

Statistics for Economists II

Lecture 12: Simple Linear Regression & OLS

Güneş A. Aşık

Department of Economics
Middle East Technical University

Spring 2025–2026

Roadmap for Today

- 1 Motivation: Why Regression?
- 2 The Simple Linear Regression Model
- 3 Ordinary Least Squares (OLS) Estimation
- 4 Worked Examples
- 5 Interpreting OLS Results
- 6 Practice Problems
- 7 Key Takeaways & Further Reading

From Descriptive Statistics to Relationships

- So far: estimating means, proportions, testing hypotheses about *single* variables

Think About It

You know the correlation between study hours and GPA is positive. But *by how much* does GPA rise per extra hour of study? Correlation alone cannot answer this—regression can.

From Descriptive Statistics to Relationships

- So far: estimating means, proportions, testing hypotheses about *single* variables
- But economists care about **relationships**:
 - How does income affect consumption?
 - Does education raise wages?
 - What happens to demand when price rises?

Think About It

You know the correlation between study hours and GPA is positive. But *by how much* does GPA rise per extra hour of study? Correlation alone cannot answer this—regression can.

From Descriptive Statistics to Relationships

- So far: estimating means, proportions, testing hypotheses about *single* variables
- But economists care about **relationships**:
 - How does income affect consumption?
 - Does education raise wages?
 - What happens to demand when price rises?
- We need a tool that quantifies **how one variable changes when another changes**

Think About It

You know the correlation between study hours and GPA is positive. But *by how much* does GPA rise per extra hour of study? Correlation alone cannot answer this—regression can.

From Descriptive Statistics to Relationships

- So far: estimating means, proportions, testing hypotheses about *single* variables
- But economists care about **relationships**:
 - How does income affect consumption?
 - Does education raise wages?
 - What happens to demand when price rises?
- We need a tool that quantifies **how one variable changes when another changes**
- That tool is **regression analysis**

Think About It

You know the correlation between study hours and GPA is positive. But *by how much* does GPA rise per extra hour of study? Correlation alone cannot answer this—regression can.

From Correlation to Causation (Carefully!)

Warning

Regression tells us about **statistical association**. Moving from association to **causation** requires additional assumptions and careful reasoning.

- **Correlation**: ice cream sales and drowning deaths are positively correlated
- Does ice cream *cause* drowning? Of course not—both are driven by **temperature**
- Regression is a *starting point*; causal inference needs more (we will revisit this later in the course)

Intuition: Regression helps us *describe* and *quantify* relationships. Whether that relationship is causal depends on the **research design**, not the statistical method itself.

Economic Model vs. Statistical Model

Economic (Theoretical) Model

- Based on economic theory
- Exact, deterministic relationship
- Example—Keynesian consumption function:

$$C = \beta_1 + \beta_2 Y$$

- Says: consumption is a *linear* function of income

Statistical (Econometric) Model

- Adds a **stochastic (random) term**
- Recognizes that the relationship is *inexact*
- Example:

$$C = \beta_1 + \beta_2 Y + \varepsilon$$

- ε captures everything else that affects C besides Y

Intuition: The error term ε is what makes economics an *empirical* science—we acknowledge that our models are simplifications of a complex reality.

The Keynesian Consumption Function: Our Running Example

- Keynes's **Fundamental Psychological Law**: as income rises, consumption rises, but by less than the increase in income
- The **marginal propensity to consume (MPC)**: $0 < \beta_2 < 1$

$$C_i = \beta_1 + \beta_2 Y_i + \varepsilon_i$$

Symbol	Meaning
C_i	Consumption expenditure of household i
Y_i	Income of household i
β_1	Autonomous consumption (intercept)
β_2	Marginal propensity to consume (slope)
ε_i	Random disturbance / error term

We will estimate β_1 and β_2 from data using **OLS**.

The Population Regression Model

The **simple linear regression model** (also called the two-variable model):

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n$$

- y_i : **dependent variable** (regressand, response, explained variable)
- x_i : **independent variable** (regressor, predictor, explanatory variable)
- β_0 : population **intercept** (constant)
- β_1 : population **slope** coefficient
- ε_i : **error term** (disturbance)

Intuition: This says: y is *approximately* a linear function of x , with some random noise ε added on top.

