Econometrics II

Linear Regression with Multiple Regressors: Introduction

Lasha Chochua

2026

Recap: Simple Linear Regression

- In Econometrics I you studied a simple regression model:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- The goal:
  - Estimate β_1, the effect of X on Y, using OLS.
- Under the assumption $\mathbb{E}[u_i \mid X_i] = 0$, the OLS estimator $\hat{\beta}_1$ is:

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2}$$

where $s_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$ is the sample covariance and $s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ is the sample variance of X.

Does Class Size Affect Student Performance?

- To see OLS in action, consider a real policy question:

  Does reducing the number of students per teacher improve test scores?

- We use data from **420 California school districts**, and set:

$$Y_i = \text{TestScore}_i, \qquad X_i = \text{STR}_i$$

- Plugging district-level data into the OLS formula yields:

$$\hat{\beta}_1 = -2.28, \qquad \hat{\beta}_0 = 698.9$$

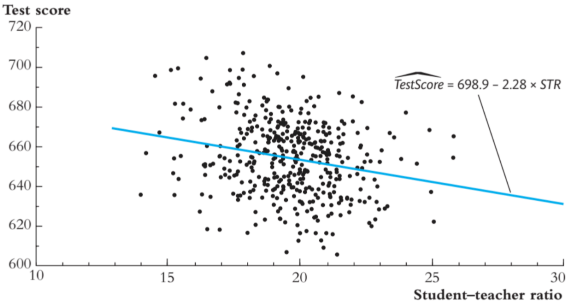
The Estimated Regression Line for the California Data

1. On the California School District Data, where TestScore is the average test score in the 420 districts and STR is the student-teacher ratio, the estimated slope is -2.28 and intercept 698.9:

$$\widehat{\text{TestScore}} = 698.9 - 2.28 \times STR$$

2. **Prediction**: For a district with 20 students per teacher:

$$\widehat{\text{TestScore}} \mid STR = 20 = 698.9 - 2.28 \times 20 = 653.3$$



Omitted Variable Bias

- But this estimate might be **biased** due to **omitted variables**:
  - Teacher quality
  - Computer access
  - **Student characteristics**, such as **English proficiency**
- These omitted variables influence test scores but were **not included** in the regression.

The Core Issue

By ignoring the percentage of English learners in the district, the OLS estimator of the effect on test scores of the student–teacher ratio could be biased.

- Why?
  - Students still learning English may **score lower** on tests.
  - Districts with **larger class sizes** tend to have **more English learners**.
  - $\Rightarrow$ The student–teacher ratio is **correlated** with an omitted determinant of test scores.

Causal vs Spurious Correlation

- The regression might capture a **spurious correlation**:

$$\text{STR} \uparrow \quad \text{and} \quad \% \quad \text{of English learners} \uparrow \quad \Rightarrow \text{TestScore} \downarrow$$

- OLS mistakenly attributes the negative effect of English proficiency to STR.
  Even if the ***true causal effect*** *of class size is small or zero,* $\hat{\beta}_1$ *may appear large and negative.*

Illustration from Data

- In the California dataset:
  - Correlation between STR and % English learners: **0.19**
  - Interpretation:
    - Districts with more English learners tend to have **larger classes**.
    - Ignoring this variable biases our estimate of β_1.

Definition of Omitted Variable Bias (OVB)

- Omitted variable bias arises when **both** of the following hold:
  1. The omitted variable is **correlated with the included regressor**, and
  2. The omitted variable is a **determinant of the dependent variable**.
- Violating one of conditions means **no OVB**.
- But if **both are true**:

$$\mathbb{E}[\hat{\beta}_1] \neq \beta_1$$

Does Omitted Variable Bias Apply?

- Let's apply the **definition of omitted variable bias** to a concrete example.
- Recall: an omitted variable biases OLS if and only if **two conditions** hold:
  1. It is **correlated with the regressor** (STR)
  2. It is a **determinant of the dependent variable** (TestScore)
- **Candidate omitted variable:** percentage of English learners ($PctEL$)

  High immigration $\rightarrow$ more English learners
- *Pause and think: does $PctEL$ satisfy both conditions?*

Let's Check the Conditions

Condition	Argument	Verdict
Correlated with STR?	Districts with more English learners tend to have larger class sizes	Yes
Determines TestScore?	Students still learning English tend to score lower on standardized tests	Yes

Let's Check the Conditions

Condition	Argument	Verdict
Correlated with STR?	Districts with more English learners tend to have larger class sizes	Yes
Determines TestScore?	Students still learning English tend to score lower on standardized tests	Yes

Conclusion

Both conditions are satisfied. Omitting $PctEL$ biases the OLS estimate of the effect of STR – the coefficient on student–teacher ratio partially picks up the negative effect of limited English proficiency.

