

Chapter 9: Regression Analysis

Descriptive Statistics

From correlation to prediction

Chapter 8 asked:

Is there a linear relationship between x and y ?

From correlation to prediction

Chapter 8 asked:

Is there a linear relationship between x and y ?

Chapter 9 asks:

Can we use x to predict y ?

From correlation to prediction

Chapter 8 asked:

Is there a linear relationship between x and y ?

Chapter 9 asks:

Can we use x to predict y ?

The main tool is the **least-squares regression line**.

Independent and dependent variables

In regression, the two variables do **not** play the same role.

Independent and dependent variables

In regression, the two variables do **not** play the same role.

x = independent or predictor variable

y = dependent or response variable

Independent and dependent variables

In regression, the two variables do **not** play the same role.

x = independent or predictor variable

y = dependent or response variable

We use a known value of x to predict a value of y .

Independent and dependent variables

In regression, the two variables do **not** play the same role.

x = independent or predictor variable

y = dependent or response variable

We use a known value of x to predict a value of y .

The regression of y on x is generally **not** the same as the regression of x on y .

Chapter roadmap

By the end of this chapter, students should be able to:

- 1 understand what least squares is minimizing;
- 2 compute a regression line from summary statistics;
- 3 use a regression equation to make predictions;
- 4 interpret residuals and prediction error;
- 5 interpret r^2 , the coefficient of determination;
- 6 estimate the spread around the regression line;
- 7 recognize extrapolation;
- 8 understand regression to the mean;
- 9 use transformations to model some nonlinear relationships.

Skill 1: What does “best-fit line” mean?

Suppose we try a line

$$\hat{y} = \hat{a} + \hat{b}x.$$

Skill 1: What does “best-fit line” mean?

Suppose we try a line

$$\hat{y} = \hat{a} + \hat{b}x.$$

For a point (x_i, y_i) , the line predicts

$$\hat{y}_i = \hat{a} + \hat{b}x_i.$$

Skill 1: What does “best-fit line” mean?

Suppose we try a line

$$\hat{y} = \hat{a} + \hat{b}x.$$

For a point (x_i, y_i) , the line predicts

$$\hat{y}_i = \hat{a} + \hat{b}x_i.$$

The vertical difference between the actual and predicted values is

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

Skill 1: What does “best-fit line” mean?

Suppose we try a line

$$\hat{y} = \hat{a} + \hat{b}x.$$

For a point (x_i, y_i) , the line predicts

$$\hat{y}_i = \hat{a} + \hat{b}x_i.$$

The vertical difference between the actual and predicted values is

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

This difference is called a **residual**.

Residuals are vertical distances with signs

For one data point:

$$e = y - \hat{y}.$$

Residuals are vertical distances with signs

For one data point:

$$e = y - \hat{y}.$$

If the point lies **above** the line:

$$y > \hat{y} \quad \Rightarrow \quad e > 0.$$

Residuals are vertical distances with signs

For one data point:

$$e = y - \hat{y}.$$

If the point lies **above** the line:

$$y > \hat{y} \quad \Rightarrow \quad e > 0.$$

If the point lies **below** the line:

$$y < \hat{y} \quad \Rightarrow \quad e < 0.$$

Residuals are vertical distances with signs

For one data point:

$$e = y - \hat{y}.$$

If the point lies **above** the line:

$$y > \hat{y} \Rightarrow e > 0.$$

If the point lies **below** the line:

$$y < \hat{y} \Rightarrow e < 0.$$

If the point lies exactly on the line:

$$e = 0.$$

Why square the residuals?

Positive and negative residuals can cancel if we simply add them.

Why square the residuals?

Positive and negative residuals can cancel if we simply add them.

Instead, we use:

$$e_1^2 + e_2^2 + \cdots + e_n^2.$$

Why square the residuals?

Positive and negative residuals can cancel if we simply add them.
Instead, we use:

$$e_1^2 + e_2^2 + \cdots + e_n^2.$$

The least-squares line is the line that makes

$$\sum e_i^2$$

as small as possible.

Why square the residuals?

Positive and negative residuals can cancel if we simply add them.
Instead, we use:

$$e_1^2 + e_2^2 + \cdots + e_n^2.$$

The least-squares line is the line that makes

$$\sum e_i^2$$

as small as possible.

Equivalently:

$$\sum_{i=1}^n \left(y_i - (\hat{a} + \hat{b}x_i) \right)^2 \text{ is minimized.}$$

Example 1A: Compare two candidate lines

Data:

$$(1, 2), \quad (2, 3), \quad (3, 5).$$

Compare:

$$L_1 : \hat{y} = x + 1$$

and

$$L_2 : \hat{y} = 1.5x + 0.5.$$

Example 1A: Compare two candidate lines

Data:

$$(1, 2), \quad (2, 3), \quad (3, 5).$$

Compare:

$$L_1 : \hat{y} = x + 1$$

and

$$L_2 : \hat{y} = 1.5x + 0.5.$$

We will compute the residuals and the sum of squared residuals for each line.

Example 1A: Candidate line L_1

For

$$L_1 : \hat{y} = x + 1,$$

the predicted values are:

2, 3, 4.

Example 1A: Candidate line L_1

For

$$L_1 : \hat{y} = x + 1,$$

the predicted values are:

2, 3, 4.

Actual values:

2, 3, 5.

Example 1A: Candidate line L_1

For

$$L_1 : \hat{y} = x + 1,$$

the predicted values are:

2, 3, 4.

Actual values:

2, 3, 5.

Residuals:

0, 0, 1.

Example 1A: Candidate line L_1

For

$$L_1 : \hat{y} = x + 1,$$

the predicted values are:

$$2, 3, 4.$$

Actual values:

$$2, 3, 5.$$

Residuals:

$$0, 0, 1.$$

Sum of squared residuals:

$$0^2 + 0^2 + 1^2 = \boxed{1}.$$

Example 1A: Candidate line L_2

For

$$L_2 : \hat{y} = 1.5x + 0.5,$$

the predicted values are:

2, 3.5, 5.

Example 1A: Candidate line L_2

For

$$L_2 : \hat{y} = 1.5x + 0.5,$$

the predicted values are:

$$2, 3.5, 5.$$

Residuals:

$$0, -0.5, 0.$$

Example 1A: Candidate line L_2

For

$$L_2 : \hat{y} = 1.5x + 0.5,$$

the predicted values are:

$$2, 3.5, 5.$$

Residuals:

$$0, -0.5, 0.$$

Sum of squared residuals:

$$0^2 + (-0.5)^2 + 0^2 = \boxed{0.25}.$$

Example 1A: Candidate line L_2

For

$$L_2 : \hat{y} = 1.5x + 0.5,$$

the predicted values are:

$$2, 3.5, 5.$$

Residuals:

$$0, -0.5, 0.$$

Sum of squared residuals:

$$0^2 + (-0.5)^2 + 0^2 = \boxed{0.25}.$$

So L_2 fits these three points better than L_1 in the least-squares sense.

Example 1B: Find a residual

Suppose the regression line is

$$\hat{y} = 12 + 3x.$$

For a data point with

$$x = 5, \quad y = 30,$$

first predict:

$$\hat{y} = 12 + 3(5).$$

Example 1B: Find a residual

Suppose the regression line is

$$\hat{y} = 12 + 3x.$$

For a data point with

$$x = 5, \quad y = 30,$$

first predict:

$$\hat{y} = 12 + 3(5).$$

$$\hat{y} = 27.$$

Example 1B: Find a residual

Suppose the regression line is

$$\hat{y} = 12 + 3x.$$

For a data point with

$$x = 5, \quad y = 30,$$

first predict:

$$\hat{y} = 12 + 3(5).$$

$$\hat{y} = 27.$$

Residual:

$$e = y - \hat{y} = 30 - 27 = \boxed{3}.$$

Example 1B: Find a residual

Suppose the regression line is

$$\hat{y} = 12 + 3x.$$

For a data point with

$$x = 5, \quad y = 30,$$

first predict:

$$\hat{y} = 12 + 3(5).$$

$$\hat{y} = 27.$$

Residual:

$$e = y - \hat{y} = 30 - 27 = \boxed{3}.$$

The actual point lies 3 units above the regression line.

