially any 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

- 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 $\hat{\beta}$

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

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

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

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

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

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

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

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

$$E[\hat{\beta}|\mathbf{X}] = \beta \text{ (zero conditional mean)}$$

$$E[E[\hat{\beta}|\mathbf{X}]] = \beta \text{ (law of iterated expectations)}$$

So, yes!

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?
- $\mathbf{I}_n$ is the $n \times n$ identity matrix, σ_u^2 is a scalar.
- Visually:

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

- 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 .

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 $\hat{\beta}$

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

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

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

To estimate $V[\hat{\beta}|\mathbf{X}]$, 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:

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

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

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:

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

- In small samples, under assumptions 1-6,

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

- Estimated standard errors are:

$$\begin{aligned}\widehat{SE}[\hat{\beta}_j] &= \sqrt{\widehat{\text{var}}[\hat{\beta}]_{jj}} \\ \widehat{\text{var}}[\hat{\beta}] &= \hat{\sigma}_u^2 (\mathbf{X}'\mathbf{X})^{-1} \\ \hat{\sigma}_u^2 &= \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n - (k + 1)}\end{aligned}$$

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

Properties of the OLS Estimator: Summary

Theorem

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

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

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

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

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

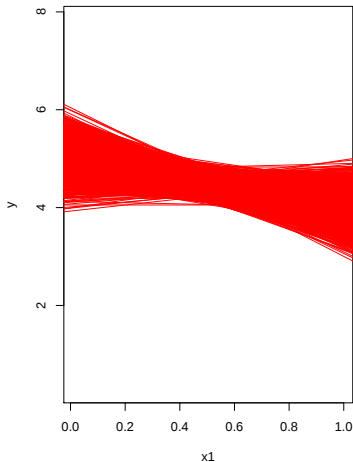
Implications of the Variance-Covariance Matrix

- 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 $\hat{\beta}$ are correlated.
- In a practical sense, this means that our uncertainty about coefficients is **correlated** across variables.

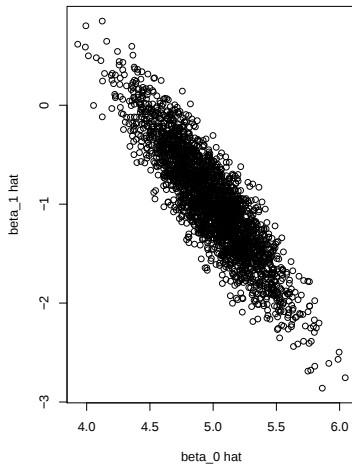
Multivariate Normal: Simulation

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

Sampling distribution of Regression Lines

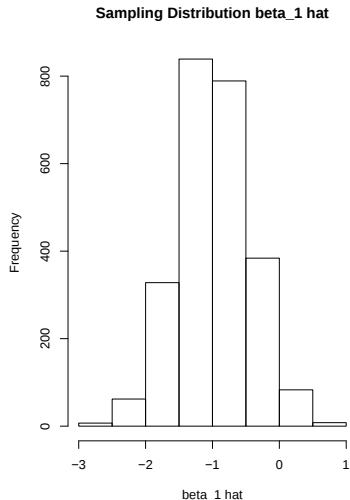
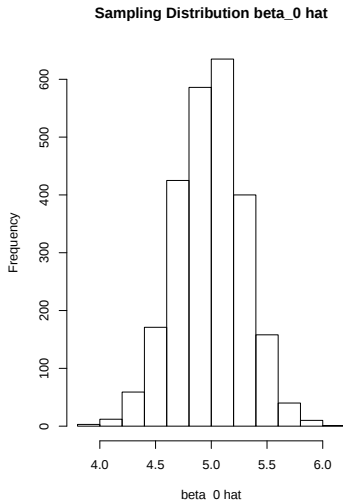


Joint sampling distribution



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 $n = 100$:



- 1 Matrix Notation of Regression
- 2 OLS Properties Under Gauss-Markov Assumptions
 - Unbiasedness
 - Classical Standard Errors
- 3 Asymptotics of OLS with Only IID Sampling
- 4 Standard Hypothesis Tests
 - t -Tests
 - Adjusted R^2
 - F Tests for Joint Significance

Asymptotics of the OLS Estimator with IID Sampling

- **Finite-sample** guarantees came at the price of **strong assumption** (CEF is linear for unbiasedness, homoskedastic-normal errors for standard errors)
- **Asymptotic results** will hold more generally under essentially only **IID sampling**, but require more data.
- Define the **Best Linear Projection** (BLP) model as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

$$E[\mathbf{X}'\mathbf{u}] = 0$$

$$E[\mathbf{X}'\mathbf{X}] > 0 \text{ (invertability)}$$

NB: ideally we would rewrite these all in terms of scalar random variable Y and random vectors for X . I've opted to keep them in matrices so the notation is parallel (you can imagine the vectors for Y are length 1 and matrices for X have one row), but the key idea remains that the expectations are over random draws.

