0.1 Learning Objectives

By the end of this lecture, you will be able to:

- Explain what the regression function $E[Y|X = t]$ represents
- Derive the least squares estimators for simple linear regression
- Interpret slope and intercept in context
- Compute a regression line in R and Python

1. Motivation

1.1 A Question

You survey 8 students before the midterm:

Student	1	2	3	4	5	6	7	8
Hours studied	2	3	4	5	5	7	8	10
Exam score	58	65	71	73	68	82	88	91

1.1 A Question

You survey 8 students before the midterm:

Student	1	2	3	4	5	6	7	8
Hours studied	2	3	4	5	5	7	8	10
Exam score	58	65	71	73	68	82	88	91

If a student studies for **6 hours**, what score would you predict?

1.1 A Question

You survey 8 students before the midterm:

Student	1	2	3	4	5	6	7	8
Hours studied	2	3	4	5	5	7	8	10
Exam score	58	65	71	73	68	82	88	91

If a student studies for **6 hours**, what score would you predict?

You'd probably look at the scatter plot and try to draw a line through the data…

1.2 Drawing a Line

Many lines *could* fit this data. Which one is “best”?

1.2 Drawing a Line

Many lines *could* fit this data. Which one is “best”?

Some options:

- Connect the first and last points?
- Eyeball it?
- Minimize the total distance from points to the line?

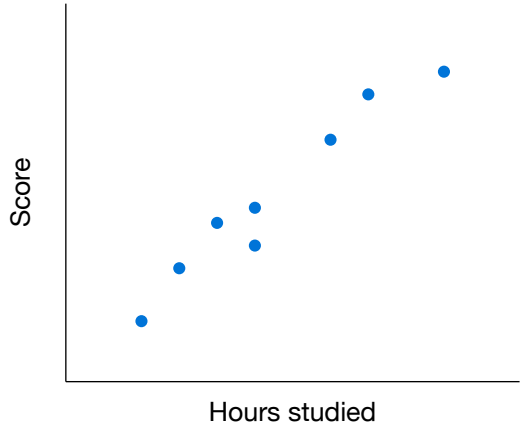
1.2 Drawing a Line

Many lines *could* fit this data. Which one is “best”?

Some options:

- Connect the first and last points?
- Eyeball it?
- Minimize the total distance from points to the line?

We need a **principled method** for choosing the best line.



1.3 What Does “Best” Mean?

1.3 What Does “Best” Mean?

For each data point (x_i, y_i), the **residual** is the vertical distance from the point to the line:

$$e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

1.3 What Does "Best" Mean?

For each data point (x_i, y_i), the **residual** is the vertical distance from the point to the line:

$$e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

Note: The **best line** minimizes the sum of **squared** residuals:

$$\text{minimize} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2$$

This is called **ordinary least squares** (OLS).

1.3 What Does "Best" Mean?

For each data point (x_i, y_i), the **residual** is the vertical distance from the point to the line:

$$e_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)$$

Note: The **best line** minimizes the sum of **squared** residuals:

$$\text{minimize} \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2$$

This is called **ordinary least squares** (OLS).

Why squared? Squaring ensures positive and negative errors don't cancel. It also penalizes large errors more than small ones.

2. The Regression Function

2.1 What Are We Estimating?

Consider a population of students. For every possible value of hours studied x, there is a *distribution* of exam scores.

2.1 What Are We Estimating?

Consider a population of students. For every possible value of hours studied x, there is a *distribution* of exam scores.

The **regression function** is the mean of Y for each value of X:

Definition: Regression Function

$$m(t) = E[Y \mid X = t]$$

The expected value of Y given that $X = t$.

2.1 What Are We Estimating?

Consider a population of students. For every possible value of hours studied x, there is a *distribution* of exam scores.

The **regression function** is the mean of Y for each value of X:

Definition: Regression Function

$$m(t) = E[Y \mid X = t]$$

The expected value of Y given that $X = t$.