Skill 2: The regression-line formulas

The regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x.$$

Skill 2: The regression-line formulas

The regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x.$$

Its slope is:

$$\hat{b} = r \frac{s_y}{s_x}.$$

Skill 2: The regression-line formulas

The regression line of y on x is

$$\hat{y} = \hat{a} + \hat{b}x.$$

Its slope is:

$$\hat{b} = r \frac{s_y}{s_x}.$$

Its intercept is:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

A second form of the regression line

The regression line always passes through

$$(\bar{x}, \bar{y}).$$

A second form of the regression line

The regression line always passes through

$$(\bar{x}, \bar{y}).$$

Therefore it can also be written in point-slope form:

$$\hat{y} - \bar{y} = \hat{b}(x - \bar{x}).$$

A second form of the regression line

The regression line always passes through

$$(\bar{x}, \bar{y}).$$

Therefore it can also be written in point-slope form:

$$\hat{y} - \bar{y} = \hat{b}(x - \bar{x}).$$

Since

$$\hat{b} = r \frac{s_y}{s_x},$$

we also have:

$$\hat{y} - \bar{y} = r \frac{s_y}{s_x} (x - \bar{x}).$$

What the slope means

Suppose

$$\hat{b} = 2.4.$$

What the slope means

Suppose

$$\hat{b} = 2.4.$$

Then an increase of 1 unit in x corresponds to an increase of about

2.4 units in the predicted y .

What the slope means

Suppose

$$\hat{b} = 2.4.$$

Then an increase of 1 unit in x corresponds to an increase of about

2.4 units in the predicted y .

If

$$\hat{b} = -2.4,$$

then an increase of 1 unit in x corresponds to a decrease of about

2.4

units in predicted y .

The sign of the slope

Because

$$\hat{b} = r \frac{s_y}{s_x}$$

and standard deviations are positive:

The sign of the slope

Because

$$\hat{b} = r \frac{s_y}{s_x}$$

and standard deviations are positive:

$$r > 0 \quad \Rightarrow \quad \hat{b} > 0,$$

$$r < 0 \quad \Rightarrow \quad \hat{b} < 0.$$

The sign of the slope

Because

$$\hat{b} = r \frac{s_y}{s_x}$$

and standard deviations are positive:

$$r > 0 \quad \Rightarrow \quad \hat{b} > 0,$$

$$r < 0 \quad \Rightarrow \quad \hat{b} < 0.$$

So the regression line has the same direction as the correlation.

Example 2A: Find the regression line

Suppose:

$$\bar{x} = 10, \quad s_x = 2,$$

$$\bar{y} = 30, \quad s_y = 6,$$

$$r = 0.50.$$

Find the regression line of y on x .

Example 2A: Step 1, find the slope

Use:

$$\hat{b} = r \frac{s_y}{s_x}.$$

Example 2A: Step 1, find the slope

Use:

$$\hat{b} = r \frac{s_y}{s_x}.$$

$$\hat{b} = 0.50 \left(\frac{6}{2} \right).$$

Example 2A: Step 1, find the slope

Use:

$$\hat{b} = r \frac{s_y}{s_x}.$$

$$\hat{b} = 0.50 \left(\frac{6}{2} \right).$$

$$\hat{b} = 0.50(3) = \boxed{1.5}.$$

Example 2A: Step 2, find the intercept

Use:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

Example 2A: Step 2, find the intercept

Use:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

$$\hat{a} = 30 - (1.5)(10).$$

Example 2A: Step 2, find the intercept

Use:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

$$\hat{a} = 30 - (1.5)(10).$$

$$\hat{a} = 30 - 15 = \boxed{15}.$$

Example 2A: Step 2, find the intercept

Use:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}.$$

$$\hat{a} = 30 - (1.5)(10).$$

$$\hat{a} = 30 - 15 = \boxed{15}.$$

Therefore:

$$\boxed{\hat{y} = 15 + 1.5x}.$$

Example 2A: Quick check

The regression line must pass through

$$(\bar{x}, \bar{y}) = (10, 30).$$

Example 2A: Quick check

The regression line must pass through

$$(\bar{x}, \bar{y}) = (10, 30).$$

Substitute $x = 10$:

$$\hat{y} = 15 + 1.5(10).$$

Example 2A: Quick check

The regression line must pass through

$$(\bar{x}, \bar{y}) = (10, 30).$$

Substitute $x = 10$:

$$\hat{y} = 15 + 1.5(10).$$

$$\hat{y} = 30.$$

Example 2A: Quick check

The regression line must pass through

$$(\bar{x}, \bar{y}) = (10, 30).$$

Substitute $x = 10$:

$$\hat{y} = 15 + 1.5(10).$$

$$\hat{y} = 30.$$

So the line passes through the required point.

Example 2B: Negative correlation

Suppose:

$$\bar{x} = 3.5, \quad s_x = 1.5,$$

$$\bar{y} = 5.5, \quad s_y = 0.75,$$

$$r = -0.8.$$

Example 2B: Negative correlation

Suppose:

$$\bar{x} = 3.5, \quad s_x = 1.5,$$

$$\bar{y} = 5.5, \quad s_y = 0.75,$$

$$r = -0.8.$$

Slope:

$$\hat{b} = -0.8 \left(\frac{0.75}{1.5} \right) = -0.8(0.5) = \boxed{-0.4}.$$

Example 2B: Negative correlation

Suppose:

$$\bar{x} = 3.5, \quad s_x = 1.5,$$

$$\bar{y} = 5.5, \quad s_y = 0.75,$$

$$r = -0.8.$$

Slope:

$$\hat{b} = -0.8 \left(\frac{0.75}{1.5} \right) = -0.8(0.5) = \boxed{-0.4}.$$

Intercept:

$$\hat{a} = 5.5 - (-0.4)(3.5) = 5.5 + 1.4 = \boxed{6.9}.$$

Example 2B: Negative correlation

Suppose:

$$\bar{x} = 3.5, \quad s_x = 1.5,$$

$$\bar{y} = 5.5, \quad s_y = 0.75,$$

$$r = -0.8.$$

Slope:

$$\hat{b} = -0.8 \left(\frac{0.75}{1.5} \right) = -0.8(0.5) = \boxed{-0.4}.$$

Intercept:

$$\hat{a} = 5.5 - (-0.4)(3.5) = 5.5 + 1.4 = \boxed{6.9}.$$

$$\boxed{\hat{y} = 6.9 - 0.4x}.$$

Example 2C: Use point-slope form directly

Suppose:

$$\bar{x} = 20, \quad \bar{y} = 50, \quad \hat{b} = 2.$$

Example 2C: Use point-slope form directly

Suppose:

$$\bar{x} = 20, \quad \bar{y} = 50, \quad \hat{b} = 2.$$

Instead of finding the intercept first:

$$\hat{y} - 50 = 2(x - 20).$$

Example 2C: Use point-slope form directly

Suppose:

$$\bar{x} = 20, \quad \bar{y} = 50, \quad \hat{b} = 2.$$

Instead of finding the intercept first:

$$\hat{y} - 50 = 2(x - 20).$$

Expand:

$$\hat{y} - 50 = 2x - 40.$$

Example 2C: Use point-slope form directly

Suppose:

$$\bar{x} = 20, \quad \bar{y} = 50, \quad \hat{b} = 2.$$

Instead of finding the intercept first:

$$\hat{y} - 50 = 2(x - 20).$$

Expand:

$$\hat{y} - 50 = 2x - 40.$$

$$\boxed{\hat{y} = 10 + 2x}.$$

Skill 3: Use the regression line to predict

Once the regression equation is known, prediction is simple.