What Does the Error Term Capture?

The error term ε_i serves **three purposes**:

- ❶ **Omitted variables**: Many factors affect y besides x . We lump them into ε .
 - Consumption depends on wealth, interest rates, expectations. . . not just income
- ❷ **Measurement error**: Economic data are often imprecisely measured.
 - Surveys have rounding, non-response, misreporting
- ❸ **Inherent randomness**: Human behavior has an unpredictable component.
 - Two households with identical income may consume different amounts

Intuition: Without ε , every data point would lie *exactly* on the line. In reality, data scatter around the line— ε explains *why*.

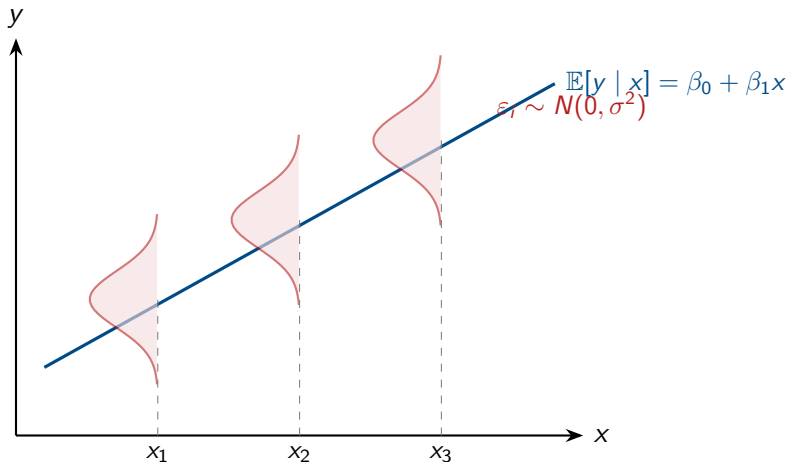
Assumptions of the Classical Linear Regression Model

The Gauss–Markov Assumptions (for the simple model)

- 1 **Linearity:** $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ (linear in *parameters*)
- 2 **Fixed regressors** (or independent of ε): x_i values are given / non-stochastic
- 3 **Zero conditional mean:** $\mathbb{E}[\varepsilon_i | x_i] = 0$ (no systematic error)
- 4 **Homoscedasticity:** $\text{Var}(\varepsilon_i) = \sigma^2$ for all i (constant variance)
- 5 **No autocorrelation:** $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$
- 6 **Normality** (optional for OLS, needed for inference): $\varepsilon_i \sim N(0, \sigma^2)$

Intuition: Under assumptions 1–5, OLS is the **Best Linear Unbiased Estimator** (BLUE)—the Gauss–Markov theorem. We will discuss what BLUE means in detail next week.

Visualizing the Assumptions



At each x -value, y is distributed around the regression line with **constant variance** σ^2 (homoscedasticity) and **zero mean** error.

The Key Question

Given data $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$,
how do we find the “best” line through the data?

The Fitted Line

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

- $\hat{\beta}_0, \hat{\beta}_1$: **estimates** of β_0, β_1
- $\hat{y}_i$: **predicted** (fitted) value

The Residual

$$e_i = y_i - \hat{y}_i$$

- Difference between **observed** y_i and **predicted** $\hat{y}_i$
- Our “mistake” for observation i

Intuition: We want the line that makes the residuals as “small” as possible overall. But how do we measure “small”?

Why Squared Residuals?

We want to minimize some measure of the overall residuals. Three candidates:

Sum of residuals?

$$\sum_{i=1}^n e_i$$

Problem: Positive and negative residuals cancel out. Could be zero even with a terrible fit!

Sum of absolute residuals?

$$\sum_{i=1}^n |e_i|$$

Works, but not differentiable everywhere—hard to optimize analytically (used in LAD estimation).

Sum of squared residuals!

$$\sum_{i=1}^n e_i^2$$

Winner: Always positive, differentiable, penalizes large errors more, leads to clean formulas.

