



Lecture 11

Statistics II

STAT-221 • UG

Dr. Rizwan Jahangir

NUST Business School

National University of Sciences & Technology

Contents

- **1. The Pivot: From Correlation to Regression:**
 - ▶ Conceptual Transition & Bivariate Modeling.
 - ▶ The Regression Line Equation ($\hat{y} = a + bx$).
- **2. The Least Squares Method:**
 - ▶ Residuals & The Best-Fit Line.
 - ▶ Formulas for Slope (b) and Intercept (a).
- **3. Solved Case Study: Copier Sales Regression ($n = 15$):**
 - ▶ Step-by-Step Manual Calculations.
- **4. Evaluating Predictive Power:**
 - ▶ The Standard Error of Estimate ($s_{y \cdot x}$).
 - ▶ Coefficient of Determination (r^2).
- **5. Testing the Significance of the Slope (β)**
- **6. Interval Estimates of Prediction:**
 - ▶ Confidence Interval of the Mean vs. Individual Prediction Interval.
 - ▶ Geometric Representation of Predictions.



The Pivot: From Correlation to Regression

Post-Midterm Course Roadmap

We are making excellent progress in the second half of STAT-221!

- In Lecture 9, we compared means of three or more groups using ANOVA.
- In Lecture 10, we examined the linear relationship between two variables using correlation (r).



Post-Midterm Course Roadmap

We are making excellent progress in the second half of STAT-221!

- In Lecture 9, we compared means of three or more groups using ANOVA.
- In Lecture 10, we examined the linear relationship between two variables using correlation (r).
- **Today's Goal:** Instead of just asking "*Are they related?*", we build a mathematical equation to **predict** the value of one variable based on the other!



Post-Midterm Course Roadmap

We are making excellent progress in the second half of STAT-221!

- In Lecture 9, we compared means of three or more groups using ANOVA.
- In Lecture 10, we examined the linear relationship between two variables using correlation (r).
- **Today's Goal:** Instead of just asking "*Are they related?*", we build a mathematical equation to **predict** the value of one variable based on the other!

Today's Learning Objectives (CLO 1, CLO 3):

- Formulate a linear regression equation using the Least Squares Method.
- Compute and interpret the Standard Error of Estimate and R-squared.
- Conduct a hypothesis test on the slope of the regression line.
- Differentiate between, calculate, and interpret Confidence Intervals and Prediction Intervals.



Real-World Icebreaker: Predictive Power in Business

Discussion

Bivariate Prediction Scenario: Real Estate and Staffing

Think about these three business scenarios:

- A real estate agency wants to estimate the selling price of a home based on its square footage.
- A sales manager wants to estimate monthly sales volume based on the number of sales calls.
- A retail store wants to forecast daily customer foot traffic based on the local temperature.

How does moving from simple "association" (correlation) to "prediction" (regression) change how a business makes operational decisions?



Back to Basics: The Straight Line from Algebra

Before looking at statistical formulas, let's recall high school algebra. Do you remember the equation of a straight line?

$$y = mx + c$$



Back to Basics: The Straight Line from Algebra

Before looking at statistical formulas, let's recall high school algebra. Do you remember the equation of a straight line?

$$y = mx + c$$

Let's redefine these algebraic terms:

- y : The vertical coordinate (dependent variable).
- x : The horizontal coordinate (independent variable).



Back to Basics: The Straight Line from Algebra

Before looking at statistical formulas, let's recall high school algebra. Do you remember the equation of a straight line?

$$y = mx + c$$

Let's redefine these algebraic terms:

- y : The vertical coordinate (dependent variable).
- x : The horizontal coordinate (independent variable).
- m : The **slope** of the line (rise over run). It tells us how much y changes for every 1-unit increase in x .



Back to Basics: The Straight Line from Algebra

Before looking at statistical formulas, let's recall high school algebra. Do you remember the equation of a straight line?

$$y = mx + c$$

Let's redefine these algebraic terms:

- y : The vertical coordinate (dependent variable).
- x : The horizontal coordinate (independent variable).
- m : The **slope** of the line (rise over run). It tells us how much y changes for every 1-unit increase in x .
- c : The **y -intercept**. It is the point where the line crosses the vertical axis (where $x = 0$).



Translating Algebra into Statistics

In statistics, we use the exact same straight-line logic, but we write the symbols slightly differently to align with standard estimation notation:

The Linear Regression Equation:

$$\hat{y} = a + bx$$

Where:



Translating Algebra into Statistics

In statistics, we use the exact same straight-line logic, but we write the symbols slightly differently to align with standard estimation notation:

The Linear Regression Equation:

$$\hat{y} = a + bx$$

Where:

- $\hat{y}$ (read "**y-hat**"): The estimated, predicted, or expected value of Y for a chosen value of x .



Translating Algebra into Statistics

In statistics, we use the exact same straight-line logic, but we write the symbols slightly differently to align with standard estimation notation:

The Linear Regression Equation:

$$\hat{y} = a + bx$$

Where:

- $\hat{y}$ (read "**y-hat**"): The estimated, predicted, or expected value of Y for a chosen value of x .
- a : The **y-intercept**. The estimated value of Y when $x = 0$.