Does Omitted Variable Bias Apply?

- Recall: an omitted variable biases OLS if and only if **two conditions** hold:
  1. It is **correlated with the regressor** (STR)
  2. It is a **determinant of the dependent variable** (TestScore)

- **Candidate omitted variable:** time of day the test was taken

- *Pause and think: does test timing satisfy both conditions?*

Let's Check the Conditions

Condition	Argument	Verdict
Correlated with STR?	Test timing varies across districts for reasons unrelated to class size	× No
Determines TestScore?	Students perform better in the morning than late in the day	Yes

Let's Check the Conditions

Condition	Argument	Verdict
Correlated with STR?	Test timing varies across districts for reasons unrelated to class size	$\times$ No
Determines TestScore?	Students perform better in the morning than late in the day	Yes

Conclusion

Only one condition is satisfied. Since test timing is uncorrelated with STR, omitting it does **not** bias the OLS estimate of the effect of class size.

Remember

Important

Not every relevant omitted variable causes bias – correlation with the regressor is key.

Does Omitted Variable Bias Apply?

- Recall: an omitted variable biases OLS if and only if **two conditions** hold:
  1. It is **correlated with the regressor** (STR)
  2. It is a **determinant of the dependent variable** (TestScore)
- **Candidate omitted variable:** parking lot space per pupil
- *Pause and think: does parking lot space satisfy both conditions?*

Let's Check the Conditions

Condition	Argument	Verdict
Correlated with STR?	More teachers per pupil $\rightarrow$ more parking space per pupil $\rightarrow$ correlated with STR	Yes
Determines TestScore?	Learning happens in the classroom, not the parking lot	$\times$ No

Let's Check the Conditions

Condition	Argument	Verdict
Correlated with STR?	More teachers per pupil $\rightarrow$ more parking space per pupil $\rightarrow$ correlated with STR	Yes
Determines TestScore?	Learning happens in the classroom, not the parking lot	$\times$ No

Conclusion

Only one condition is satisfied. Since parking lot space does not affect test scores, omitting it does **not** bias the OLS estimate of the effect of class size.

Remember

Important

Not every variable correlated with a regressor needs to be controlled for – it must also be a determinant of the outcome.

OVB and the First Least Squares Assumption

- **Key Assumption in the Simple Linear regression model:**

$$\mathbb{E}[u_i \mid X_i] = 0$$

- This is a **mean independence** assumption: the expected value of the error is zero for **any value** of X_i – ruling out any systematic relationship between u_i and X_i.

OVB and the First Least Squares Assumption (Cont.)

- Note the hierarchy of assumptions:

$$u_i \perp X_i \implies \mathbb{E}[u_i \mid X_i] = 0 \implies \text{Cov}(u_i, X_i) = 0$$

- Mean independence is **stronger** than zero correlation, but **weaker** than full independence:
  - It restricts the **mean** of u_i given X_i, but allows the **variance** of u_i to depend on X_i (i.e., heteroskedasticity is permitted).

- **Why does OVB violate this?** If a relevant variable Z_i is omitted and correlated with X_i, it enters u_i – making $\mathbb{E}[u_i \mid X_i] \neq 0$.

How OVB Violates the Assumption

- In a simple regression, the error term u_i includes **all omitted determinants** of Y_i.
- If one of these omitted variables is:
  - A **true determinant** of Y_i, and
  - **Correlated** with X_i,

  then u_i and X_i are correlated.