Intuition: Squaring kills two birds with one stone: it prevents cancellation *and* it punishes outliers (big mistakes) more heavily.

The OLS Objective

Ordinary Least Squares (OLS) chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize:

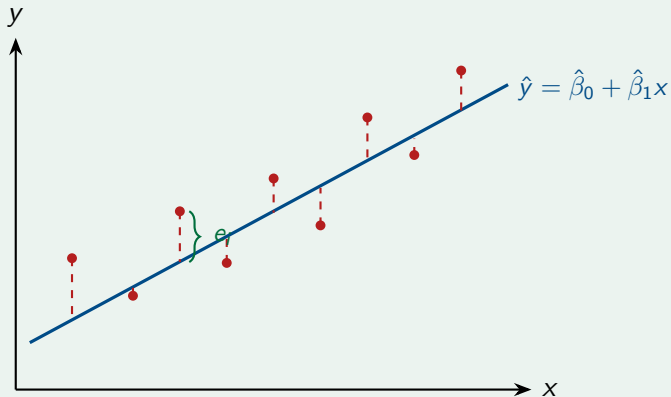
$$S(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

OLS minimizes the sum of the **squared vertical distances** from each point to the line.

The OLS Objective

Ordinary Least Squares (OLS) chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize:

Geometric Interpretation



Deriving the OLS Formulas: Step 1 — First-Order Conditions

Minimize $S = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$ by taking partial derivatives and setting them to zero:

FOC w.r.t. $\hat{\beta}_0$:

$$\begin{aligned}\frac{\partial S}{\partial \hat{\beta}_0} &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \Rightarrow \sum_{i=1}^n y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^n x_i\end{aligned}\quad (\text{Eq. 1})$$

FOC w.r.t. $\hat{\beta}_1$:

$$\begin{aligned}\frac{\partial S}{\partial \hat{\beta}_1} &= -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \Rightarrow \sum_{i=1}^n x_i y_i &= \hat{\beta}_0 \sum_{i=1}^n x_i + \hat{\beta}_1 \sum_{i=1}^n x_i^2\end{aligned}\quad (\text{Eq. 2})$$

Deriving the OLS Formulas: Step 2 — Solving for $\hat{\beta}_0$

From Equation 1, divide both sides by n :

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

Solving for $\hat{\beta}_0$:

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Intuition: This tells us something beautiful: the OLS regression line **always passes through the point** $(\bar{x}, \bar{y})$ —the “center of gravity” of the data.

Deriving the OLS Formulas: Step 3 — Solving for $\hat{\beta}_1$

Substitute $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ into Equation 2:

$$\sum x_i y_i = (\bar{y} - \hat{\beta}_1 \bar{x}) \sum x_i + \hat{\beta}_1 \sum x_i^2$$

After algebra (expanding, collecting terms with $\hat{\beta}_1$):

$$\sum x_i y_i - n \bar{x} \bar{y} = \hat{\beta}_1 \left(\sum x_i^2 - n \bar{x}^2 \right)$$

Recognizing the **deviation formulas**:

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n \bar{x} \bar{y}, \quad \sum (x_i - \bar{x})^2 = \sum x_i^2 - n \bar{x}^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

The OLS Formulas: Summary

OLS Estimators for Simple Linear Regression

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Alternative expression for $\hat{\beta}_1$ (useful for computation):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n x_i^2 - n \bar{x}^2} = \frac{\text{sample covariance of } x \text{ and } y}{\text{sample variance of } x} = \frac{s_{xy}}{s_x^2}$$

Example: Keynesian Consumption Function — Data

Consider the following data on weekly income (Y , in \$) and consumption (C , in \$) for 10 households:

i	1	2	3	4	5	6	7	8	9	10
Y_i	80	100	120	140	160	180	200	220	240	260
C_i	55	65	79	80	102	110	120	135	137	150

Step 1: Compute sample means:

$$\bar{Y} = \frac{80 + 100 + \cdots + 260}{10} = 170, \quad \bar{C} = \frac{55 + 65 + \cdots + 150}{10} = 103.3$$

Our goal: estimate $C_i = \hat{\beta}_0 + \hat{\beta}_1 Y_i$.