Translating Algebra into Statistics

In statistics, we use the exact same straight-line logic, but we write the symbols slightly differently to align with standard estimation notation:

The Linear Regression Equation:

$$\hat{y} = a + bx$$

Where:

- $\hat{y}$ (read "**y-hat**"): The estimated, predicted, or expected value of Y for a chosen value of x .
- a : The **y-intercept**. The estimated value of Y when $x = 0$.
- b : The **slope** of the line. The estimated change in $\hat{y}$ for each 1-unit change in x .



Translating Algebra into Statistics

In statistics, we use the exact same straight-line logic, but we write the symbols slightly differently to align with standard estimation notation:

The Linear Regression Equation:

$$\hat{y} = a + bx$$

Where:

- $\hat{y}$ (read "**y-hat**"): The estimated, predicted, or expected value of Y for a chosen value of x .
- a : The **y-intercept**. The estimated value of Y when $x = 0$.
- b : The **slope** of the line. The estimated change in $\hat{y}$ for each 1-unit change in x .
- x : Any chosen value of the independent predictor variable.



The Least Squares Method

The Challenge of Scatter Plots

If we plot bivariate data, the points do not fall perfectly on a single straight line. Instead, they form a scattered cloud of points.

- We could draw dozens of different lines through the cloud of data points.
- Some lines would be closer to some points, while others would fit different points better.



The Challenge of Scatter Plots

If we plot bivariate data, the points do not fall perfectly on a single straight line. Instead, they form a scattered cloud of points.

- We could draw dozens of different lines through the cloud of data points.
- Some lines would be closer to some points, while others would fit different points better.
- **The Question:** How do we mathematically find the single “**best-fitting**” line possible?



The Challenge of Scatter Plots

If we plot bivariate data, the points do not fall perfectly on a single straight line. Instead, they form a scattered cloud of points.

- We could draw dozens of different lines through the cloud of data points.
- Some lines would be closer to some points, while others would fit different points better.
- **The Question:** How do we mathematically find the single “**best-fitting**” line possible?

We use the **Least Squares Method** (also known as Ordinary Least Squares, or OLS) to determine the unique line that minimizes the total error across all data points.



Defining Residuals (Prediction Errors)

For any actual data point (x, y) , the model predicts an estimated value $\hat{y}$ on the line. The vertical distance between the actual point and our predicted line is the **residual** (error):

$$\text{Residual} = y - \hat{y}$$



Defining Residuals (Prediction Errors)

For any actual data point (x, y) , the model predicts an estimated value $\hat{y}$ on the line. The vertical distance between the actual point and our predicted line is the **residual** (error):

$$\text{Residual} = y - \hat{y}$$

- If a point is above the line, the residual is positive ($y - \hat{y} > 0$).
- If a point is below the line, the residual is negative ($y - \hat{y} < 0$).



Defining Residuals (Prediction Errors)

For any actual data point (x, y) , the model predicts an estimated value $\hat{y}$ on the line. The vertical distance between the actual point and our predicted line is the **residual** (error):

$$\text{Residual} = y - \hat{y}$$

- If a point is above the line, the residual is positive ($y - \hat{y} > 0$).
- If a point is below the line, the residual is negative ($y - \hat{y} < 0$).
- If we simply sum the raw residuals, the positive and negative errors cancel out to 0, which is mathematically useless!



The Least Squares Principle

To prevent positive and negative errors from canceling each other out, we square each residual before adding them up.

The Least Squares Principle: The best-fitting regression line is the one that minimizes the sum of the squared vertical distances between the actual y values and the predicted $\hat{y}$ values:

$$\text{Minimize } \Sigma(y - \hat{y})^2$$



The Least Squares Principle

To prevent positive and negative errors from canceling each other out, we square each residual before adding them up.

The Least Squares Principle: The best-fitting regression line is the one that minimizes the sum of the squared vertical distances between the actual y values and the predicted $\hat{y}$ values:

$$\text{Minimize } \Sigma(y - \hat{y})^2$$

By squaring the errors, we also penalize large outliers heavily, forcing our line to fit the entire dataset as consistently as possible.



Visualizing Least Squares and Residuals

Look at the plot on the right:

- The solid blue line is our calculated regression line $\hat{y} = a + bx$.



Visualizing Least Squares and Residuals

Look at the plot on the right:

- The solid blue line is our calculated regression line $\hat{y} = a + bx$.
- The black dots are the actual observed data points (x, y) .



Visualizing Least Squares and Residuals

Look at the plot on the right:

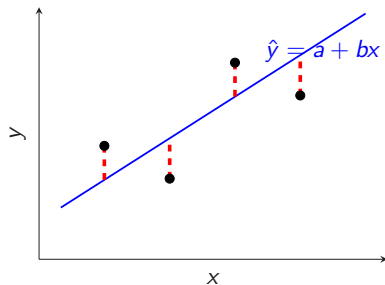
- The solid blue line is our calculated regression line $\hat{y} = a + bx$.
- The black dots are the actual observed data points (x, y) .
- The vertical dashed red lines represent the **residuals** $(y - \hat{y})$.



Visualizing Least Squares and Residuals

Look at the plot on the right:

- The solid blue line is our calculated regression line $\hat{y} = a + bx$.
- The black dots are the actual observed data points (x, y) .
- The vertical dashed red lines represent the **residuals** $(y - \hat{y})$.
- The Least Squares method positions the line so that the sum of the squared lengths of these red dashed lines is minimized!