Skill 3: Use the regression line to predict

Once the regression equation is known, prediction is simple.

If:

$$\hat{y} = \hat{a} + \hat{b}x,$$

and we are given a value of x :

Skill 3: Use the regression line to predict

Once the regression equation is known, prediction is simple.
If:

$$\hat{y} = \hat{a} + \hat{b}x,$$

and we are given a value of x :

- 1 substitute the given x ;
- 2 simplify;
- 3 label the result as a **predicted** value.

Example 3A: Basic prediction

Suppose:

$$\hat{y} = 15 + 1.5x.$$

Predict y when

$$x = 14.$$

Example 3A: Basic prediction

Suppose:

$$\hat{y} = 15 + 1.5x.$$

Predict y when

$$x = 14.$$

$$\hat{y} = 15 + 1.5(14).$$

Example 3A: Basic prediction

Suppose:

$$\hat{y} = 15 + 1.5x.$$

Predict y when

$$x = 14.$$

$$\hat{y} = 15 + 1.5(14).$$

$$\hat{y} = 15 + 21.$$

Example 3A: Basic prediction

Suppose:

$$\hat{y} = 15 + 1.5x.$$

Predict y when

$$x = 14.$$

$$\hat{y} = 15 + 1.5(14).$$

$$\hat{y} = 15 + 21.$$

$$\boxed{\hat{y} = 36}.$$

Predicted value versus actual value

The notation

$$\hat{y}$$

means a **predicted** value.

Predicted value versus actual value

The notation

$$\hat{y}$$

means a **predicted** value.

It does not mean that every observation with that x -value will actually have exactly that y -value.

Predicted value versus actual value

The notation

$$\hat{y}$$

means a **predicted** value.

It does not mean that every observation with that x -value will actually have exactly that y -value.

The actual value

$$y$$

may be above or below the line.

Predicted value versus actual value

The notation

$$\hat{y}$$

means a **predicted** value.

It does not mean that every observation with that x -value will actually have exactly that y -value.

The actual value

$$y$$

may be above or below the line.

Regression predicts a typical or average response near that x .

Example 3B: Interpret the prediction

Suppose the regression line relating hours studied to exam score is

$$\hat{y} = 50 + 6x.$$

For a student who studies 4 hours:

$$\hat{y} = 50 + 6(4) = 74.$$

Example 3B: Interpret the prediction

Suppose the regression line relating hours studied to exam score is

$$\hat{y} = 50 + 6x.$$

For a student who studies 4 hours:

$$\hat{y} = 50 + 6(4) = 74.$$

Interpretation:

Students studying about 4 hours are predicted to score about 74 on average.

Example 3B: Interpret the prediction

Suppose the regression line relating hours studied to exam score is

$$\hat{y} = 50 + 6x.$$

For a student who studies 4 hours:

$$\hat{y} = 50 + 6(4) = 74.$$

Interpretation:

Students studying about 4 hours are predicted to score about 74 on average.

It is not a guarantee that a particular student will score exactly 74.

Example 3C: Reverse thinking

Suppose:

$$\hat{y} = 20 + 0.6x.$$

A student has $x = 80$.

Example 3C: Reverse thinking

Suppose:

$$\hat{y} = 20 + 0.6x.$$

A student has $x = 80$.

Prediction:

$$\hat{y} = 20 + 0.6(80) = 68.$$

Example 3C: Reverse thinking

Suppose:

$$\hat{y} = 20 + 0.6x.$$

A student has $x = 80$.

Prediction:

$$\hat{y} = 20 + 0.6(80) = 68.$$

If the actual response is

$$y = 75,$$

then the residual is:

$$75 - 68 = \boxed{7}.$$

Example 3C: Reverse thinking

Suppose:

$$\hat{y} = 20 + 0.6x.$$

A student has $x = 80$.

Prediction:

$$\hat{y} = 20 + 0.6(80) = 68.$$

If the actual response is

$$y = 75,$$

then the residual is:

$$75 - 68 = \boxed{7}.$$

The actual value is 7 units above the prediction.

Skill 4: How accurate is a prediction?

The regression line gives a center for predicted y -values.

Skill 4: How accurate is a prediction?

The regression line gives a center for predicted y -values.
But actual values scatter around the line.

Skill 4: How accurate is a prediction?

The regression line gives a center for predicted y -values.

But actual values scatter around the line.

The chapter estimates this remaining spread by:

$$s_{y|x} \approx s_y \sqrt{1 - r^2}.$$

Skill 4: How accurate is a prediction?

The regression line gives a center for predicted y -values.

But actual values scatter around the line.

The chapter estimates this remaining spread by:

$$s_{y|x} \approx s_y \sqrt{1 - r^2}.$$

This is the approximate standard deviation of y around the regression prediction when x is known.

Understanding the formula

$$s_{y|x} \approx s_y \sqrt{1 - r^2}.$$

Understanding the formula

$$s_{y|x} \approx s_y \sqrt{1 - r^2}.$$

If

$$|r| \approx 1,$$

then

$$1 - r^2 \approx 0,$$

so:

$$s_{y|x} \approx 0.$$

Understanding the formula

$$s_{y|x} \approx s_y \sqrt{1 - r^2}.$$

If

$$|r| \approx 1,$$

then

$$1 - r^2 \approx 0,$$

so:

$$s_{y|x} \approx 0.$$

Strong correlation means points cluster closely around the regression line.

What if $r = 0$?

If:

$$r = 0,$$

then:

$$s_{y|x} \approx s_y \sqrt{1 - 0} = s_y.$$

What if $r = 0$?

If:

$$r = 0,$$

then:

$$s_{y|x} \approx s_y \sqrt{1 - 0} = s_y.$$

Knowing x gives essentially no reduction in uncertainty about y .

What if $r = 0$?

If:

$$r = 0,$$

then:

$$s_{y|x} \approx s_y \sqrt{1 - 0} = s_y.$$

Knowing x gives essentially no reduction in uncertainty about y .

The regression prediction is then no more informative than using the overall mean of y .

Example 4A: Estimate prediction spread

Suppose:

$$s_y = 12, \quad r = 0.60.$$

Example 4A: Estimate prediction spread

Suppose:

$$s_y = 12, \quad r = 0.60.$$

$$s_{y|x} \approx 12\sqrt{1 - (0.60)^2}.$$

Example 4A: Estimate prediction spread

Suppose:

$$s_y = 12, \quad r = 0.60.$$

$$s_{y|x} \approx 12\sqrt{1 - (0.60)^2}.$$

$$= 12\sqrt{1 - 0.36} = 12\sqrt{0.64}.$$

Example 4A: Estimate prediction spread

Suppose:

$$s_y = 12, \quad r = 0.60.$$

$$s_{y|x} \approx 12\sqrt{1 - (0.60)^2}.$$

$$= 12\sqrt{1 - 0.36} = 12\sqrt{0.64}.$$

$$= 12(0.8) = \boxed{9.6}.$$

Example 4B: Stronger correlation

Suppose:

$$s_y = 12, \quad r = 0.90.$$

Example 4B: Stronger correlation

Suppose:

$$s_y = 12, \quad r = 0.90.$$

$$s_{y|x} \approx 12\sqrt{1 - 0.81}.$$

Example 4B: Stronger correlation

Suppose:

$$s_y = 12, \quad r = 0.90.$$

$$s_{y|x} \approx 12\sqrt{1 - 0.81}.$$

$$= 12\sqrt{0.19}.$$

Example 4B: Stronger correlation

Suppose:

$$s_y = 12, \quad r = 0.90.$$

$$s_{y|x} \approx 12\sqrt{1 - 0.81}.$$

$$= 12\sqrt{0.19}.$$

$$\approx 12(0.436) \approx \boxed{5.23}.$$

Example 4B: Stronger correlation

Suppose:

$$s_y = 12, \quad r = 0.90.$$

$$s_{y|x} \approx 12\sqrt{1 - 0.81}.$$

$$= 12\sqrt{0.19}.$$

$$\approx 12(0.436) \approx \boxed{5.23}.$$

The stronger correlation produces a much smaller spread around the regression line.