Example: Consumption Function — Computing $\hat{\beta}_1$

Step 2: Build the computation table (showing key columns):

Y_i	C_i	$Y_i - \bar{Y}$	$C_i - \bar{C}$	$(Y_i - \bar{Y})(C_i - \bar{C})$	$(Y_i - \bar{Y})^2$
80	55	-90	-48.3	4347	8100
100	65	-70	-38.3	2681	4900
120	79	-50	-24.3	1215	2500
140	80	-30	-23.3	699	900
160	102	-10	-1.3	13	100
180	110	10	6.7	67	100
200	120	30	16.7	501	900
220	135	50	31.7	1585	2500
240	137	70	33.7	2359	4900
260	150	90	46.7	4203	8100
				17 670	33 000

$$\hat{\beta}_1 = \frac{\sum(Y_i - \bar{Y})(C_i - \bar{C})}{\sum(Y_i - \bar{Y})^2} = \frac{17,670}{33,000} \approx \mathbf{0.5355}$$

Example: Consumption Function — Results

Step 3: Compute $\hat{\beta}_0$:

$$\hat{\beta}_0 = \bar{C} - \hat{\beta}_1 \bar{Y} = 103.3 - 0.5355 \times 170 = 103.3 - 91.03 \approx \mathbf{12.27}$$

Estimated Consumption Function

$$\hat{C}_i = 12.27 + 0.5355 Y_i$$

Interpretation:

- $\hat{\beta}_0 = 12.27$: When income is zero, predicted consumption is \$12.27 (autonomous consumption)
- $\hat{\beta}_1 = 0.5355$: For each additional dollar of income, consumption increases by approximately **54 cents**, *on average*
- The estimated MPC is 0.5355—consistent with Keynes's prediction that $0 < \text{MPC} < 1$

Classical Example: MPC = 0.7064 (Textbook Data)

From the textbook (Gujarati, Example 6.1), using US data:

$$\hat{C}_i = 24.45 + 0.7064 Y_i$$

Reading the Result

- MPC = 0.7064: out of every extra dollar of income, about **71 cents** goes to consumption
- Autonomous consumption = \$24.45
- Consistent with Keynesian theory

Caution

- Is the relationship truly *causal*?
- Are there omitted variables?
- We will address these questions when we discuss **inference** and **model diagnostics**

Example: Study Hours and Test Scores

A professor collects data on hours studied (x) and test scores (y) for 8 students:

x_i	2	3	4	5	6	7	8	9
y_i	50	55	62	68	71	78	82	88

Compute: $\bar{x} = 5.5$, $\bar{y} = 69.25$

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{196.5}{42} = 4.679$$

$$\hat{\beta}_0 = 69.25 - 4.679 \times 5.5 = 43.52$$

Result

$\hat{y}_i = 43.52 + 4.68 x_i$ — Each additional hour of study is associated with a **4.68 point** increase in test score, on average.

Interpreting $\hat{\beta}_0$ (the Intercept)

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Interpretation: $\hat{\beta}_0$ is the predicted value of y when $x = 0$.

- In the consumption example: $\hat{\beta}_0 = 12.27$ means predicted consumption when income is zero
- In the study hours example: $\hat{\beta}_0 = 43.52$ is the predicted test score with zero hours of study

Practical Warning

The intercept often lacks a meaningful economic interpretation, especially if $x = 0$ is far outside the range of the data. No one in our sample has zero income—so $\hat{\beta}_0$ is an *extrapolation*.

Intuition: The intercept is often just a “mathematical anchor” that positions the line correctly. The slope is usually more economically interesting.

Interpreting $\hat{\beta}_1$ (the Slope)

$$\hat{\beta}_1 = \frac{\Delta \hat{y}}{\Delta x}$$

Interpretation: A one-unit increase in x is associated with a $\hat{\beta}_1$ -unit change in $\hat{y}$, **on average**.