Formulas for Slope (b) and Intercept (a)

Using calculus to minimize the sum of squared residuals, we derive the following equations for our regression coefficients:

The Slope (b):

$$b = r \left(\frac{s_y}{s_x} \right)$$

Where:

- r = Correlation coefficient.
- s_x, s_y = Standard deviations of x and y .



Formulas for Slope (b) and Intercept (a)

Using calculus to minimize the sum of squared residuals, we derive the following equations for our regression coefficients:

The Slope (b):

$$b = r \left(\frac{s_y}{s_x} \right)$$

Where:

- r = Correlation coefficient.
- s_x, s_y = Standard deviations of x and y .

The Intercept (a):

$$a = \bar{Y} - b\bar{X}$$

Where:

- $\bar{X}, \bar{Y}$ = Sample means of the two variables.
- b = Calculated slope.



A Beautiful Property: The Point of Means

Look closely at the intercept formula: $a = \bar{Y} - b\bar{X}$. If we rearrange this algebraically, we get:

$$\bar{Y} = a + b\bar{X}$$

What does this tell us?



A Beautiful Property: The Point of Means

Look closely at the intercept formula: $a = \bar{Y} - b\bar{X}$. If we rearrange this algebraically, we get:

$$\bar{Y} = a + b\bar{X}$$

What does this tell us?

A unique mathematical property of the Ordinary Least Squares (OLS) line is that it **always** passes exactly through the coordinates of the sample means:

$$(x, y) = (\bar{X}, \bar{Y})$$

This acts as an excellent sanity check when plotting your regression line manually!



Solved Case Study: Copier Sales

Case Study: North American Copier Sales

Textbook Dataset: Expanded Copier Sales Calls ($n = 15$)

Nadia Bancer, sales manager at North American Copier Sales, wants to model the relationship between the number of sales calls made last month (X) and the number of copiers sold (Y).

The study has been expanded to a sample of **15 salespeople** ($n = 15$).



Case Study: North American Copier Sales

Textbook Dataset: Expanded Copier Sales Calls ($n = 15$)

Nadia Bancer, sales manager at North American Copier Sales, wants to model the relationship between the number of sales calls made last month (X) and the number of copiers sold (Y).

The study has been expanded to a sample of **15 salespeople** ($n = 15$).

Summary Statistics calculated from the raw dataset:

- Sample size: $n = 15$
- Sum of Calls: $\Sigma X = 1,440 \implies \bar{X} = 96.00$ calls
- Sum of Copiers Sold: $\Sigma Y = 675 \implies \bar{Y} = 45.00$ copiers
- Sample Standard Deviation of Calls (X): $s_x = 22.40$
- Sample Standard Deviation of Sales (Y): $s_y = 6.41$
- Calculated Correlation Coefficient: $r = 0.909$



Copier Sales: Step 1 - Calculate the Slope (b)

Let's use our summary statistics to calculate the slope (b) first:

$$b = r \left(\frac{s_y}{s_x} \right)$$



Copier Sales: Step 1 - Calculate the Slope (b)

Let's use our summary statistics to calculate the slope (b) first:

$$b = r \left(\frac{s_y}{s_x} \right)$$

Plug in the values:

- $r = 0.909$
- $s_y = 6.41$
- $s_x = 22.40$



Copier Sales: Step 1 - Calculate the Slope (b)

Let's use our summary statistics to calculate the slope (b) first:

$$b = r \left(\frac{s_y}{s_x} \right)$$

Plug in the values:

- $r = 0.909$
- $s_y = 6.41$
- $s_x = 22.40$

$$b = 0.909 \times \left(\frac{6.41}{22.40} \right)$$



Copier Sales: Step 1 - Calculate the Slope (b)

Let's use our summary statistics to calculate the slope (b) first:

$$b = r \left(\frac{s_y}{s_x} \right)$$

Plug in the values:

- $r = 0.909$
- $s_y = 6.41$
- $s_x = 22.40$

$$b = 0.909 \times \left(\frac{6.41}{22.40} \right)$$

$$b = 0.909 \times 0.28616 = \mathbf{0.2601}$$



Copier Sales: Step 2 - Calculate Intercept (a)

Now, use the calculated slope $b = 0.2601$ and the sample means ($\bar{X} = 96.00$, $\bar{Y} = 45.00$) to find the intercept (a):

$$a = \bar{Y} - b\bar{X}$$



Copier Sales: Step 2 - Calculate Intercept (a)

Now, use the calculated slope $b = 0.2601$ and the sample means ($\bar{X} = 96.00$, $\bar{Y} = 45.00$) to find the intercept (a):

$$a = \bar{Y} - b\bar{X}$$

Plug in the values:

$$a = 45.00 - (0.2601 \times 96.00)$$



Copier Sales: Step 2 - Calculate Intercept (a)

Now, use the calculated slope $b = 0.2601$ and the sample means ($\bar{X} = 96.00$, $\bar{Y} = 45.00$) to find the intercept (a):

$$a = \bar{Y} - b\bar{X}$$

Plug in the values:

$$a = 45.00 - (0.2601 \times 96.00)$$

$$a = 45.00 - 24.9696$$



Copier Sales: Step 2 - Calculate Intercept (a)