Skill 5: The coefficient of determination

The quantity

$$r^2$$

is called the **coefficient of determination**.

Skill 5: The coefficient of determination

The quantity

$$r^2$$

is called the **coefficient of determination**.

It measures the fraction of the variation in y that is accounted for by the linear regression on x .

Skill 5: The coefficient of determination

The quantity

$$r^2$$

is called the **coefficient of determination**.

It measures the fraction of the variation in y that is accounted for by the linear regression on x .

Because:

$$-1 \leq r \leq 1,$$

we always have:

$$0 \leq r^2 \leq 1.$$

Translate r^2 into a percentage

If:

$$r = 0.70,$$

then:

$$r^2 = (0.70)^2 = 0.49.$$

Translate r^2 into a percentage

If:

$$r = 0.70,$$

then:

$$r^2 = (0.70)^2 = 0.49.$$

So:

49%

of the variation in y is accounted for by the linear regression.

Translate r^2 into a percentage

If:

$$r = 0.70,$$

then:

$$r^2 = (0.70)^2 = 0.49.$$

So:

49%

of the variation in y is accounted for by the linear regression.

The remaining:

$$100\% - 49\% = 51\%$$

is not accounted for by that linear model.

Example 5A: Find and interpret r^2

Suppose:

$$r = -0.80.$$

Example 5A: Find and interpret r^2

Suppose:

$$r = -0.80.$$

Square it:

$$r^2 = (-0.80)^2 = 0.64.$$

Example 5A: Find and interpret r^2

Suppose:

$$r = -0.80.$$

Square it:

$$r^2 = (-0.80)^2 = 0.64.$$

Interpretation:

64% of the variation in y is accounted for by the linear regression on x .

Example 5A: Find and interpret r^2

Suppose:

$$r = -0.80.$$

Square it:

$$r^2 = (-0.80)^2 = 0.64.$$

Interpretation:

64% of the variation in y is accounted for by the linear regression on x .

The negative sign of r tells direction.

The square r^2 tells explained variation.

Example 5B: Work backward

Suppose:

$$r^2 = 0.36$$

and the scatter plot has a negative slope.

Example 5B: Work backward

Suppose:

$$r^2 = 0.36$$

and the scatter plot has a negative slope.

Then:

$$|r| = \sqrt{0.36} = 0.60.$$

Example 5B: Work backward

Suppose:

$$r^2 = 0.36$$

and the scatter plot has a negative slope.

Then:

$$|r| = \sqrt{0.36} = 0.60.$$

Because the trend is negative:

$$r = -0.60.$$

Do not confuse r and r^2

If:

$$r = 0.50,$$

then:

$$r^2 = 0.25.$$

Do not confuse r and r^2

If:

$$r = 0.50,$$

then:

$$r^2 = 0.25.$$

So the regression accounts for about:

25%

of the variation, **not** 50%.

Do not confuse r and r^2

If:

$$r = 0.50,$$

then:

$$r^2 = 0.25.$$

So the regression accounts for about:

25%

of the variation, **not** 50%.

This distinction matters.

Skill 6: Stay inside the data range

Suppose the observed x -values range from:

$$10 \leq x \leq 50.$$

Skill 6: Stay inside the data range

Suppose the observed x -values range from:

$$10 \leq x \leq 50.$$

Using the regression line to predict at

$$x = 30$$

is prediction **within** the observed range.

Skill 6: Stay inside the data range

Suppose the observed x -values range from:

$$10 \leq x \leq 50.$$

Using the regression line to predict at

$$x = 30$$

is prediction **within** the observed range.

Using it at

$$x = 100$$

is called:

extrapolation.

Why extrapolation is risky

A regression line describes the pattern seen in the available data.

Why extrapolation is risky

A regression line describes the pattern seen in the available data.
Outside that range:

- ▶ the relationship may change;
- ▶ the trend may curve;
- ▶ the physical situation may no longer make sense.

Why extrapolation is risky

A regression line describes the pattern seen in the available data.

Outside that range:

- ▶ the relationship may change;
- ▶ the trend may curve;
- ▶ the physical situation may no longer make sense.

Therefore:

A good fit inside the data range does not guarantee a good prediction outside it.

Example 6A: Which prediction is safer?

Observed data:

$$5 \leq x \leq 25.$$

Regression line:

$$\hat{y} = 10 + 2x.$$

Predict at:

$$x = 20$$

or

$$x = 60.$$

Example 6A: Which prediction is safer?

Observed data:

$$5 \leq x \leq 25.$$

Regression line:

$$\hat{y} = 10 + 2x.$$

Predict at:

$$x = 20$$

or

$$x = 60.$$

The prediction at $x = 20$ is inside the observed range.

Example 6A: Which prediction is safer?

Observed data:

$$5 \leq x \leq 25.$$

Regression line:

$$\hat{y} = 10 + 2x.$$

Predict at:

$$x = 20$$

or

$$x = 60.$$

The prediction at $x = 20$ is inside the observed range.

The prediction at $x = 60$ is extrapolation.

Example 6A: Which prediction is safer?

Observed data:

$$5 \leq x \leq 25.$$

Regression line:

$$\hat{y} = 10 + 2x.$$

Predict at:

$$x = 20$$

or

$$x = 60.$$

The prediction at $x = 20$ is inside the observed range.

The prediction at $x = 60$ is extrapolation.

So the $x = 20$ prediction is much more defensible.

Example 6B: A physically impossible extrapolation

Suppose a regression model for a plant's height is:

$$\hat{y} = 4 + 3x,$$

where x is age in weeks, based on data from weeks 1 through 8.

Example 6B: A physically impossible extrapolation

Suppose a regression model for a plant's height is:

$$\hat{y} = 4 + 3x,$$

where x is age in weeks, based on data from weeks 1 through 8.

Using this line to predict the plant's height at

$$x = 200$$

weeks assumes the same growth pattern continues for years.

Example 6B: A physically impossible extrapolation

Suppose a regression model for a plant's height is:

$$\hat{y} = 4 + 3x,$$

where x is age in weeks, based on data from weeks 1 through 8.

Using this line to predict the plant's height at

$$x = 200$$

weeks assumes the same growth pattern continues for years.

That assumption is not supported by the original data.

Example 6B: A physically impossible extrapolation

Suppose a regression model for a plant's height is:

$$\hat{y} = 4 + 3x,$$

where x is age in weeks, based on data from weeks 1 through 8.

Using this line to predict the plant's height at

$$x = 200$$

weeks assumes the same growth pattern continues for years.

That assumption is not supported by the original data.

The arithmetic can be correct while the prediction is unreasonable.

Skill 7: Regression to the mean

Suppose two measurements are positively correlated, but not perfectly.

Skill 7: Regression to the mean

Suppose two measurements are positively correlated, but not perfectly. Individuals who are extremely high on the first measurement tend, on average, to be less extreme on the second.

Skill 7: Regression to the mean

Suppose two measurements are positively correlated, but not perfectly.

Individuals who are extremely high on the first measurement tend, on average, to be less extreme on the second.

Likewise, individuals who are extremely low on the first tend, on average, to be less extreme on the second.