A Formula for Omitted Variable Bias

- Let $\rho_{Xu} = \text{corr}(X_i, u_i)$ be the correlation between the regressor and the error term. Suppose:
  - The **first least squares assumption fails**,
  - But other OLS assumptions hold.
- Then the large-sample limit of the OLS estimator is:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \left(\frac{\sigma_u}{\sigma_X} \right) \tag{1}$$

Bias Formula - I

- Start with the OLS estimator for the slope in a simple linear regression:

$$\hat{\beta}_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$$

- Substitute the model: $Y_i = \beta_0 + \beta_1 X_i + u_i$
- Then: $Y_i - \bar{Y} = \beta_1(X_i - \bar{X}) + (u_i - \bar{u})$
- So the numerator becomes:

$$\sum(X_i - \bar{X})(Y_i - \bar{Y}) = \beta_1 \sum(X_i - \bar{X})^2 + \sum(X_i - \bar{X})(u_i - \bar{u})$$

- Therefore:

$$\hat{\beta}_1 = \beta_1 + \frac{\sum(X_i - \bar{X})(u_i - \bar{u})}{\sum(X_i - \bar{X})^2}$$

Bias Formula - II

- Divide numerator and denominator by n:

$$\hat{\beta}_1 = \beta_1 + \frac{\frac{1}{n}\sum(X_i - \bar{X})(u_i - \bar{u})}{\frac{1}{n}\sum(X_i - \bar{X})^2}$$

- By the **Law of Large Numbers**, as $n \to \infty$:

$$\frac{1}{n}\sum(X_i - \bar{X})(u_i - \bar{u}) \xrightarrow{p} \text{Cov}(X, u)$$

$$\frac{1}{n}\sum(X_i - \bar{X})^2 \xrightarrow{p} \text{Var}(X) = \sigma_X^2$$

- By **Slutsky's theorem**:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\text{Cov}(X, u)}{\sigma_X^2}$$

Bias Formula - III

- Use the identity:

$$\text{Cov}(X, u) = \rho_{Xu} \cdot \sigma_X \cdot \sigma_u$$

- So the bias term becomes:

$$\frac{\text{Cov}(X, u)}{\sigma_X^2} = \rho_{Xu} \cdot \frac{\sigma_u}{\sigma_X}$$

- Final result:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu} \left(\frac{\sigma_u}{\sigma_X} \right)$$

- This shows the OLS estimator is **biased and inconsistent** when $\rho_{Xu} \neq 0$.

Key Insights from Equation (1)

1. **Bias does not vanish** with large n
   - $\hat{\beta}_1$ is not a consistent estimator if $\rho_{Xu} \neq 0$
2. **Magnitude of bias** depends on $|\rho_{Xu}|$
   - Greater correlation $\rightarrow$ **larger bias**
3. **Sign of the bias** depends on the direction of correlation:
   - Suppose:
     - English learners $\downarrow$ test scores $\rightarrow$ effect enters u_i negatively
     - English learners $\uparrow$ STR $\rightarrow$ positive correlation with X_i
   - Then: X_i (STR) is **negatively correlated** with u_i
   - So: $\rho_{Xu} < 0$
   - Therefore: $\hat{\beta}_1$ is **biased downward**

The Multiple Regression Model

- The multiple regression model **extends** the simple linear model to include **multiple regressors**.
- It allows us to estimate the **effect of one variable** (e.g., X_{1i}) on the outcome (Y_i) **while holding other variables constant** (e.g., $X_{2i}, X_{3i}, ...$).

Why Use Multiple Regression?

- In the class size example:
  - Y_i: test scores
  - X_{1i}: student–teacher ratio
  - X_{2i}: % of English learners
- Multiple regression allows us to estimate the effect of **student–teacher ratio while controlling for** English proficiency and other factors.

Two Key Uses

1. **Causal Inference**:
   - Isolate the effect of X_{1i} on Y_i by holding $X_{2i}, X_{3i}, \ldots$ constant
   - Solve omitted variable bias
2. **Prediction**:
   - Improve accuracy by including multiple predictors

The Population Regression Line

- Suppose we have two regressors: X_{1i} and X_{2i}.
- The multiple regression model defines the **conditional expectation** of Y_i as:

$$\mathbb{E}[Y_i \mid X_{1i} = x_1, X_{2i} = x_2] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{2}$$

- This is called the **population regression function**
- It gives the expected value of Y_i **given specific values** of both regressors

Interpreting the Coefficients

- β_0: intercept (expected value of Y_i when $X_{1i} = X_{2i} = 0$)
- β_1: **partial effect** of X_{1i} on Y_i, holding X_{2i} constant
- β_2: partial effect of X_{2i} on Y_i, holding X_{1i} constant