Now, use the calculated slope $b = 0.2601$ and the sample means ($\bar{X} = 96.00$, $\bar{Y} = 45.00$) to find the intercept (a):

$$a = \bar{Y} - b\bar{X}$$

Plug in the values:

$$a = 45.00 - (0.2601 \times 96.00)$$

$$a = 45.00 - 24.9696$$

$$a = \mathbf{20.03}$$



Copier Sales: The Final Regression Equation

By combining our slope and intercept, we write our estimated regression model:

Copier Sales Prediction Model:

$$\hat{y} = 20.03 + 0.2601x$$



Copier Sales: The Final Regression Equation

By combining our slope and intercept, we write our estimated regression model:

Copier Sales Prediction Model:

$$\hat{y} = 20.03 + 0.2601x$$

How do we interpret these coefficients for the CEO?



Copier Sales: The Final Regression Equation

By combining our slope and intercept, we write our estimated regression model:

Copier Sales Prediction Model:

$$\hat{y} = 20.03 + 0.2601x$$

How do we interpret these coefficients for the CEO?

- **The y -intercept (20.03):** If a salesperson makes 0 sales calls ($x = 0$), they are still expected to sell approximately **20.03 copiers** last month (perhaps due to walk-ins or phone-in orders).



Copier Sales: The Final Regression Equation

By combining our slope and intercept, we write our estimated regression model:

Copier Sales Prediction Model:

$$\hat{y} = 20.03 + 0.2601x$$

How do we interpret these coefficients for the CEO?

- **The y-intercept (20.03):** If a salesperson makes 0 sales calls ($x = 0$), they are still expected to sell approximately **20.03 copiers** last month (perhaps due to walk-ins or phone-in orders).
- **The Slope (0.2601):** For each additional sales call a representative makes, sales are expected to increase by **0.2601 copiers** (or about 1 extra copier for every 4 calls made).



Copier Sales: Making a Prediction

Business Forecast Task:

A new salesperson, Ahmed, is hired at North American Copier Sales. Nadia Bancer instructs him to make exactly **100 sales calls** next month.

How many copiers is Ahmed predicted to sell?



Copier Sales: Making a Prediction

Business Forecast Task:

A new salesperson, Ahmed, is hired at North American Copier Sales. Nadia Bancer instructs him to make exactly **100 sales calls** next month.

How many copiers is Ahmed predicted to sell?

Using our regression equation, substitute $x = 100$:

$$\hat{y} = 20.03 + 0.2601(100)$$



Copier Sales: Making a Prediction

Business Forecast Task:

A new salesperson, Ahmed, is hired at North American Copier Sales. Nadia Bancer instructs him to make exactly **100 sales calls** next month.

How many copiers is Ahmed predicted to sell?

Using our regression equation, substitute $x = 100$:

$$\hat{y} = 20.03 + 0.2601(100)$$

$$\hat{y} = 20.03 + 26.01 = \mathbf{46.04} \text{ copiers}$$

Ahmed is predicted to sell approximately 46 copiers next month.



Evaluating Predictive Power

Measuring Prediction Accuracy

Our model predicts that a salesperson making 100 calls sells 46 copiers. But actual sales will vary around this estimate. How do we measure the accuracy (or precision) of our predictions?



Measuring Prediction Accuracy

Our model predicts that a salesperson making 100 calls sells 46 copiers. But actual sales will vary around this estimate. How do we measure the accuracy (or precision) of our predictions?

To evaluate predictive power, we use two primary statistical metrics:

- 1 **The Standard Error of Estimate ($s_{y \cdot x}$):** Measures the absolute distance that points fall from the regression line.
- 2 **The Coefficient of Determination (r^2):** Measures the relative percentage of variation explained by the line.



The Standard Error of Estimate ($s_{y \cdot x}$)

Standard Error of Estimate Formula:

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{SSE}{n - 2}}$$

Where:

- SSE = Sum of Squares Error (the sum of squared residuals $\Sigma(y - \hat{y})^2$).
- n = Number of paired observations.



The Standard Error of Estimate ($s_{y \cdot x}$)

Standard Error of Estimate Formula:

$$s_{y \cdot x} = \sqrt{\frac{\Sigma(y - \hat{y})^2}{n - 2}} = \sqrt{\frac{SSE}{n - 2}}$$

Where:

- SSE = Sum of Squares Error (the sum of squared residuals $\Sigma(y - \hat{y})^2$).
- n = Number of paired observations.

Degrees of Freedom Alert!

Notice that we divide by $n - 2$ in the denominator. Why? Because we have estimated **two** population parameters from our sample data: the intercept a and the slope b !