Key Phrases

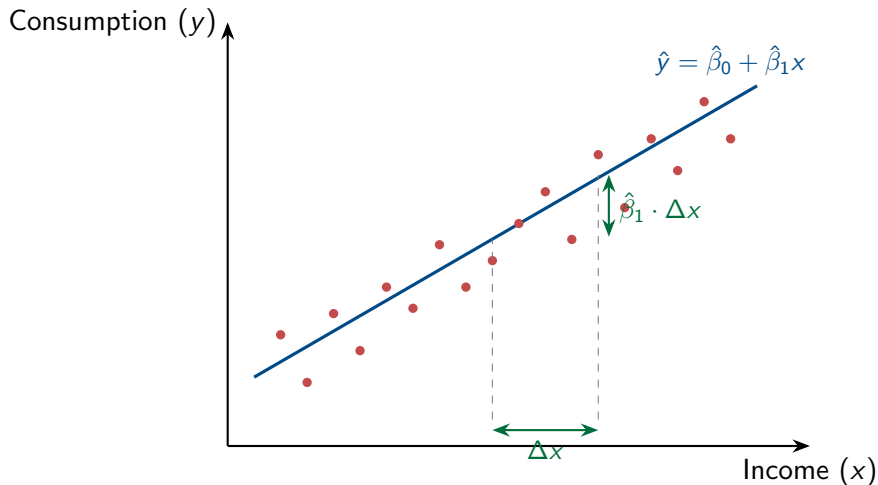
- "...is *associated with*..."
- "...*on average*..."
- "...holding other factors constant..."
(for multiple regression)

Avoid Saying

- "x *causes* y to change by $\hat{\beta}_1$ "
- "y *will definitely* change by $\hat{\beta}_1$ "
- "Every individual changes by exactly $\hat{\beta}_1$ "

Intuition: The word "*on average*" is crucial. The regression line tells us about the **average relationship**—individual observations scatter around this average due to the error term.

The “On Average” Interpretation: Visualized



The line represents the **average** relationship. Individual households (red dots) deviate from the average.

Properties of OLS Residuals

Two useful properties follow directly from the first-order conditions:

Property 1: Residuals Sum to Zero

$$\sum_{i=1}^n e_i = \sum_{i=1}^n (y_i - \hat{y}_i) = 0$$

The regression line is positioned so that positive and negative residuals exactly cancel.

Property 2: Residuals Are Uncorrelated with x

$$\sum_{i=1}^n x_i e_i = 0$$

OLS “extracts” all the linear information from x —nothing systematic is left in the residuals.

The Fitted Line Always Passes Through $(\bar{x}, \bar{y})$

Recall: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

Plug $x = \bar{x}$ into the fitted equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} = (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 \bar{x} = \bar{y} \quad \checkmark$$

What This Means

No matter what the data look like, the OLS line always goes through the “center of mass” of the scatter plot—the point $(\bar{x}, \bar{y})$.

Intuition: Think of $(\bar{x}, \bar{y})$ as the “pivot point” of the regression line. The slope $\hat{\beta}_1$ determines how the line tilts around this point.

Practice Problem 1

Problem

Given the following data on advertising expenditure (x , in thousands \$) and sales (y , in thousands \$):

x_i	1	2	3	4	5
y_i	10	14	17	21	24

Find $\hat{\beta}_0$ and $\hat{\beta}_1$. Interpret the slope.

Practice Problem 1

Problem

Given the following data on advertising expenditure (x , in thousands \$) and sales (y , in thousands \$):

x_i	1	2	3	4	5
y_i	10	14	17	21	24

Find $\hat{\beta}_0$ and $\hat{\beta}_1$. Interpret the slope.

Solution

$$\bar{x} = 3, \quad \bar{y} = 17.2$$

$$\sum (x_i - \bar{x})(y_i - \bar{y}) = (-2)(-7.2) + (-1)(-3.2) + (0)(-0.2) + (1)(3.8) + (2)(6.8) = 34.0$$

$$\sum (x_i - \bar{x})^2 = 4 + 1 + 0 + 1 + 4 = 10$$

$$\hat{\beta}_1 = 34/10 = 3.4, \quad \hat{\beta}_0 = 17.2 - 3.4(3) = 7.0$$

Interpretation: Each additional \$1,000 in advertising is associated with \$3,400 more in sales,

Practice Problem 2

Problem

For the regression $\hat{y}_i = 5 + 2x_i$ estimated from 20 observations with $\bar{x} = 10$ and $\bar{y} = 25$:

- (a) Verify that the line passes through $(\bar{x}, \bar{y})$.
- (b) Predict y when $x = 15$.
- (c) If $y_{15} = 38$ (actual value when $x = 15$), find the residual.