Skill 7: Regression to the mean

Suppose two measurements are positively correlated, but not perfectly.

Individuals who are extremely high on the first measurement tend, on average, to be less extreme on the second.

Likewise, individuals who are extremely low on the first tend, on average, to be less extreme on the second.

This is:

regression to the mean.

Why does regression to the mean happen?

From the regression equation:

$$\hat{y} - \bar{y} = r \frac{s_y}{s_x} (x - \bar{x}).$$

Why does regression to the mean happen?

From the regression equation:

$$\hat{y} - \bar{y} = r \frac{s_y}{s_x} (x - \bar{x}).$$

Standardize both sides:

$$\boxed{\frac{\hat{y} - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x}}.$$

Why does regression to the mean happen?

From the regression equation:

$$\hat{y} - \bar{y} = r \frac{s_y}{s_x} (x - \bar{x}).$$

Standardize both sides:

$$\boxed{\frac{\hat{y} - \bar{y}}{s_y} = r \frac{x - \bar{x}}{s_x}}.$$

If

$$0 < r < 1,$$

the predicted y is fewer standard deviations from its mean than x was from its mean.

Example 7A: Standardized interpretation

Suppose:

$$r = 0.60.$$

A person's x -value is:

2

standard deviations above the x -mean.

Example 7A: Standardized interpretation

Suppose:

$$r = 0.60.$$

A person's x -value is:

$$2$$

standard deviations above the x -mean.

The predicted y -value is:

$$0.60(2) = 1.2$$

standard deviations above the y -mean.

Example 7A: Standardized interpretation

Suppose:

$$r = 0.60.$$

A person's x -value is:

$$2$$

standard deviations above the x -mean.

The predicted y -value is:

$$0.60(2) = 1.2$$

standard deviations above the y -mean.

So the prediction is still above average, but less extreme.

Example 7B: A concrete version

Suppose:

$$\bar{x} = 70, \quad s_x = 10,$$

$$\bar{y} = 72, \quad s_y = 8,$$

$$r = 0.50.$$

A student scores:

$$x = 90.$$

Example 7B: A concrete version

Suppose:

$$\bar{x} = 70, \quad s_x = 10,$$

$$\bar{y} = 72, \quad s_y = 8,$$

$$r = 0.50.$$

A student scores:

$$x = 90.$$

That is:

$$\frac{90 - 70}{10} = 2$$

standard deviations above the x -mean.

Example 7B: Predict the second score

The predicted standardized y -score is:

$$0.50(2) = 1$$

standard deviation above average.

Example 7B: Predict the second score

The predicted standardized y -score is:

$$0.50(2) = 1$$

standard deviation above average.

Therefore:

$$\hat{y} = 72 + 1(8).$$

Example 7B: Predict the second score

The predicted standardized y -score is:

$$0.50(2) = 1$$

standard deviation above average.

Therefore:

$$\hat{y} = 72 + 1(8).$$

$$\boxed{\hat{y} = 80}.$$

Example 7B: Predict the second score

The predicted standardized y -score is:

$$0.50(2) = 1$$

standard deviation above average.

Therefore:

$$\hat{y} = 72 + 1(8).$$

$$\boxed{\hat{y} = 80}.$$

The first score was extremely high relative to its distribution.

The predicted second score is still high, but less extreme.

The regression fallacy

Suppose subjects are selected because they have unusually extreme measurements.

The regression fallacy

Suppose subjects are selected because they have unusually extreme measurements. On a later measurement, some movement toward the mean is expected even if **nothing has changed**.

The regression fallacy

Suppose subjects are selected because they have unusually extreme measurements.

On a later measurement, some movement toward the mean is expected even if **nothing has changed**.

If a treatment occurs between the two measurements, it is easy to mistakenly credit the treatment for this movement.

The regression fallacy

Suppose subjects are selected because they have unusually extreme measurements. On a later measurement, some movement toward the mean is expected even if **nothing has changed**.

If a treatment occurs between the two measurements, it is easy to mistakenly credit the treatment for this movement.

That mistake is called the:

regression fallacy.

Example 7C: Blood pressure

Suppose patients are enrolled in a study because their blood pressure is unusually high on one particular day.

Example 7C: Blood pressure

Suppose patients are enrolled in a study because their blood pressure is unusually high on one particular day.

A second measurement is likely to be somewhat less extreme even without an effective treatment.

Example 7C: Blood pressure

Suppose patients are enrolled in a study because their blood pressure is unusually high on one particular day.

A second measurement is likely to be somewhat less extreme even without an effective treatment.

Therefore a before-and-after comparison alone may exaggerate treatment success.

Example 7C: Blood pressure

Suppose patients are enrolled in a study because their blood pressure is unusually high on one particular day.

A second measurement is likely to be somewhat less extreme even without an effective treatment.

Therefore a before-and-after comparison alone may exaggerate treatment success.

Regression to the mean must be considered.

Skill 8: Why transform data?

Linear regression works best when the scatter plot is approximately linear.

Skill 8: Why transform data?

Linear regression works best when the scatter plot is approximately linear. Sometimes the original scatter plot is strongly curved.

Skill 8: Why transform data?

Linear regression works best when the scatter plot is approximately linear.

Sometimes the original scatter plot is strongly curved.

A transformation can sometimes turn the curved pattern into a straight-line pattern.

Skill 8: Why transform data?

Linear regression works best when the scatter plot is approximately linear.

Sometimes the original scatter plot is strongly curved.

A transformation can sometimes turn the curved pattern into a straight-line pattern.

Then we can apply linear regression to the **transformed variables**.

Three transformations in the chapter

The chapter emphasizes:

① **log-log:**

$\log y$ versus $\log x$;

Three transformations in the chapter

The chapter emphasizes:

① **log-log:**

$\log y$ versus $\log x$;

② **semi-log:**

$\log y$ versus x ;

Three transformations in the chapter

The chapter emphasizes:

① **log-log:**

$\log y$ versus $\log x$;

② **semi-log:**

$\log y$ versus x ;

③ **reciprocal:**

y versus $\frac{1}{x}$.

Log-log transformation

If

$$\log y = a + b \log x,$$

then:

$$y = cx^b$$

for some positive constant c .

Log-log transformation

If

$$\log y = a + b \log x,$$

then:

$$y = cx^b$$

for some positive constant c .

So a straight line on a log-log plot corresponds to a **power law** in the original variables.

Log-log transformation

If

$$\log y = a + b \log x,$$

then:

$$y = cx^b$$

for some positive constant c .

So a straight line on a log-log plot corresponds to a **power law** in the original variables.

The slope of the log-log line is the power:

$$b.$$

Example 8A: Read a power law from a log-log line

Suppose:

$$\log_{10} y = 0.30 + 1.50 \log_{10} x.$$

Example 8A: Read a power law from a log-log line

Suppose:

$$\log_{10} y = 0.30 + 1.50 \log_{10} x.$$

Exponentiate:

$$y = 10^{0.30} x^{1.50}.$$

Example 8A: Read a power law from a log-log line

Suppose:

$$\log_{10} y = 0.30 + 1.50 \log_{10} x.$$

Exponentiate:

$$y = 10^{0.30} x^{1.50}.$$

Since:

$$10^{0.30} \approx 2,$$

we obtain:

$$\boxed{y \approx 2x^{1.5}}.$$

Example 8B: Start with a power law

Suppose:

$$y = 5x^2.$$

Example 8B: Start with a power law

Suppose:

$$y = 5x^2.$$

Take logarithms:

$$\log y = \log(5x^2).$$

Example 8B: Start with a power law

Suppose:

$$y = 5x^2.$$

Take logarithms:

$$\log y = \log(5x^2).$$

$$\log y = \log 5 + 2 \log x.$$

Example 8B: Start with a power law

Suppose:

$$y = 5x^2.$$

Take logarithms:

$$\log y = \log(5x^2).$$

$$\log y = \log 5 + 2 \log x.$$

Therefore a plot of

$\log y$ versus $\log x$

is a straight line with:

$$\boxed{\text{slope } 2}.$$

Semi-log transformation

If:

$$\log_{10} y = a + bx,$$

then:

$$y = 10^a 10^{bx}.$$

Semi-log transformation

If:

$$\log_{10} y = a + bx,$$

then:

$$y = 10^a 10^{bx}.$$

Writing:

$$c = 10^a,$$

we get:

$$\boxed{y = c 10^{bx}}.$$

Semi-log transformation

If:

$$\log_{10} y = a + bx,$$

then:

$$y = 10^a 10^{bx}.$$

Writing:

$$c = 10^a,$$

we get:

$$\boxed{y = c 10^{bx}}.$$

So a straight line on a semi-log plot corresponds to **exponential behavior**.