Calculating the Standard Error for Copier Sales

For our $n = 15$ copier sales dataset, suppose the computer calculated the Sum of Squares Error as:

$$SSE = \Sigma(y - \hat{y})^2 = 111.45$$



Calculating the Standard Error for Copier Sales

For our $n = 15$ copier sales dataset, suppose the computer calculated the Sum of Squares Error as:

$$SSE = \Sigma(y - \hat{y})^2 = 111.45$$

Let's calculate the Standard Error of Estimate ($s_{y \cdot x}$):

$$s_{y \cdot x} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{\frac{111.45}{15 - 2}}$$



Calculating the Standard Error for Copier Sales

For our $n = 15$ copier sales dataset, suppose the computer calculated the Sum of Squares Error as:

$$SSE = \Sigma(y - \hat{y})^2 = 111.45$$

Let's calculate the Standard Error of Estimate ($s_{y \cdot x}$):

$$s_{y \cdot x} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{\frac{111.45}{15 - 2}}$$

$$s_{y \cdot x} = \sqrt{\frac{111.45}{13}} = \sqrt{8.573} = \mathbf{2.928} \text{ copiers}$$



Calculating the Standard Error for Copier Sales

For our $n = 15$ copier sales dataset, suppose the computer calculated the Sum of Squares Error as:

$$SSE = \Sigma(y - \hat{y})^2 = 111.45$$

Let's calculate the Standard Error of Estimate ($s_{y \cdot x}$):

$$s_{y \cdot x} = \sqrt{\frac{SSE}{n - 2}} = \sqrt{\frac{111.45}{15 - 2}}$$

$$s_{y \cdot x} = \sqrt{\frac{111.45}{13}} = \sqrt{8.573} = \mathbf{2.928} \text{ copiers}$$

What does this standard error mean?

- It measures the typical deviation of the data points from our regression line.
- A standard error of **2.928** means our predicted values will typically be off by about 2.93 copiers from actual sales.



The Coefficient of Determination (r^2) in Regression

We can also evaluate our regression line's fit by looking at the proportion of total variation it explains:

Coefficient of Determination (r^2):

$$r^2 = \frac{SSR}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

Where:

- SSR = Regression Sum of Squares (variation explained by the line).
- $SSTotal$ = Total Sum of Squares (total variation in Y).



The Coefficient of Determination (r^2) in Regression

We can also evaluate our regression line's fit by looking at the proportion of total variation it explains:

Coefficient of Determination (r^2):

$$r^2 = \frac{SSR}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

Where:

- SSR = Regression Sum of Squares (variation explained by the line).
- $SSTotal$ = Total Sum of Squares (total variation in Y).

For our copiers, since $r = 0.909 \implies r^2 = (0.909)^2 = \mathbf{0.8263}$ (**82.63%**).



The Coefficient of Determination (r^2) in Regression

We can also evaluate our regression line's fit by looking at the proportion of total variation it explains:

Coefficient of Determination (r^2):

$$r^2 = \frac{SSR}{SSTotal} = 1 - \frac{SSE}{SSTotal}$$

Where:

- SSR = Regression Sum of Squares (variation explained by the line).
- $SSTotal$ = Total Sum of Squares (total variation in Y).

For our copiers, since $r = 0.909 \implies r^2 = (0.909)^2 = \mathbf{0.8263}$ (**82.63%**).

Exactly 82.63% of the variation in copier sales is explained by the number of sales calls!



Pop-Up Question 1: Assessing Two Models

Discussion

Pop-Up Question!

A retail logistics firm builds two models to estimate warehouse delivery delays (Y):

- **Model A:** $s_{y \cdot x} = 4.2$ minutes, $r^2 = 85\%$
- **Model B:** $s_{y \cdot x} = 8.5$ minutes, $r^2 = 52\%$

Which model is superior for prediction, and why?



Pop-Up Question 1: Assessing Two Models

Discussion

Pop-Up Question!

A retail logistics firm builds two models to estimate warehouse delivery delays (Y):

- **Model A:** $s_{y \cdot x} = 4.2$ minutes, $r^2 = 85\%$
- **Model B:** $s_{y \cdot x} = 8.5$ minutes, $r^2 = 52\%$

Which model is superior for prediction, and why?

Answer:

- **Model A is superior.**
- It has a much **smaller Standard Error of Estimate** (predictions are off by only 4.2 minutes vs. 8.5 minutes).
- It has a much **higher R-squared** (explaining 85% of the variance in delay times vs. 52% for Model B).



Testing the Significance of the Slope

Is X Actually a Significant Predictor of Y ?

Even if our sample slope is $b = 0.2601$, could this relationship occur purely by random chance from a population where the independent variable actually has **zero** effect on the dependent variable? To prove that our predictor variable is statistically useful, we must test whether the true population slope (denoted by the Greek letter β , "**beta**") is significantly different from zero.



Is X Actually a Significant Predictor of Y ?

Even if our sample slope is $b = 0.2601$, could this relationship occur purely by random chance from a population where the independent variable actually has **zero** effect on the dependent variable? To prove that our predictor variable is statistically useful, we must test whether the true population slope (denoted by the Greek letter β , "**beta**") is significantly different from zero.

Standard Hypotheses for Slope Significance:

- $H_0 : \beta = 0$ (The population slope is zero; X cannot predict Y)
- $H_1 : \beta \neq 0$ (The population slope is non-zero; X is a useful predictor)



The t -Test Statistic for the Slope

To test our slope, we transform the sample slope b into a t -statistic using its standard error (s_b):

t -Test Statistic for the Slope:

$$t = \frac{b - 0}{s_b}$$

Where s_b is the standard error of the slope:

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\sum (x - \bar{x})^2}}$$

The test statistic follows Student's t -distribution with $df = n - 2$ degrees of freedom.



Testing Copier Sales Slope: Steps 1 to 4

Testing the Copier Sales Slope ($n = 15, b = 0.2601, s_{y \cdot x} = 2.928$)

A computer calculation of our copier calls dataset yields $\Sigma(x - \bar{x})^2 = 7,014.5$.