We're modeling how the **average** outcome changes with X

2.1 What Are We Estimating?

not predicting individual outcomes exactly.

2.2 Why the Word “Regression”?

Francis Galton (1886) studied heights of parents and children.

2.2 Why the Word “Regression”?

Francis Galton (1886) studied heights of parents and children.

He found: very tall parents tend to have children who are tall, but **not as tall** as the parents. Very short parents have children who are short, but **not as short**.

2.2 Why the Word “Regression”?

Francis Galton (1886) studied heights of parents and children.

He found: very tall parents tend to have children who are tall, but **not as tall** as the parents. Very short parents have children who are short, but **not as short**.

Heights “regress toward the mean”

2.2 Why the Word “Regression”?

and the name stuck.

2.2 Why the Word “Regression”?

and the name stuck.

Note: In modern usage, “regression” simply means modeling $E[Y|X]$ as a function of X. The original meaning about “regression to the mean” is a separate concept.

3. Simple Linear Regression

3.1 The Model

We assume the regression function is **linear**:

$$m(t) = \beta_0 + \beta_1 t$$

3.1 The Model

We assume the regression function is **linear**:

$$m(t) = \beta_0 + \beta_1 t$$

For each observation:

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

3.1 The Model

We assume the regression function is **linear**:

$$m(t) = \beta_0 + \beta_1 t$$

For each observation:

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

- β_0 = **intercept** (value of Y when $X = 0$)
- β_1 = **slope** (change in Y per unit change in X)
- ε_i = **error** (random noise, things we can't predict)

Assumptions about ε_i:

- $E[\varepsilon_i] = 0$
- $\mathrm{Var}(\varepsilon_i) = \sigma^2$ (constant)
- Errors are independent

3.2 Parameters vs. Estimates

β_0 and β_1 are the **true population parameters**. We never observe them directly.

3.2 Parameters vs. Estimates

β_0 and β_1 are the **true population parameters**. We never observe them directly.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are our **estimates** from sample data.

3.2 Parameters vs. Estimates

β_0 and β_1 are the **true population parameters**. We never observe them directly.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are our **estimates** from sample data.

Recall: This is the same distinction as μ vs. $\overline{x}$, or σ vs. s. We use data to estimate population quantities.

4. Finding the Best Line

4.1 The Least Squares Objective

We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$L\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2$$

4.1 The Least Squares Objective

We want to find $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize:

$$L\left(\hat{\beta}_0, \hat{\beta}_1\right) = \sum_{i=1}^{n} \left[y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_i\right)\right]^2$$

This is a calculus optimization problem. Take partial derivatives, set them to zero.

4.2 Deriving $\hat{\beta}_1$

Take $\partial L/\partial \hat{\beta}_1 = 0$ and $\partial L/\partial \hat{\beta}_0 = 0$:

4.2 Deriving $\hat{\beta}_1$

Take $\partial L/\partial \hat{\beta}_1 = 0$ and $\partial L/\partial \hat{\beta}_0 = 0$:

From $\partial L/\partial \hat{\beta}_0 = 0$:

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

4.2 Deriving $\hat{\beta}_1$

Take $\partial L/\partial \hat{\beta}_1 = 0$ and $\partial L/\partial \hat{\beta}_0 = 0$:

From $\partial L/\partial \hat{\beta}_0 = 0$:

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

This says: the regression line always passes through the point $(\overline{x}, \overline{y})$.

4.3 Deriving $\hat{\beta}_1$ (continued)

Substituting back and solving:

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

4.3 Deriving $\hat{\beta}_1$ (continued)

Substituting back and solving:

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

Recognize anything?

4.3 Deriving $\hat{\beta}_1$ (continued)

Substituting back and solving:

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

Recognize anything?

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

4.3 Deriving $\hat{\beta}_1$ (continued)

Substituting back and solving:

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

Recognize anything?

$$\hat{\beta}_1 = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Var}(X)}$$

Note: The slope is the covariance of X and Y divided by the variance of X. If X and Y tend to increase together (positive covariance), the slope is positive.