Estimator

- We know the population value of β is:

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

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

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

Asymptotic OLS inference

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

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

- Core idea: $\mathbf{X}'\mathbf{u}$ is the sum of random variables 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.
- Note that we have only used IID sampling and no perfect collinearity so far.

Theorem (Asymptotic Normality of OLS Under Random Sampling)

Under the Best Linear Projection model,

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

where

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

and

- $\hat{\beta}$ is approximately normal asymptotically with mean β and variance Ω/n .
- Ω/n is the **asymptotic covariance matrix** of $\hat{\beta}$ and the square root of the diagonal are the standard errors.
- Asymptotically this provides us with all the necessary infrastructure for confidence intervals and tests.

Comparing to the Classical Approach (Homoskedasticity)

- Remember what we did before:

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

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

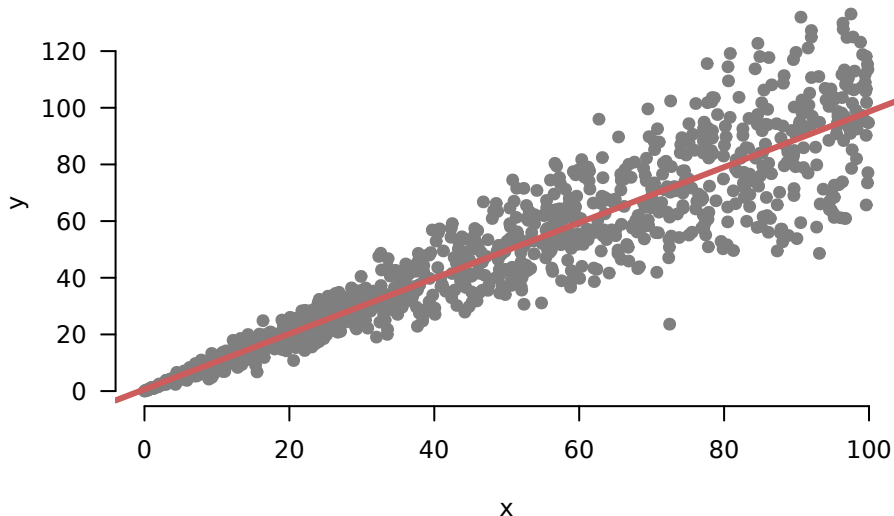
$$\text{Var}[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{\Sigma}\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}$$

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

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

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

What Does This Rule Out?



Non-constant Error Variance

- Homoskedastic:

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

- Heteroskedastic:

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

- 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 Model

- Standard error estimates incorrect:

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

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

Heteroskedasticity Consistent Estimator

- Under non-constant error variance:

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

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

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

- 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):

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

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

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

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 $\hat{\Sigma}$ is an $n \times n$ matrix, $\mathbf{X}'\hat{\Sigma}\mathbf{X}$ 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*.

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

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

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the “sandwich” formula:

- 1 Fit the regression and obtain the residuals $\hat{u}$
- 2 Construct the “meat” matrix $\hat{\Sigma}$ with squared residuals in diagonal:

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

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

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

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

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

Notes on White's Robust Standard Errors

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

- 1 Matrix Notation of Regression
- 2 OLS Properties Under Gauss-Markov Assumptions
 - Unbiasedness
 - Classical Standard Errors
- 3 Asymptotics of OLS with Only IID Sampling
- 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.
- 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 :

$$H_0 : \beta_j = c$$

- 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 $\hat{SE}(\hat{\beta}_j)$?

$$\hat{SE}(\hat{\beta}_j) = \sqrt{\widehat{V}(\hat{\beta}_j)} = \sqrt{\widehat{V}(\hat{\beta})_{(j,j)}} = \sqrt{\hat{\sigma}^2(\mathbf{X}'\mathbf{X})_{(j,j)}^{-1}}$$

where $\mathbf{A}_{(j,j)}$ is the (j,j) element of matrix $\mathbf{A}$.