Task: Test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ at the **0.05 level of significance**.



Testing Copier Sales Slope: Steps 1 to 4

Testing the Copier Sales Slope ($n = 15$, $b = 0.2601$, $s_{y \cdot x} = 2.928$)

A computer calculation of our copier calls dataset yields $\Sigma(x - \bar{x})^2 = 7,014.5$.

Task: Test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ at the **0.05 level of significance**.

- **Step 1:** $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$ (Two-Tailed)
- **Step 2:** $\alpha = 0.05$
- **Step 3:** Use t -statistic.



Testing Copier Sales Slope: Steps 1 to 4

Testing the Copier Sales Slope ($n = 15$, $b = 0.2601$, $s_{y \cdot x} = 2.928$)

A computer calculation of our copier calls dataset yields $\Sigma(x - \bar{x})^2 = 7,014.5$.

Task: Test $H_0 : \beta = 0$ against $H_1 : \beta \neq 0$ at the **0.05 level of significance**.

- **Step 1:** $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$ (Two-Tailed)
- **Step 2:** $\alpha = 0.05$
- **Step 3:** Use t -statistic.
- **Step 4: Decision Rule.**
 - ▶ $df = n - 2 = 15 - 2 = 13$.
 - ▶ Two-Tailed lookup at $\alpha = 0.05 \implies$ **Critical** $t = \pm 2.160$ (Table B.5).



Testing Copier Sales Slope: Steps 5 & 6

Step 5: Compute the standard error of the slope s_b and the t -statistic.

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\sum(x - \bar{x})^2}} = \frac{2.928}{\sqrt{7,014.5}} = \frac{2.928}{83.75} = \mathbf{0.03496}$$



Testing Copier Sales Slope: Steps 5 & 6

Step 5: Compute the standard error of the slope s_b and the t -statistic.

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\sum(x - \bar{x})^2}} = \frac{2.928}{\sqrt{7,014.5}} = \frac{2.928}{83.75} = \mathbf{0.03496}$$

Now find the t -statistic:

$$t = \frac{b - 0}{s_b} = \frac{0.2601 - 0}{0.03496} = \mathbf{7.44}$$



Testing Copier Sales Slope: Steps 5 & 6

Step 5: Compute the standard error of the slope s_b and the t -statistic.

$$s_b = \frac{s_{y \cdot x}}{\sqrt{\sum(x - \bar{x})^2}} = \frac{2.928}{\sqrt{7,014.5}} = \frac{2.928}{83.75} = \mathbf{0.03496}$$

Now find the t -statistic:

$$t = \frac{b - 0}{s_b} = \frac{0.2601 - 0}{0.03496} = \mathbf{7.44}$$

Step 6: Make a Decision and Interpret.

Because computed $t = 7.44 > 2.160$, we **Reject H_0** . The slope is highly statistically significant! Sales call volume is a valid predictor of copier sales.



Interval Estimates of Prediction

Prediction Uncertainty: The Point vs. Interval

Ahmed's predicted sales value for 100 calls was a single number: **46.04 copiers**. But in reality, predictions contain natural uncertainty. We should build an interval around our estimate.



Prediction Uncertainty: The Point vs. Interval

Ahmed's predicted sales value for 100 calls was a single number: **46.04 copiers**. But in reality, predictions contain natural uncertainty. We should build an interval around our estimate. To do this correctly, we must identify **which** of two distinct business questions we are trying to answer:



Prediction Uncertainty: The Point vs. Interval

Ahmed's predicted sales value for 100 calls was a single number: **46.04 copiers**. But in reality, predictions contain natural uncertainty. We should build an interval around our estimate. To do this correctly, we must identify **which** of two distinct business questions we are trying to answer:

Question A:

*"What is the average sales volume for **all** representatives who make 100 calls?"*



Prediction Uncertainty: The Point vs. Interval

Ahmed's predicted sales value for 100 calls was a single number: **46.04 copiers**. But in reality, predictions contain natural uncertainty. We should build an interval around our estimate. To do this correctly, we must identify **which** of two distinct business questions we are trying to answer:

Question A:

*"What is the average sales volume for **all** representatives who make 100 calls?"*

Question B:

*"What is the predicted sales volume for **one specific** salesperson (Ahmed) who makes 100 calls?"*



Confidence Interval vs. Prediction Interval

To answer these two different questions, statisticians construct two different intervals:

A. Confidence Interval (CI): Estimates the mean (average) value of Y for the entire population at a chosen value of x .



Confidence Interval vs. Prediction Interval

To answer these two different questions, statisticians construct two different intervals:

A. Confidence Interval (CI): Estimates the mean (average) value of Y for the entire population at a chosen value of x .

B. Prediction Interval (PI): Estimates the predicted Y value for a **single, specific** individual at a chosen value of x .



Confidence Interval vs. Prediction Interval

To answer these two different questions, statisticians construct two different intervals:

A. Confidence Interval (CI): Estimates the mean (average) value of Y for the entire population at a chosen value of x .

B. Prediction Interval (PI): Estimates the predicted Y value for a **single, specific** individual at a chosen value of x .

Which is wider? An individual's behavior is always far more variable and harder to predict than the average of a group. Therefore, the **Prediction Interval is always wider** than the Confidence Interval!