These are also called ***slope coefficients***

Meaning of β_1 as a Partial Effect

- To interpret β_1:
- Compare two observations where X_2 is the same, but X_1 differs by ΔX_1
- Then the change in Y is:

$$\Delta Y = \beta_1 \cdot \Delta X_1 \tag{3}$$

So:

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \quad \text{holding } X_2 \text{ constant}$$

Summary of Interpretation

- β_1 measures the **change in Y for a one-unit change in X_1, holding all other regressors constant**
- This differs from the simple regression interpretation, where **no other variables are held fixed**
- The term **partial effect** emphasizes the idea that we are **controlling for other variables**
- β_0 is the predicted value of Y_i when all X variables are 0
- Like in simple regression, it represents where the regression plane **intercepts the Y-axis**

The Population Multiple Regression Model

- The **population regression function** represents the average relationship between Y, X_1, and X_2.
- But in reality, Y_i is influenced by **other factors**:
  - School characteristics
  - Family background
  - Random variation ("luck")
- So, we account for these by adding an **error term**:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, \ldots, n \tag{4}$$

Generalizing to k Regressors

- In practice, we often use more than two regressors. The general model becomes:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i \quad (5)$$

- This is the **multiple linear regression model** with k regressors
- All coefficients β_j are interpreted as **partial effects** (holding other regressors constant)

Homoskedasticity vs. Heteroskedasticity

- The model is **homoskedastic** if:

$$\text{Var}(u_i \mid X_{1i}, \dots, X_{ki}) = \sigma^2$$

  – i.e., constant variance across all observations

- If this variance **depends on X values**, the model is **heteroskedastic**

Homoskedasticity simplifies inference. But in real data, heteroskedasticity is common and must be tested and/or accounted for.

OLS Estimator in Multiple Regression

- To estimate the unknown population coefficients $\beta_0, \dots, \beta_k$, we use **Ordinary Least Squares (OLS)**.
- The idea: Choose estimates $b_0, b_1, \dots, b_k$ that **minimize the sum of squared prediction errors**.

OLS Minimization Problem

- Minimize the following:

$$SSR = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{1i} - \cdots - b_k X_{ki})^2 \tag{6}$$

- This is the **sum of squared residuals**
- The values of $b_0, \ldots, b_k$ that minimize this are the **OLS estimators**:

$$\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$$

OLS Terminology

- **OLS regression line**:
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_k X_{ki}$$
- **Predicted value**: $\hat{Y}_i$ – what the model says Y_i should be
- **Residual**:
$$\hat{u}_i = Y_i - \hat{Y}_i$$
  – the difference between actual and predicted Y
- OLS could be solved by trial-and-error, but...
- Much easier using **matrix algebra and calculus**
- These formulas are built into **modern software** (e.g., R, Python, Stata)