That is, take the variance-covariance matrix of $\hat{\beta}$ 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 classical variance-covariance matrix 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

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

$$P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = 1 - \alpha$$

- 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}

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:

$$P\left(-t_{\alpha/2} \leq \frac{\hat{\beta}_j - \beta_j}{\hat{SE}[\hat{\beta}_j]} \leq t_{\alpha/2}\right) = 1 - \alpha$$

We rearrange:

$$\left[\hat{\beta}_j - t_{\alpha/2} \hat{SE}[\hat{\beta}_j], \hat{\beta}_j + t_{\alpha/2} \hat{SE}[\hat{\beta}_j]\right]$$

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

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:

$$T = \frac{\hat{\beta}_{LAm} - \hat{\beta}_{Asia}}{\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})}$$

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:

$$T = \frac{\hat{\beta}_{LAm} - \hat{\beta}_{Asia}}{\hat{SE}(\hat{\beta}_{LAm} - \hat{\beta}_{Asia})} \sim t_{n-k-1}$$

- 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:

$$\hat{SE}(\hat{\beta}_1 \pm \hat{\beta}_2) = \sqrt{\hat{V}(\hat{\beta}_1) + \hat{V}(\hat{\beta}_2) \pm 2\widehat{\text{Cov}}[\hat{\beta}_1, \hat{\beta}_2]}$$

which we can calculate from the estimated covariance matrix of $\hat{\beta}$.

- Since the estimates of the coefficients are correlated, we need the covariance term.

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:

```
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
```

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

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

We cannot reject the null.

NB: this same strategy can be used to get standard errors for interactions at any value of the other variables.

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

Adjusted R^2

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

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

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 .

Adjusted R^2

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

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

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

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

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

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

- Some algebra gets us to

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

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

Limits of 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 on that next week.
- Regardless the intuition here is useful: as we add more variables R^2 must stay constant or go up mechanically and that creates complications.

- 1 Matrix Notation of Regression
- 2 OLS Properties Under Gauss-Markov Assumptions
 - Unbiasedness
 - Classical Standard Errors
- 3 Asymptotics of OLS with Only IID Sampling
- 4 Standard Hypothesis Tests
 - t -Tests
 - Adjusted R^2
 - F Tests for Joint Significance

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:

$$\text{Voted} = \beta_0 + \gamma_1 \text{FEMALE} + \beta_1 \text{EDUCATION} + \gamma_2 (\text{FEMALE} \cdot \text{EDUCATION}) + \beta_2 \text{AGE} + \gamma_3 (\text{FEMALE} \cdot \text{AGE}) + u$$

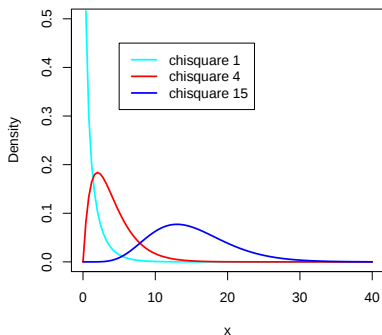
and we want to test

$$H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0.$$

- Substantively, what question are we asking?
→ 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 test **joint hypothesis**

The χ^2 Distribution

- To test more than one hypothesis jointly we need to introduce some new probability distributions.
- Suppose $Z_1, \dots, Z_n$ 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()`

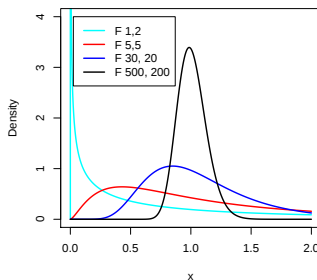
The F distribution

The **F distribution** 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 $X_1 \sim \chi_{df_1}^2$, $X_2 \sim \chi_{df_2}^2$, 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 $H_0 : \gamma_1 = \gamma_2 = \gamma_3 = 0$.