Comparing the Interval Formulas

Let's compare the equations side-by-side:

Confidence Interval of the Mean (Group Average):

$$\hat{y} \pm ts_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$



Comparing the Interval Formulas

Let's compare the equations side-by-side:

Confidence Interval of the Mean (Group Average):

$$\hat{y} \pm ts_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Prediction Interval of an Individual (Y):

$$\hat{y} \pm ts_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$



Comparing the Interval Formulas

Let's compare the equations side-by-side:

Confidence Interval of the Mean (Group Average):

$$\hat{y} \pm ts_{y \cdot x} \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Prediction Interval of an Individual (Y):

$$\hat{y} \pm ts_{y \cdot x} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum (x - \bar{x})^2}}$$



Visualizing CI and PI Bands Geometrically

Look at the bands on the right:

- The central straight line is our regression line $\hat{y}$.



Visualizing CI and PI Bands Geometrically

Look at the bands on the right:

- The central straight line is our regression line $\hat{y}$.
- The inner curved bands represent the **Confidence Interval of the Mean**.



Visualizing CI and PI Bands Geometrically

Look at the bands on the right:

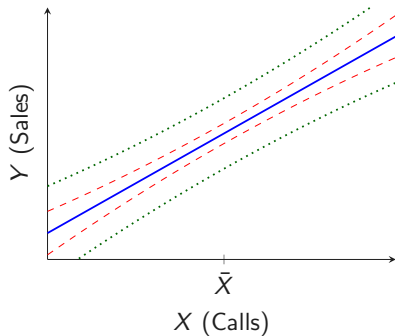
- The central straight line is our regression line $\hat{y}$.
- The inner curved bands represent the **Confidence Interval of the Mean**.
- The outer curved bands represent the **Individual Prediction Interval**.



Visualizing CI and PI Bands Geometrically

Look at the bands on the right:

- The central straight line is our regression line $\hat{y}$.
- The inner curved bands represent the **Confidence Interval of the Mean**.
- The outer curved bands represent the **Individual Prediction Interval**.
- Notice that both intervals are **narrowest at the sample mean $\bar{X}$** (where $x - \bar{x} = 0$), and flare out as we move further away from our data center.



Case Study: Calculating the 95% Confidence Interval

Task: Calculate a 95% confidence interval for the **mean sales** of all representatives who make exactly $x = 100$ sales calls.

$n = 15, \bar{X} = 96.00, s_{y \cdot x} = 2.928, \Sigma(x - \bar{x})^2 = 7,014.5, t_{13} = 2.160.$

Note: $\hat{y}$ for 100 calls is 46.04 copiers.



Case Study: Calculating the 95% Confidence Interval

Task: Calculate a 95% confidence interval for the **mean sales** of all representatives who make exactly $x = 100$ sales calls.

$n = 15, \bar{X} = 96.00, s_{y \cdot x} = 2.928, \Sigma(x - \bar{x})^2 = 7,014.5, t_{13} = 2.160.$

Note: $\hat{y}$ for 100 calls is 46.04 copiers.

Step 1: Calculate the radical term.

$$\text{Radical} = \sqrt{\frac{1}{15} + \frac{(100 - 96.00)^2}{7,014.5}} = \sqrt{0.0667 + \frac{16}{7,014.5}} = \sqrt{0.0667 + 0.00228} = \sqrt{0.06898} = 0.2626$$



Case Study: Calculating the 95% Confidence Interval

Task: Calculate a 95% confidence interval for the **mean sales** of all representatives who make exactly $x = 100$ sales calls.

$n = 15, \bar{X} = 96.00, s_{y \cdot x} = 2.928, \Sigma(x - \bar{x})^2 = 7,014.5, t_{13} = 2.160.$

Note: $\hat{y}$ for 100 calls is 46.04 copiers.

Step 1: Calculate the radical term.

$$\text{Radical} = \sqrt{\frac{1}{15} + \frac{(100 - 96.00)^2}{7,014.5}} = \sqrt{0.0667 + \frac{16}{7,014.5}} = \sqrt{0.0667 + 0.00228} = \sqrt{0.06898} = 0.2626$$

Step 2: Compute Interval limits.

$$CI = 46.04 \pm (2.160 \times 2.928 \times 0.2626) = 46.04 \pm 1.66 = [44.38, 47.70] \text{ copiers}$$



Case Study: Calculating the 95% Prediction Interval

Task: Calculate a 95% prediction interval for the **individual sales** of Ahmed, who plans to make exactly $x = 100$ sales calls.



Case Study: Calculating the 95% Prediction Interval

Task: Calculate a 95% prediction interval for the **individual sales** of Ahmed, who plans to make exactly $x = 100$ sales calls.

Step 1: Calculate the radical term (with the extra "+1").

$$\text{Radical} = \sqrt{1 + \frac{1}{15} + \frac{(100 - 96.00)^2}{7,014.5}} = \sqrt{1 + 0.0667 + 0.00228} = \sqrt{1.06898} = \mathbf{1.0339}$$



Case Study: Calculating the 95% Prediction Interval

Task: Calculate a 95% prediction interval for the **individual sales** of Ahmed, who plans to make exactly $x = 100$ sales calls.