4.4 The Least Squares Formulas

Slope:

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

Intercept:

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

4.4 The Least Squares Formulas

Slope:

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

Intercept:

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

Once we have $\hat{\beta}_0$ and $\hat{\beta}_1$, our **predicted values** are:

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

and the **residuals** are:

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

5. Worked Example

5.1 Study Hours Data

Student	1	2	3	4	5	6	7	8
x_i **(hours)**	2	3	4	5	5	7	8	10
y_i **(score)**	58	65	71	73	68	82	88	91

5.1 Study Hours Data

Student	1	2	3	4	5	6	7	8
x_i **(hours)**	2	3	4	5	5	7	8	10
y_i **(score)**	58	65	71	73	68	82	88	91

First, compute the means: $\overline{x} = 44/8 = 5.5, \quad \overline{y} = 596/8 = 74.5$

5.2 Computing the Slope

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

5.2 Computing the Slope

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

Numerator: $\sum (x_i - 5.5)(y_i - 74.5)$

$= (2 - 5.5)(58 - 74.5) + \cdots + (10 - 5.5)(91 - 74.5)$

5.2 Computing the Slope

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

Numerator: $\sum(x_i - 5.5)(y_i - 74.5)$

$$= (2 - 5.5)(58 - 74.5) + \cdots + (10 - 5.5)(91 - 74.5)$$

$$= 57.75 + 23.75 + 5.25 + 0.75 + 3.25 + 11.25 + 33.75 + 74.25$$

5.2 Computing the Slope

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

Numerator: $\sum(x_i - 5.5)(y_i - 74.5)$

$$= (2 - 5.5)(58 - 74.5) + \cdots + (10 - 5.5)(91 - 74.5)$$

$$= 57.75 + 23.75 + 5.25 + 0.75 + 3.25 + 11.25 + 33.75 + 74.25$$

$$= 210$$

5.2 Computing the Slope

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

Numerator: $\sum(x_i - 5.5)(y_i - 74.5)$

$$= (2 - 5.5)(58 - 74.5) + \cdots + (10 - 5.5)(91 - 74.5)$$

$$= 57.75 + 23.75 + 5.25 + 0.75 + 3.25 + 11.25 + 33.75 + 74.25$$

$$= 210$$

Denominator: $\sum (x_i - 5.5)^2 = 50$

5.2 Computing the Slope

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

Numerator: $\sum(x_i - 5.5)(y_i - 74.5)$

$= (2 - 5.5)(58 - 74.5) + \cdots + (10 - 5.5)(91 - 74.5)$

$= 57.75 + 23.75 + 5.25 + 0.75 + 3.25 + 11.25 + 33.75 + 74.25$

$= 210$

Denominator: $\sum(x_i - 5.5)^2 = 50$

$$\hat{\beta}_1 = \frac{210}{50} = 4.2$$

i	x_i	y_i
1	2	58
2	3	65
3	4	71
4	5	73
5	5	68
6	7	82
7	8	88
8	10	91

$\overline{x} = 5.5\ \overline{y} = 74.5$

5.3 Computing the Intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 74.5 - 4.2 \times 5.5 = 74.5 - 23.1 = 51.4$$

5.3 Computing the Intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 74.5 - 4.2 \times 5.5 = 74.5 - 23.1 = 51.4$$

The fitted regression line is:

$$\hat{y} = 51.4 + 4.2x$$

5.3 Computing the Intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 74.5 - 4.2 \times 5.5 = 74.5 - 23.1 = 51.4$$

The fitted regression line is:

$$\hat{y} = 51.4 + 4.2x$$

Interpretation:

- **Slope:** Each additional hour of studying is associated with a 4.2-point increase in exam score, on average
- **Intercept:** A student who studies 0 hours would score about 51.4 (extrapolation — use cautiously!)