Example 8C: Semi-log model

Suppose:

$$\log_{10} y = 1 + 0.2x.$$

Example 8C: Semi-log model

Suppose:

$$\log_{10} y = 1 + 0.2x.$$

Exponentiate:

$$y = 10^{1+0.2x}.$$

Example 8C: Semi-log model

Suppose:

$$\log_{10} y = 1 + 0.2x.$$

Exponentiate:

$$y = 10^{1+0.2x}.$$

$$y = 10 \cdot 10^{0.2x}.$$

Example 8C: Semi-log model

Suppose:

$$\log_{10} y = 1 + 0.2x.$$

Exponentiate:

$$y = 10^{1+0.2x}.$$

$$y = 10 \cdot 10^{0.2x}.$$

Therefore:

$$y = 10(10^{0.2})^x.$$

Reciprocal transformation

Suppose a plot of

y versus $\frac{1}{x}$

is linear.

Reciprocal transformation

Suppose a plot of

y versus $\frac{1}{x}$

is linear.

Then:

$$y = a + b \left(\frac{1}{x} \right).$$

Reciprocal transformation

Suppose a plot of

y versus $\frac{1}{x}$

is linear.

Then:

$$y = a + b \left(\frac{1}{x} \right).$$

So the original relationship has the form:

$$y = a + \frac{b}{x}.$$

Example 8D: Reciprocal model

Suppose regression on the transformed variable gives:

$$\hat{y} = 4 + \frac{12}{x}.$$

Predict y when:

$$x = 3.$$

Example 8D: Reciprocal model

Suppose regression on the transformed variable gives:

$$\hat{y} = 4 + \frac{12}{x}.$$

Predict y when:

$$x = 3.$$

$$\hat{y} = 4 + \frac{12}{3}.$$

Example 8D: Reciprocal model

Suppose regression on the transformed variable gives:

$$\hat{y} = 4 + \frac{12}{x}.$$

Predict y when:

$$x = 3.$$

$$\hat{y} = 4 + \frac{12}{3}.$$

$$\boxed{\hat{y} = 8}.$$

Log transformations require positive values

To compute:

$$\log x$$

or

$$\log y,$$

the transformed variable must be positive.

Log transformations require positive values

To compute:

$$\log x$$

or

$$\log y,$$

the transformed variable must be positive.

So a standard log transformation cannot be directly applied to:

$$x \leq 0$$

or

$$y \leq 0.$$

Log transformations require positive values

To compute:

$$\log x$$

or

$$\log y,$$

the transformed variable must be positive.

So a standard log transformation cannot be directly applied to:

$$x \leq 0$$

or

$$y \leq 0.$$

Always check the domain before taking logarithms.

Skill 9: A high r does not guarantee a useful line

A few points far from the rest of the data can strongly control the regression line.

Skill 9: A high r does not guarantee a useful line

A few points far from the rest of the data can strongly control the regression line. Such points can be **high-leverage points**.

Skill 9: A high r does not guarantee a useful line

A few points far from the rest of the data can strongly control the regression line. Such points can be **high-leverage points**. They may produce:

$$|r| \approx 1$$

even when a straight-line model is poor for most of the data.

Skill 9: A high r does not guarantee a useful line

A few points far from the rest of the data can strongly control the regression line. Such points can be **high-leverage points**. They may produce:

$$|r| \approx 1$$

even when a straight-line model is poor for most of the data. Always inspect the scatter plot.

Example 9A: Why a picture matters

Suppose most points have:

$$0 < x < 2, \quad 0 < y < 3,$$

but one point is:

$$(100, 150).$$

Example 9A: Why a picture matters

Suppose most points have:

$$0 < x < 2, \quad 0 < y < 3,$$

but one point is:

$$(100, 150).$$

The distant point can dominate:

- ▶ the slope,
- ▶ the correlation,
- ▶ the apparent quality of the fit.

Example 9A: Why a picture matters

Suppose most points have:

$$0 < x < 2, \quad 0 < y < 3,$$

but one point is:

$$(100, 150).$$

The distant point can dominate:

- ▶ the slope,
- ▶ the correlation,
- ▶ the apparent quality of the fit.

A large r^2 should not replace visual inspection.

Residual thinking

A useful regression model should leave residuals that do not show an obvious pattern.

Residual thinking

A useful regression model should leave residuals that do not show an obvious pattern.

If residuals are:

$+, +, 0, -, -, 0, +, +$

in a clear curved pattern as x increases, then the line is missing structure.

Residual thinking

A useful regression model should leave residuals that do not show an obvious pattern.

If residuals are:

$+, +, 0, -, -, 0, +, +$

in a clear curved pattern as x increases, then the line is missing structure.

A curved residual pattern suggests:

the relationship may not be linear.

Example 9B: A bad linear model

Suppose the data follow approximately:

$$y = x^2.$$

Example 9B: A bad linear model

Suppose the data follow approximately:

$$y = x^2.$$

A straight regression line may have modest or even misleading summary statistics.

Example 9B: A bad linear model

Suppose the data follow approximately:

$$y = x^2.$$

A straight regression line may have modest or even misleading summary statistics.
But the scatter plot shows a curve.

Example 9B: A bad linear model

Suppose the data follow approximately:

$$y = x^2.$$

A straight regression line may have modest or even misleading summary statistics.

But the scatter plot shows a curve.

A transformation or nonlinear model is more appropriate.

Skill 10: How changing units affects a regression line

Suppose:

$$\hat{y} = \hat{a} + \hat{b}x.$$

Skill 10: How changing units affects a regression line

Suppose:

$$\hat{y} = \hat{a} + \hat{b}x.$$

Changing the units of x or y can change:

- ▶ the slope,
- ▶ the intercept.

Skill 10: How changing units affects a regression line

Suppose:

$$\hat{y} = \hat{a} + \hat{b}x.$$

Changing the units of x or y can change:

- ▶ the slope,
- ▶ the intercept.

But positive unit conversions do not change:

$$r$$

or:

$$r^2.$$

Example 10A: Change the units of y

Suppose:

$$\hat{y} = 20 + 4x,$$

where y is measured in centimeters.

Convert y to meters:

$$Y = \frac{y}{100}.$$

Example 10A: Change the units of y

Suppose:

$$\hat{y} = 20 + 4x,$$

where y is measured in centimeters.

Convert y to meters:

$$Y = \frac{y}{100}.$$

Then:

$$\hat{Y} = \frac{20 + 4x}{100}.$$

Example 10A: Change the units of y

Suppose:

$$\hat{y} = 20 + 4x,$$

where y is measured in centimeters.

Convert y to meters:

$$Y = \frac{y}{100}.$$

Then:

$$\hat{Y} = \frac{20 + 4x}{100}.$$

$$\boxed{\hat{Y} = 0.20 + 0.04x}.$$

Example 10A: Change the units of y

Suppose:

$$\hat{y} = 20 + 4x,$$

where y is measured in centimeters.