Step 1: Calculate the radical term (with the extra "+1").

$$\text{Radical} = \sqrt{1 + \frac{1}{15} + \frac{(100 - 96.00)^2}{7,014.5}} = \sqrt{1 + 0.0667 + 0.00228} = \sqrt{1.06898} = \mathbf{1.0339}$$

Step 2: Compute Interval limits.

$$\text{PI} = 46.04 \pm (2.160 \times 2.928 \times 1.0339) = 46.04 \pm 6.54 = \mathbf{[39.50, 52.58]} \text{ copiers}$$



Case Study: Calculating the 95% Prediction Interval

Task: Calculate a 95% prediction interval for the **individual sales** of Ahmed, who plans to make exactly $x = 100$ sales calls.

Step 1: Calculate the radical term (with the extra "+1").

$$\text{Radical} = \sqrt{1 + \frac{1}{15} + \frac{(100 - 96.00)^2}{7,014.5}} = \sqrt{1 + 0.0667 + 0.00228} = \sqrt{1.06898} = \mathbf{1.0339}$$

Step 2: Compute Interval limits.

$$\text{PI} = 46.04 \pm (2.160 \times 2.928 \times 1.0339) = 46.04 \pm 6.54 = \mathbf{[39.50, 52.58]} \text{ copiers}$$

- Group Average CI: **[44.38 to 47.70]** (Width: 3.32 copiers)
- Individual Ahmed PI: **[39.50 to 52.58]** (Width: 13.08 copiers)



Windows Excel Software Application

Excel for Windows: Running Linear Regression

We will now run this analysis using Microsoft Excel on our Windows computers.

- 1 Ensure your data is entered in two columns (Calls in Column B, Sales in Column C).
- 2 Click on the **Data** tab on the top menu ribbon.
- 3 On the far right, select **Data Analysis**.
- 4 Scroll down, select **Regression**, and click **OK**.



Excel for Windows: Running Linear Regression

We will now run this analysis using Microsoft Excel on our Windows computers.

- 1 Ensure your data is entered in two columns (Calls in Column B, Sales in Column C).
- 2 Click on the **Data** tab on the top menu ribbon.
- 3 On the far right, select **Data Analysis**.
- 4 Scroll down, select **Regression**, and click **OK**.

The Menu Input Requirements:

- **Input Y Range:** Select your dependent variable (*C1 : C16*, sales).
- **Input X Range:** Select your independent variable (*B1 : B16*, calls).
- Check the **Labels** box (since row 1 has headings).
- Select **Output Range** and type a cell coordinate (e.g., *E1*).



Reading the Excel Output - Regression Statistics

Excel outputs three tables of data. Let's learn how to read them:

Regression Statistics Table	
Multiple R	0.909
R Square	0.826
Adjusted R Square	0.813
Standard Error	2.928
Observations	15



Reading the Excel Output - Regression Statistics

Excel outputs three tables of data. Let's learn how to read them:

Regression Statistics Table	
Multiple R	0.909
R Square	0.826
Adjusted R Square	0.813
Standard Error	2.928
Observations	15

What do we recognize?

- **Multiple R:** The correlation coefficient ($r = 0.909$).
- **R Square:** The coefficient of determination ($r^2 = 82.6\%$).
- **Standard Error:** The Standard Error of Estimate ($s_{y \cdot x} = 2.928$).



Reading the Excel Output - Coefficient Table

Now look at the bottom table, which gives us the regression coefficients:

	Coefficients	Standard Error	t Stat	P-value
Intercept	20.031	3.421	5.855	0.0001
Sales Calls	0.2601	0.035	7.442	0.0000



Reading the Excel Output - Coefficient Table

Now look at the bottom table, which gives us the regression coefficients:

	Coefficients	Standard Error	t Stat	P-value
Intercept	20.031	3.421	5.855	0.0001
Sales Calls	0.2601	0.035	7.442	0.0000

Extracting the Equations and Hypothesis Tests:

- Under **Coefficients**, we read our intercept ($a = 20.031$) and slope ($b = 0.2601$).
- For the Sales Calls variable, the computed **t Stat** is **7.442**.
- The **P-value** is listed as **0.0000**. Since $0.0000 < 0.05$, we reject H_0 immediately!



Pop-Up Question 2: Reading Excel Output

Discussion

You run a regression in Excel to estimate ice cream sales (Y , in dollars) based on daily temperature (X , in degrees Fahrenheit).

The Excel output shows:

Sales intercept coefficient: -150.00

Temperature slope coefficient: 8.50

Temperature P-value: 0.012

1. What is the regression equation?
2. Is temperature a significant predictor of sales at $\alpha = 0.05$?



Pop-Up Question 2: Reading Excel Output

Discussion

You run a regression in Excel to estimate ice cream sales (Y , in dollars) based on daily temperature (X , in degrees Fahrenheit).

The Excel output shows:

Sales intercept coefficient: -150.00

Temperature slope coefficient: 8.50

Temperature P-value: 0.012

1. What is the regression equation?
2. Is temperature a significant predictor of sales at $\alpha = 0.05$?

Answers:

- **1:** $\hat{y} = -150.00 + 8.50x$.
- **2: Yes.** Since the P-value (0.012) is less than $\alpha = 0.05$, we reject H_0 and conclude that temperature is a highly significant predictor.