5.3 Computing the Intercept

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} = 74.5 - 4.2 \times 5.5 = 74.5 - 23.1 = 51.4$$

The fitted regression line is:

$$\hat{y} = 51.4 + 4.2x$$

Interpretation:

- **Slope:** Each additional hour of studying is associated with a 4.2-point increase in exam score, on average
- **Intercept:** A student who studies 0 hours would score about 51.4 (extrapolation — use cautiously!)

Prediction: For a student who studies 6 hours: $\hat{y} = 51.4 + 4.2(6) = 76.6$

5.4 Computing Regression in R and Python

Vim

```r
x <- c(2,3,4,5,5,7,8,10)
y <- c(58,65,71,73,68,82,88,91)
model <- lm(y ~ x)
summary(model)
# Coefficients:
#   (Intercept)  51.40
#   x             4.20
```

Python

```python
import numpy as np
from scipy import stats
x = [2,3,4,5,5,7,8,10]
y = [58,65,71,73,68,82,88,91]
result = stats.linregress(x, y)
print(f"slope:
{result.slope:.2f}")
print(f"intercept:
{result.intercept:.2f}")
# slope: 4.20, intercept: 51.40
```

5.4 Computing Regression in R and Python

Vim

```r
x <- c(2,3,4,5,5,7,8,10)
y <- c(58,65,71,73,68,82,88,91)
model <- lm(y ~ x)
summary(model)
# Coefficients:
#   (Intercept)  51.40
#   x             4.20
```

Python

```python
import numpy as np
from scipy import stats
x = [2,3,4,5,5,7,8,10]
y = [58,65,71,73,68,82,88,91]
result = stats.linregress(x, y)
print(f"slope:
{result.slope:.2f}")
print(f"intercept:
{result.intercept:.2f}")
# slope: 4.20, intercept: 51.40
```

In practice, you won't compute regression by hand

5.4 Computing Regression in R and Python

but understanding the formula helps you interpret the output.

6. Slope and Correlation

6.1 Relationship Between $\hat{\beta}_1$ and r

The **sample correlation coefficient** r measures the strength of the linear relationship:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} \sqrt{\sum (y_i - \overline{y})^2}}$$

6.1 Relationship Between $\hat{\beta}_1$ and r

6.1 Relationship Between $\hat{\beta}_1$ and r

The **sample correlation coefficient** r measures the strength of the linear relationship:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} \sqrt{\sum (y_i - \overline{y})^2}}$$

Compare to the slope:

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

6.1 Relationship Between $\hat{\beta}_1$ and r

6.1 Relationship Between $\hat{\beta}_1$ and r

The **sample correlation coefficient** r measures the strength of the linear relationship:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} \sqrt{\sum (y_i - \overline{y})^2}}$$

Compare to the slope:

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

They are related:

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}$$

6.1 Relationship Between $\hat{\beta}_1$ and r

6.1 Relationship Between $\hat{\beta}_1$ and r

The **sample correlation coefficient** r measures the strength of the linear relationship:

$$r = \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum (x_i - \overline{x})^2} \sqrt{\sum (y_i - \overline{y})^2}}$$

Compare to the slope:

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

They are related:

$$\hat{\beta}_1 = r \cdot \frac{s_y}{s_x}$$

6.1 Relationship Between $\hat{\beta}_1$ and r

Note: r tells you the **strength and direction** of the relationship (-1 to $+1$). The slope $\hat{\beta}_1$ also encodes the **scale** — how many units Y changes per unit of X.

7. Exercise

7.1 Try It: Server Response Time

A web developer measures response time (ms) vs. number of concurrent users:

Users (x)	10	25	40	55	70	85
Response (ms) (y)	120	155	210	260	295	350

Compute the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

7.1 Try It: Server Response Time

A web developer measures response time (ms) vs. number of concurrent users:

Users (x)	10	25	40	55	70	85
Response (ms) (y)	120	155	210	260	295	350

Compute the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Think (2 minutes): Compute $\overline{x}$, $\overline{y}$, then the slope and intercept.