The **F statistic** can be calculated by the following procedure:

- 1 Fit the **Unrestricted Model (UR)** which *does not* impose H_0 :

$$\text{Vote} = \beta_0 + \gamma_1 \text{FEM} + \beta_1 \text{EDUC} + \gamma_2 (\text{FEM} * \text{EDUC}) + \beta_2 \text{AGE} + \gamma_3 (\text{FEM} * \text{AGE}) + u$$

- 2 Fit the **Restricted Model (R)** which *does* impose H_0 :

$$\text{Vote} = \beta_0 + \beta_1 \text{EDUC} + \beta_2 \text{AGE} + u$$

- 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.
- Asymptotically under IID sampling, $F_0 \stackrel{d}{\sim} \mathcal{F}_{q, n-k-1}$

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

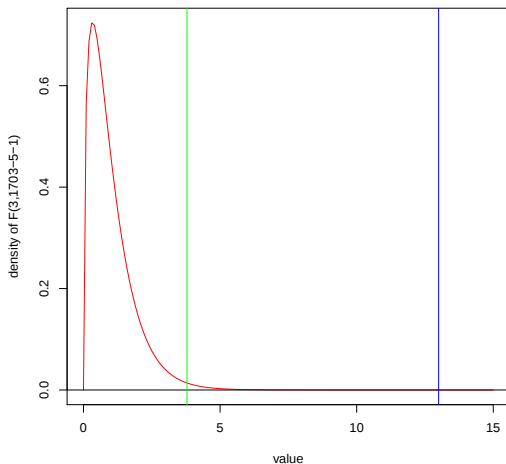
> qf(0.99, DFnum, DFdenom)
[1] 3.793171
```

Given above, what do we conclude?

$F_0 = 13$ 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

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + u$$

We may want to test:

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

- 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

The F test can be used to test various joint hypotheses which involve multiple linear restrictions. Consider the regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_k X_k + u$$

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:

$$Y = \beta_0 + \beta_1(X_1 + X_2 + X_3) + \dots + \beta_k X_k + u$$

So we create a new variable $X^* = X_1 + X_2 + X_3$ and fit:

$$Y = \beta_0 + \beta_1 X^* + \dots + \beta_k X_k + u$$

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?

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?

- Option 1: Compute $T = (\hat{\beta}_1 - \hat{\beta}_2) / \hat{SE}(\hat{\beta}_1 - \hat{\beta}_2)$ and do the **t test**.
- Option 2: Create $X^* = X_1 + X_2$, fit the restricted model, compute $F = (SSR_R - SSR_{UR}) / (SSR_R / (n - k - 1))$ and do the **F test**.

It turns out these two tests give **identical** results. This is because

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

- So, for testing a single hypothesis it does not matter whether one does a t test or an F test.
- Usually, the t test is used for single hypotheses and the F test is used for joint hypotheses.

Some More Notes on F Tests

- The F-value can also be calculated from R^2 :

$$F = \frac{(R_{UR}^2 - R_R^2)/q}{(1 - R_{UR}^2)/(n - k - 1)}$$

- 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

$$Y = \beta_0 + \beta_1 X_1 \quad + \beta_3 X_3 + u$$

against

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

Some More Notes on F Tests

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

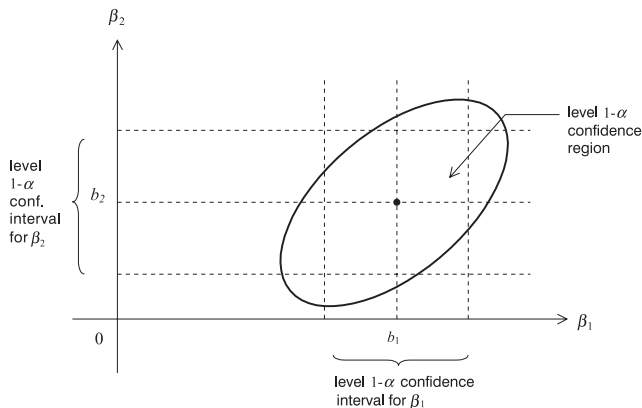


Figure 1.5: t - versus F -Tests

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!

- **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**.
- Under **strong** classical linear model assumptions we have lots of nice **finite sample** properties.
- Under the **BLP** model, we can be more agnostic and assume only IID and no-perfect collinearity. This leads to different **standard errors** and properties only hold **asymptotically** (and converge to the linear projection rather than the true CEF).
- Much of the **hypothesis test/CI** infrastructure ports over nicely, plus there are new **joint tests** we can use.

Next week: Diagnostics!