Convert y to meters:

$$Y = \frac{y}{100}.$$

Then:

$$\hat{Y} = \frac{20 + 4x}{100}.$$

$$\boxed{\hat{Y} = 0.20 + 0.04x}.$$

The numerical slope changes because the units changed.

Example 10B: Add a constant to y

Suppose:

$$\hat{y} = 5 + 2x.$$

Now every recorded y -value is increased by 10.

Example 10B: Add a constant to y

Suppose:

$$\hat{y} = 5 + 2x.$$

Now every recorded y -value is increased by 10.

The new regression predictions are also increased by 10:

$$\hat{Y} = 15 + 2x.$$

Example 10B: Add a constant to y

Suppose:

$$\hat{y} = 5 + 2x.$$

Now every recorded y -value is increased by 10.

The new regression predictions are also increased by 10:

$$\hat{Y} = 15 + 2x.$$

So:

slope unchanged,

intercept increases by 10.

Mixed Example 1

Given:

$$\bar{x} = 5, \quad s_x = 2,$$

$$\bar{y} = 20, \quad s_y = 4,$$

$$r = 0.75,$$

find the slope.

Mixed Example 1

Given:

$$\bar{x} = 5, \quad s_x = 2,$$

$$\bar{y} = 20, \quad s_y = 4,$$

$$r = 0.75,$$

find the slope.

$$\hat{b} = 0.75 \left(\frac{4}{2} \right).$$

Mixed Example 1

Given:

$$\bar{x} = 5, \quad s_x = 2,$$

$$\bar{y} = 20, \quad s_y = 4,$$

$$r = 0.75,$$

find the slope.

$$\hat{b} = 0.75 \left(\frac{4}{2} \right).$$

$$\boxed{\hat{b} = 1.5}.$$

Mixed Example 2

Continue with:

$$\bar{x} = 5, \quad \bar{y} = 20, \quad \hat{b} = 1.5.$$

Find the intercept.

Mixed Example 2

Continue with:

$$\bar{x} = 5, \quad \bar{y} = 20, \quad \hat{b} = 1.5.$$

Find the intercept.

$$\hat{a} = 20 - (1.5)(5).$$

Mixed Example 2

Continue with:

$$\bar{x} = 5, \quad \bar{y} = 20, \quad \hat{b} = 1.5.$$

Find the intercept.

$$\hat{a} = 20 - (1.5)(5).$$

$$\hat{a} = 20 - 7.5 = \boxed{12.5}.$$

Mixed Example 2

Continue with:

$$\bar{x} = 5, \quad \bar{y} = 20, \quad \hat{b} = 1.5.$$

Find the intercept.

$$\hat{a} = 20 - (1.5)(5).$$

$$\hat{a} = 20 - 7.5 = \boxed{12.5}.$$

Regression line:

$$\boxed{\hat{y} = 12.5 + 1.5x}.$$

Mixed Example 3

Use:

$$\hat{y} = 12.5 + 1.5x.$$

Predict when:

$$x = 8.$$

Mixed Example 3

Use:

$$\hat{y} = 12.5 + 1.5x.$$

Predict when:

$$x = 8.$$

$$\hat{y} = 12.5 + 1.5(8).$$

Mixed Example 3

Use:

$$\hat{y} = 12.5 + 1.5x.$$

Predict when:

$$x = 8.$$

$$\hat{y} = 12.5 + 1.5(8).$$

$$= 12.5 + 12 = \boxed{24.5}.$$

Mixed Example 4

The predicted value is:

$$\hat{y} = 24.5.$$

The actual value is:

$$y = 22.$$

Find the residual.

Mixed Example 4

The predicted value is:

$$\hat{y} = 24.5.$$

The actual value is:

$$y = 22.$$

Find the residual.

$$e = y - \hat{y}.$$

Mixed Example 4

The predicted value is:

$$\hat{y} = 24.5.$$

The actual value is:

$$y = 22.$$

Find the residual.

$$e = y - \hat{y}.$$

$$e = 22 - 24.5 = \boxed{-2.5}.$$

Mixed Example 4

The predicted value is:

$$\hat{y} = 24.5.$$

The actual value is:

$$y = 22.$$

Find the residual.

$$e = y - \hat{y}.$$

$$e = 22 - 24.5 = \boxed{-2.5}.$$

The point lies 2.5 units below the regression line.

Mixed Example 5

Suppose:

$$r = 0.75.$$

Find:

$$r^2.$$

Mixed Example 5

Suppose:

$$r = 0.75.$$

Find:

$$r^2.$$

$$r^2 = (0.75)^2 = 0.5625.$$

Mixed Example 5

Suppose:

$$r = 0.75.$$

Find:

$$r^2.$$

$$r^2 = (0.75)^2 = 0.5625.$$

Interpretation:

56.25% of the variation in y is accounted for by the linear regression.

Mixed Example 6

Suppose:

$$s_y = 10, \quad r = 0.75.$$

Estimate:

$$s_{y|x}.$$

Mixed Example 6

Suppose:

$$s_y = 10, \quad r = 0.75.$$

Estimate:

$$s_{y|x}.$$

$$s_{y|x} \approx 10\sqrt{1 - (0.75)^2}.$$

Mixed Example 6

Suppose:

$$s_y = 10, \quad r = 0.75.$$

Estimate:

$$s_{y|x}.$$

$$s_{y|x} \approx 10\sqrt{1 - (0.75)^2}.$$

$$= 10\sqrt{0.4375} \approx 10(0.6614).$$

Mixed Example 6

Suppose:

$$s_y = 10, \quad r = 0.75.$$

Estimate:

$$s_{y|x}.$$

$$s_{y|x} \approx 10\sqrt{1 - (0.75)^2}.$$

$$= 10\sqrt{0.4375} \approx 10(0.6614).$$

$$\boxed{s_{y|x} \approx 6.61}.$$

Mixed Example 7

The observed x -values run from:

2 to 15.

Which prediction is extrapolation?

$$x = 7, \quad x = 14, \quad x = 40.$$

Mixed Example 7

The observed x -values run from:

2 to 15.

Which prediction is extrapolation?

$x = 7,$ $x = 14,$ $x = 40.$

$x = 40.$

Mixed Example 7

The observed x -values run from:

2 to 15.

Which prediction is extrapolation?

$x = 7,$ $x = 14,$ $x = 40.$

$x = 40.$

It lies outside the observed range.

Mixed Example 8

A log-log regression gives:

$$\log y = 0.7 + 2.3 \log x.$$

What is the basic form of the original model?

Mixed Example 8

A log-log regression gives:

$$\log y = 0.7 + 2.3 \log x.$$

What is the basic form of the original model?

A log-log line gives a power law:

$$y = cx^b.$$

Mixed Example 8

A log-log regression gives:

$$\log y = 0.7 + 2.3 \log x.$$

What is the basic form of the original model?

A log-log line gives a power law:

$$y = cx^b.$$

The slope gives:

$$b = 2.3.$$

Mixed Example 8

A log-log regression gives:

$$\log y = 0.7 + 2.3 \log x.$$

What is the basic form of the original model?

A log-log line gives a power law:

$$y = cx^b.$$

The slope gives:

$$b = 2.3.$$

Therefore:

$$\boxed{y = cx^{2.3}}.$$

The chapter in one slide

Regression line:

$$\hat{y} = \hat{a} + \hat{b}x$$

$$\hat{b} = r \frac{s_y}{s_x}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Residual:

$$e = y - \hat{y}$$

Coefficient of determination:

$$r^2$$

Approximate spread around the line:

$$s_{y|x} \approx s_y \sqrt{1 - r^2}$$

The chapter in one slide

Regression line:

$$\hat{y} = \hat{a} + \hat{b}x$$

$$\hat{b} = r \frac{s_y}{s_x}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Residual:

$$e = y - \hat{y}$$

Coefficient of determination:

$$r^2$$

Approximate spread around the line:

$$s_{y|x} \approx s_y \sqrt{1 - r^2}$$

And:

Avoid extrapolation when possible.