7.1 Try It: Server Response Time

A web developer measures response time (ms) vs. number of concurrent users:

Users (x)	10	25	40	55	70	85
Response (ms) (y)	120	155	210	260	295	350

Compute the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Think (2 minutes): Compute $\overline{x}$, $\overline{y}$, then the slope and intercept.

Discuss with your neighbor (1 minute): What does the slope mean in this context?

7.1 Try It: Server Response Time

A web developer measures response time (ms) vs. number of concurrent users:

Users (x)	10	25	40	55	70	85
Response (ms) (y)	120	155	210	260	295	350

Compute the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Think (2 minutes): Compute $\overline{x}$, $\overline{y}$, then the slope and intercept.

Discuss with your neighbor (1 minute): What does the slope mean in this context?

Try it yourself

Talk to your neighbor and try to solve this problem.

7.2 Solution

$$\overline{x} = 285/6 = 47.5, \quad \overline{y} = 1390/6 \approx 231.7$$

7.2 Solution

$$\overline{x} = 285/6 = 47.5, \quad \overline{y} = 1390/6 \approx 231.7$$

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = 25437.5$$

$$\sum (x_i - \overline{x})^2 = 3437.5$$

7.2 Solution

$$\overline{x} = 285/6 = 47.5, \quad \overline{y} = 1390/6 \approx 231.7$$

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = 25437.5$$

$$\sum (x_i - \overline{x})^2 = 3437.5$$

$$\hat{\beta}_1 = \frac{25437.5}{3437.5} \approx 3.0$$

$$\hat{\beta}_0 = 231.7 - 3.0 \times 47.5 = 231.7 - 142.5 = 89.2$$

7.2 Solution

$$\overline{x} = 285/6 = 47.5, \quad \overline{y} = 1390/6 \approx 231.7$$

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = 25437.5$$

$$\sum (x_i - \overline{x})^2 = 3437.5$$

$$\hat{\beta}_1 = \frac{25437.5}{3437.5} \approx 3.0$$

$$\hat{\beta}_0 = 231.7 - 3.0 \times 47.5 = 231.7 - 142.5 = 89.2$$

$$\hat{y} = 89.2 + 3.0x$$

7.2 Solution

$$\overline{x} = 285/6 = 47.5, \quad \overline{y} = 1390/6 \approx 231.7$$

$$\sum (x_i - \overline{x})(y_i - \overline{y}) = 25437.5$$

$$\sum (x_i - \overline{x})^2 = 3437.5$$

$$\hat{\beta}_1 = \frac{25437.5}{3437.5} \approx 3.0$$

$$\hat{\beta}_0 = 231.7 - 3.0 \times 47.5 = 231.7 - 142.5 = 89.2$$

$$\hat{y} = 89.2 + 3.0x$$

Interpretation: Each additional concurrent user adds about **3 milliseconds** to the response time, on average.

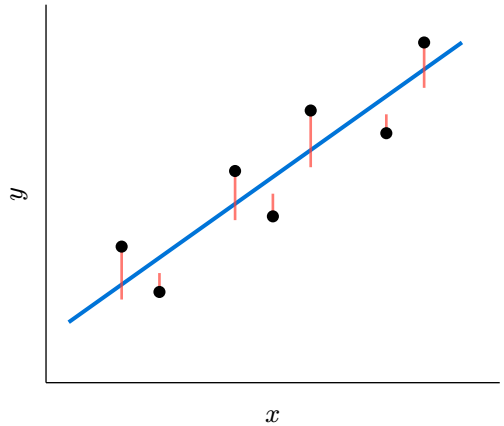
8. Visualizing Residuals

8.1 What Do Residuals Look Like?

The residual $e_i = y_i - \hat{y}_i$ is the **vertical distance** from each data point to the regression line.