Geometric Interpretation - I

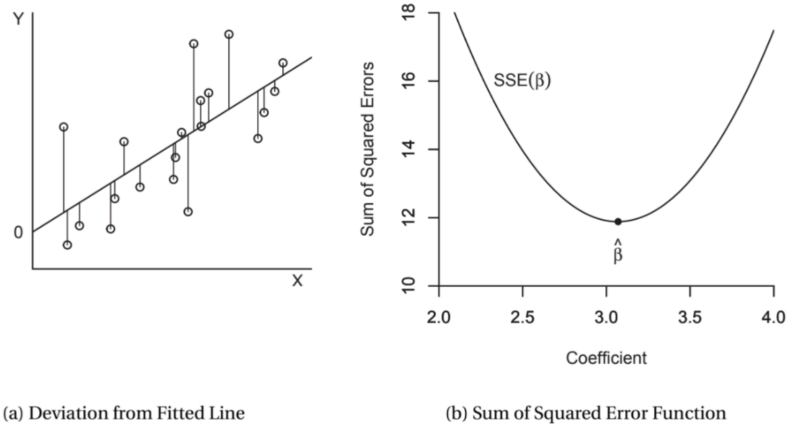


Figure 1: Simple OLS

Geometric Interpretation - II

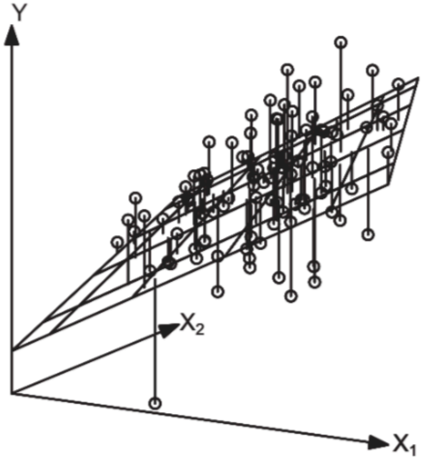


Figure 2: OLS with Two Regressands

Geometric Interpretation - II

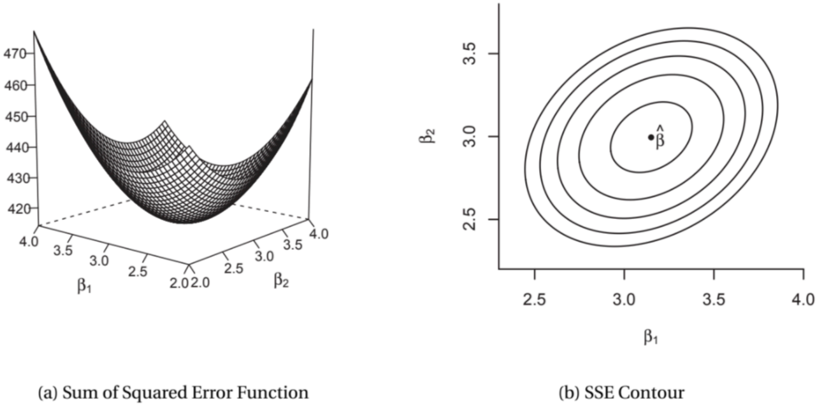


Figure 3: OLS with Two Regressands

Measures of Fit in Multiple Regression

- We use 3 common statistics to measure how well the model fits the data:
- **Standard Error of the Regression (SER)**
- R^2 (regression R-squared)
- **Adjusted** R^2
- All describe how closely the OLS line “fits” the data points.

The Standard Error of the Regression (SER)

- The SER estimates the **standard deviation of the error term** u_i.
- It measures the **typical size of a residual** — i.e., how far Y_i is from $\hat{Y}_i$.

$$SER = s_{\hat{u}} = \sqrt{s_{\hat{u}}^2} \quad \text{where} \quad s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1} \tag{7}$$

- $SSR = \sum_{i=1}^{n}\hat{u}_i^2$ is the **Sum of Squared Residuals**
- k is the number of regressors
- $n-k-1$ is the **degrees of freedom**

The R^2 in Multiple Regression

- The regression R^2 tells us: **What fraction of the variation in Y_i is explained by the regressors**
- Equivalently: $R^2 = 1-$ fraction **not** explained (i.e., from residuals)

Definition of R^2

The formula is the same as in simple regression:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{8}$$

Where:

- $ESS = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2$ = **explained sum of squares**
- $TSS = \sum_{i=1}^{n}(Y_i - \bar{Y})^2$ = **total sum of squares**
- $SSR = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$ = **sum of squared residuals**

Behavior of R^2

- Adding a new regressor **never decreases** R^2
- Why? OLS always minimizes SSR
- So unless the new variable is totally irrelevant ($\hat{\beta} = 0$), SSR drops $\rightarrow$ R^2 rises
- In practice: R^2 nearly always increases when you add a regressor.
- R^2 **always increases** with more regressors
  $\rightarrow$ this can lead to **overfitting**
- This is why we also use **adjusted** R^2 — coming up next!

The Adjusted R^2

- Unlike R^2, the **adjusted** R^2 does **not always increase** when you add a regressor.
- It adjusts R^2 to account for the number of regressors k.
- **Adjusted** R^2 **answers:** *"Does the new regressor improve the model enough to justify its inclusion?"*

Formula for Adjusted R^2

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2} \tag{9}$$

Where:

- $s_{\hat{u}}^2$ = variance of residuals
- s_Y^2 = sample variance of Y
- n = number of observations
- k = number of regressors

Interpretation

- Adjusted R^2 is **always less than** R^2
- It can **decrease** when a new variable is added
- Penalizes model complexity (i.e., adding too many predictors)

Three Key Facts about $\bar{R}^2$

1. The adjustment factor $\frac{n-1}{n-k-1} > 1$
   $\bar{R}^2 < R^2$
2. Adding a regressor:
   - Decreases SSR (increases R^2)
   - Increases the penalty factor
   - So $\bar{R}^2$ **may rise or fall**
3. Adjusted R^2 **can be negative**
   - When added variables explain **very little**
   - Rare, but it means: *"This model is worse than just using the mean"*

When to Use Adjusted R^2

- Best for **model comparison**
- Helps prevent **overfitting**
- Especially useful when you're testing whether adding variables truly helps
- **Rule of thumb:** Use R^2 to describe fit, but use $\bar{R}^2$ to compare models.