Practice Problem 2

Problem

For the regression $\hat{y}_i = 5 + 2x_i$ estimated from 20 observations with $\bar{x} = 10$ and $\bar{y} = 25$:

- (a) Verify that the line passes through $(\bar{x}, \bar{y})$.
- (b) Predict y when $x = 15$.
- (c) If $y_{15} = 38$ (actual value when $x = 15$), find the residual.

Solution

(a) $\hat{y} = 5 + 2(10) = 25 = \bar{y} \quad \checkmark$

(b) $\hat{y}_{15} = 5 + 2(15) = 35$

(c) $e_{15} = y_{15} - \hat{y}_{15} = 38 - 35 = 3$

The model under-predicts by 3 units for this observation.

Practice Problem 3

Problem

A researcher estimates the relationship between price (x , in TL) and quantity demanded (y , in units):

$$\hat{y} = 200 - 4.5x$$

- (a) Interpret $\hat{\beta}_0 = 200$ and $\hat{\beta}_1 = -4.5$.
- (b) Is the sign of $\hat{\beta}_1$ consistent with the law of demand?
- (c) Predict quantity demanded when price is 20 TL.

Practice Problem 3

Solution

- (a) $\hat{\beta}_0 = 200$: When price is zero, predicted demand is 200 units.
 $\hat{\beta}_1 = -4.5$: Each 1 TL price increase is associated with a decrease of 4.5 units demanded, on average.
- (b) Yes—the law of demand says price and quantity demanded are **inversely related**. The negative slope is consistent with theory.
- (c) $\hat{y} = 200 - 4.5(20) = 200 - 90 = 110$ units.

Practice Problem 4

Problem

Show that the OLS slope $\hat{\beta}_1$ can be written as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Hint: Expand $(x_i - \bar{x})(y_i - \bar{y})$ and $(x_i - \bar{x})^2$.

Practice Problem 4

Problem

Show that the OLS slope $\hat{\beta}_1$ can be written as:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}$$

Hint: Expand $(x_i - \bar{x})(y_i - \bar{y})$ and $(x_i - \bar{x})^2$.

Solution

Numerator: $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \bar{y} \sum x_i - \bar{x} \sum y_i + n\bar{x}\bar{y}$

Since $\sum x_i = n\bar{x}$ and $\sum y_i = n\bar{y}$: $= \sum x_i y_i - n\bar{x}\bar{y} - n\bar{x}\bar{y} + n\bar{x}\bar{y} = \sum x_i y_i - n\bar{x}\bar{y}$ ✓

Denominator: $\sum (x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x} \sum x_i + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2$ ✓

Key Takeaways

- 1 **Regression** quantifies how one variable changes when another changes—going beyond correlation.
- 2 The **simple linear regression model** is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The error term captures omitted variables, measurement error, and inherent randomness.
- 3 **OLS** finds the line that minimizes the sum of squared residuals:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- 4 Interpretation is about **averages**: “A one-unit increase in x is associated with a $\hat{\beta}_1$ -unit change in y , on average.”
- 5 The OLS line always passes through $(\bar{x}, \bar{y})$ and the residuals sum to zero.

What's Coming Next

Week 13: Properties of OLS Estimators

- Unbiasedness of $\hat{\beta}_0$ and $\hat{\beta}_1$
- Variance of OLS estimators
- The Gauss–Markov theorem (BLUE)
- Estimating σ^2

Week 14: Inference in Regression

- Hypothesis testing for β_1
- Confidence intervals
- Goodness of fit: R^2

Further Reading

- **Gujarati**: Chapters 1–3 (“The Nature of Regression Analysis” and “Two-Variable Regression Model”)
- **Maddala**: Chapter 2 (“Simple Regression”)
- Lecture notes: Sections on the two-variable model and OLS