8.1 What Do Residuals Look Like?

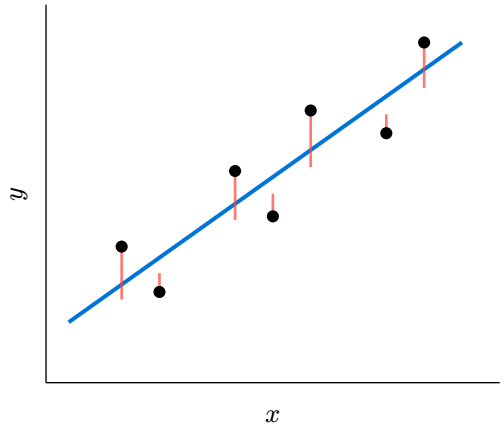
The residual $e_i = y_i - \hat{y}_i$ is the **vertical distance** from each data point to the regression line.



Each red line is a residual.

8.1 What Do Residuals Look Like?

The residual $e_i = y_i - \hat{y}_i$ is the **vertical distance** from each data point to the regression line.

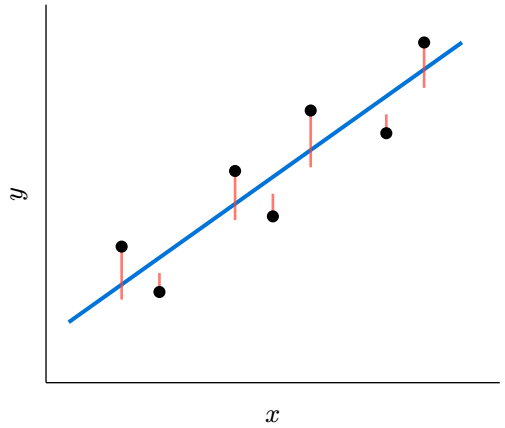


Each red line is a residual.

- Positive residual: point is **above** the line
- Negative residual: point is **below** the line
- Least squares minimizes $\sum e_i^2$

8.1 What Do Residuals Look Like?

The residual $e_i = y_i - \hat{y}_i$ is the **vertical distance** from each data point to the regression line.



Each red line is a residual.

- Positive residual: point is **above** the line
- Negative residual: point is **below** the line
- Least squares minimizes $\sum e_i^2$

The residuals always sum to zero: $\sum e_i = 0$.

8.2 Why Squared Residuals?

Why minimize the **sum of squared** residuals instead of, say, the sum of absolute values?

8.2 Why Squared Residuals?

Why minimize the **sum of squared** residuals instead of, say, the sum of absolute values?

Sum of absolute residuals $\sum |e_i|$:

- Treats all errors equally
- Less sensitive to outliers
- No closed-form solution

Sum of squared residuals $\sum e_i^2$:

- Penalizes large errors more heavily
- Has a clean, closed-form solution
- Connects to variance and conditional expectation

8.2 Why Squared Residuals?

Why minimize the **sum of squared** residuals instead of, say, the sum of absolute values?

Sum of absolute residuals $\sum|e_i|$:

- Treats all errors equally
- Less sensitive to outliers
- No closed-form solution

Sum of squared residuals $\sum e_i^2$:

- Penalizes large errors more heavily
- Has a clean, closed-form solution
- Connects to variance and conditional expectation

Note: From probability theory (Bertsekas Ch. 4.6): the function $g(X)$ that minimizes $E[(Y - g(X))^2]$ is exactly $g(X) = E[Y|X]$ — the conditional expectation. Least squares is the natural estimator of the regression function.

8.3 Interactive Demo

Try it yourself:

PhET: Least-Squares Regression

8.3 Interactive Demo

drag data points and watch the best-fit line, residuals, and R^2 update in real time.

8.3 Interactive Demo

drag data points and watch the best-fit line, residuals, and R^2 update in real time.