Transformation summary

$$\log y \text{ vs. } \log x \text{ linear} \iff y = cx^b$$

Transformation summary

$$\log y \text{ vs. } \log x \text{ linear} \iff y = cx^b$$

$$\log y \text{ vs. } x \text{ linear} \iff y = c 10^{bx}$$

Transformation summary

$$\log y \text{ vs. } \log x \text{ linear} \iff y = cx^b$$

$$\log y \text{ vs. } x \text{ linear} \iff y = c 10^{bx}$$

$$y \text{ vs. } 1/x \text{ linear} \iff y = a + \frac{b}{x}$$

A practical regression workflow

Given a regression problem:

- 1 Identify which variable is x and which is y .

A practical regression workflow

Given a regression problem:

- 1 Identify which variable is x and which is y .
- 2 Inspect the scatter plot.

A practical regression workflow

Given a regression problem:

- ① Identify which variable is x and which is y .
- ② Inspect the scatter plot.
- ③ Confirm that a linear model is reasonable.

A practical regression workflow

Given a regression problem:

- 1 Identify which variable is x and which is y .
- 2 Inspect the scatter plot.
- 3 Confirm that a linear model is reasonable.
- 4 Find $\hat{b} = r(s_y/s_x)$.

A practical regression workflow

Given a regression problem:

- 1 Identify which variable is x and which is y .
- 2 Inspect the scatter plot.
- 3 Confirm that a linear model is reasonable.
- 4 Find $\hat{b} = r(s_y/s_x)$.
- 5 Find $\hat{a} = \bar{y} - \hat{b}\bar{x}$.

A practical regression workflow

Given a regression problem:

- 1 Identify which variable is x and which is y .
- 2 Inspect the scatter plot.
- 3 Confirm that a linear model is reasonable.
- 4 Find $\hat{b} = r(s_y/s_x)$.
- 5 Find $\hat{a} = \bar{y} - \hat{b}\bar{x}$.
- 6 Write $\hat{y} = \hat{a} + \hat{b}x$.

A practical regression workflow

Given a regression problem:

- 1 Identify which variable is x and which is y .
- 2 Inspect the scatter plot.
- 3 Confirm that a linear model is reasonable.
- 4 Find $\hat{b} = r(s_y/s_x)$.
- 5 Find $\hat{a} = \bar{y} - \hat{b}\bar{x}$.
- 6 Write $\hat{y} = \hat{a} + \hat{b}x$.
- 7 Substitute the requested x -value.

A practical regression workflow

Given a regression problem:

- 1 Identify which variable is x and which is y .
- 2 Inspect the scatter plot.
- 3 Confirm that a linear model is reasonable.
- 4 Find $\hat{b} = r(s_y/s_x)$.
- 5 Find $\hat{a} = \bar{y} - \hat{b}\bar{x}$.
- 6 Write $\hat{y} = \hat{a} + \hat{b}x$.
- 7 Substitute the requested x -value.
- 8 Check whether the prediction is extrapolation.

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?
- ▶ Did I use s_y/s_x , not s_x/s_y , in the slope?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?
- ▶ Did I use s_y/s_x , not s_x/s_y , in the slope?
- ▶ Does the line pass through $(\bar{x}, \bar{y})$?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?
- ▶ Did I use s_y/s_x , not s_x/s_y , in the slope?
- ▶ Does the line pass through $(\bar{x}, \bar{y})$?
- ▶ Did I distinguish $\hat{y}$ from the actual value y ?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?
- ▶ Did I use s_y/s_x , not s_x/s_y , in the slope?
- ▶ Does the line pass through $(\bar{x}, \bar{y})$?
- ▶ Did I distinguish $\hat{y}$ from the actual value y ?
- ▶ Did I compute residual as $y - \hat{y}$?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?
- ▶ Did I use s_y/s_x , not s_x/s_y , in the slope?
- ▶ Does the line pass through $(\bar{x}, \bar{y})$?
- ▶ Did I distinguish $\hat{y}$ from the actual value y ?
- ▶ Did I compute residual as $y - \hat{y}$?
- ▶ Did I square r before interpreting explained variation?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?
- ▶ Did I use s_y/s_x , not s_x/s_y , in the slope?
- ▶ Does the line pass through $(\bar{x}, \bar{y})$?
- ▶ Did I distinguish $\hat{y}$ from the actual value y ?
- ▶ Did I compute residual as $y - \hat{y}$?
- ▶ Did I square r before interpreting explained variation?
- ▶ Did I check for extrapolation?

Common errors checklist

Before submitting an answer, check:

- ▶ Did I regress the correct variable on the correct variable?
- ▶ Did I use s_y/s_x , not s_x/s_y , in the slope?
- ▶ Does the line pass through $(\bar{x}, \bar{y})$?
- ▶ Did I distinguish $\hat{y}$ from the actual value y ?
- ▶ Did I compute residual as $y - \hat{y}$?
- ▶ Did I square r before interpreting explained variation?
- ▶ Did I check for extrapolation?
- ▶ Did I inspect the scatter plot before trusting the regression?

Suppose:

$$\bar{x} = 40, \quad s_x = 5,$$

$$\bar{y} = 70, \quad s_y = 10,$$

$$r = 0.60.$$

Find the regression line and predict y when $x = 45$.

Final practice: Step 1

Slope:

$$\hat{b} = 0.60 \left(\frac{10}{5} \right) .$$

Final practice: Step 1

Slope:

$$\hat{b} = 0.60 \left(\frac{10}{5} \right) .$$

$$\hat{b} = 0.60(2) = \boxed{1.2} .$$

Final practice: Step 1

Slope:

$$\hat{b} = 0.60 \left(\frac{10}{5} \right).$$

$$\hat{b} = 0.60(2) = \boxed{1.2}.$$

Intercept:

$$\hat{a} = 70 - (1.2)(40).$$

Final practice: Step 1

Slope:

$$\hat{b} = 0.60 \left(\frac{10}{5} \right).$$

$$\hat{b} = 0.60(2) = \boxed{1.2}.$$

Intercept:

$$\hat{a} = 70 - (1.2)(40).$$

$$\hat{a} = 70 - 48 = \boxed{22}.$$

Final practice: Step 2

Regression line:

$$\hat{y} = 22 + 1.2x.$$

Final practice: Step 2

Regression line:

$$\hat{y} = 22 + 1.2x.$$

At:

$$x = 45,$$

we get:

$$\hat{y} = 22 + 1.2(45).$$

Final practice: Step 2

Regression line:

$$\hat{y} = 22 + 1.2x.$$

At:

$$x = 45,$$

we get:

$$\hat{y} = 22 + 1.2(45).$$

$$\hat{y} = 22 + 54.$$

Final practice: Step 2

Regression line:

$$\hat{y} = 22 + 1.2x.$$

At:

$$x = 45,$$

we get:

$$\hat{y} = 22 + 1.2(45).$$

$$\hat{y} = 22 + 54.$$

$$\hat{y} = 76.$$

Final practice: Step 2

Regression line:

$$\hat{y} = 22 + 1.2x.$$

At:

$$x = 45,$$

we get:

$$\hat{y} = 22 + 1.2(45).$$

$$\hat{y} = 22 + 54.$$

$$\hat{y} = 76.$$

Since $r^2 = 0.36$, the linear regression accounts for:

$$36\%$$

of the variation in y .