Least Squares Assumptions for Causal Inference

Key Concept 1

Model: $Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$, for $i = 1, \ldots, n$.

can be used for casual inference if

Assumptions:

1. **Zero conditional mean:** $\mathbb{E}[u_i \mid X_{1i}, X_{2i}, \ldots, X_{ki}] = 0$
2. **i.i.d. sampling:** $(X_{1i}, \ldots, X_{ki}, Y_i)$ are i.i.d. draws from the joint distribution
3. **No large outliers:** All variables have nonzero finite 4th moments
4. **No perfect multicollinearity**

hold.

Distribution of OLS Estimators in Multiple Regression

- Each sample of data gives different estimates: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$
- This sampling variation reflects **uncertainty** in estimating the population regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$

Key Result: Sampling Distribution

- Under the least squares assumptions:
- $\hat{\beta}_j$ is an **unbiased and consistent** estimator of β_j
- For large n, the sampling distribution of each $\hat{\beta}_j$ is well approximated by a **normal distribution**

From Simple to Multiple Regression

- In simple regression, $\hat{\beta}_0$, $\hat{\beta}_1$ $\rightarrow$ bivariate normal in large n
- In multiple regression: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are jointly approximately **multivariate normal**
- This follows from the Central Limit Theorem: as $n \rightarrow \infty$, the sampling distribution becomes normal.

Key Concept 2: Large-Sample Distribution of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$

If the least squares assumptions (Key Concept 1) hold, then in large samples the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are jointly normally distributed, and each $\hat{\beta}_j$ is distributed as:

$$\hat{\beta}_j \sim N(\beta_j, \sigma^2_{\hat{\beta}_j}), \quad j = 0, \ldots, k$$

OLS Estimator Distribution with $k = 2$ and Homoskedasticity

- When $k = 2$ and errors are **homoskedastic**, the variance of $\hat{\beta}_1$ simplifies:

$$\text{Var}(u_i \mid X_{1i}, X_{2i}) = \sigma_u^2$$

- Then the sampling distribution of $\hat{\beta}_1$ is approximately:

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

Large Sample Variance of $\hat{\beta}_1$

$$\sigma^2_{\hat{\beta}_1} = \frac{1}{n\left(1 - \rho^2_{X_1,X_2}\right)} \cdot \frac{\sigma^2_u}{\sigma^2_{X_1}} \tag{10}$$

Where:

- ρ_{X_1,X_2} is the **population correlation** between regressors
- $\sigma^2_{X_1}$ is the **variance** of regressor X_1

Control Variables

Least Squares Assumptions (with Control Variables):

Consider the model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i$$

where $\beta_1, \ldots, \beta_k$ are causal effects, and W's are control variables. The assumptions:

1. **Conditional Mean Independence:**

$$\mathbb{E}(u_i \mid X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}) = \mathbb{E}(u_i \mid W_{1i}, \ldots, W_{ri})$$

2. The data $(X_{1i}, \ldots, X_{ki}, W_{1i}, \ldots, W_{ri}, Y_i)$ are i.i.d.
3. Large outliers are unlikely: all regressors and Y_i have nonzero finite fourth moments.
4. No perfect multicollinearity among regressors.

Multicollinearity and Causal Inference

- One of the key assumptions for valid causal inference in multiple regression is:

No perfect multicollinearity

- **Multicollinearity** refers to the situation where one regressor is a linear combination of the others.
- This prevents the OLS estimator from being computed.
  - **Perfect multicollinearity** $\Rightarrow$ OLS breaks
  - **High (imperfect) multicollinearity** $\Rightarrow$ imprecise estimates, large SEs
- Now let's switch to the **Jupyter Notebook** and explore multicollinearity using simulated data.

Main Reading

- Stock & Watson (2020) - Chapter 6