Observable: Interactive Linear Regression

8.3 Interactive Demo

toggle between viewing **absolute error** (vertical lines) and **squared error** (colored squares) to see why squaring matters.

8.3 Interactive Demo

toggle between viewing **absolute error** (vertical lines) and **squared error** (colored squares) to see why squaring matters.

Things to try:

- What happens when you add an **outlier** far from the line?
- Where does an outlier have the most influence — near $\overline{x}$ or far from $\overline{x}$?
- When does R^2 approach 1? When does it approach 0?

9. Always Plot Your Data

9.1 Anscombe's Quartet

Four datasets with **identical** summary statistics:

	Dataset I	Dataset II	Dataset III	Dataset IV
$\overline{x}$	9.0	9.0	9.0	9.0
$\overline{y}$	7.50	7.50	7.50	7.50
$\hat{\beta}_1$	0.500	0.500	0.500	0.500
$\hat{\beta}_0$	3.00	3.00	3.00	3.00
R^2	0.67	0.67	0.67	0.67

9.1 Anscombe's Quartet

Four datasets with **identical** summary statistics:

	Dataset I	Dataset II	Dataset III	Dataset IV
$\overline{x}$	9.0	9.0	9.0	9.0
$\overline{y}$	7.50	7.50	7.50	7.50
$\hat{\beta}_1$	0.500	0.500	0.500	0.500
$\hat{\beta}_0$	3.00	3.00	3.00	3.00
R^2	0.67	0.67	0.67	0.67

Same means, same slope, same intercept, same R^2.

9.1 Anscombe's Quartet

Four datasets with **identical** summary statistics:

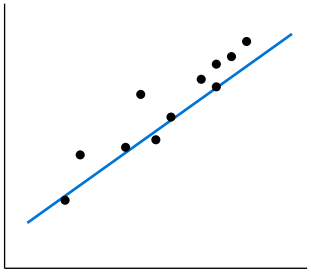
	Dataset I	Dataset II	Dataset III	Dataset IV
$\overline{x}$	9.0	9.0	9.0	9.0
$\overline{y}$	7.50	7.50	7.50	7.50
$\hat{\beta}_1$	0.500	0.500	0.500	0.500
$\hat{\beta}_0$	3.00	3.00	3.00	3.00
R^2	0.67	0.67	0.67	0.67

Same means, same slope, same intercept, same R^2.

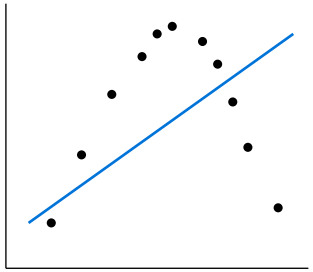
Are these datasets the same?

9.2 Anscombe's Quartet: The Scatter Plots

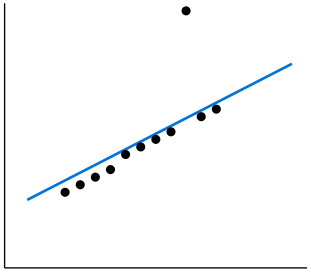
I. Linear relationship (regression is appropriate)



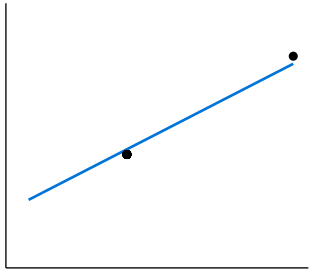
II. Curved relationship (need a nonlinear model)



III. Outlier inflates residuals (one extreme point)



IV. Leverage point drives the line (one influential point)



9.3 The Lesson

Summary statistics and regression coefficients are not enough.

9.3 The Lesson

Summary statistics and regression coefficients are not enough.

Always visualize your data before fitting a model.

9.3 The Lesson

Summary statistics and regression coefficients are not enough.

Always visualize your data before fitting a model.

Note:

- Dataset II has a curved relationship — a linear model is wrong
- Dataset III has one outlier pulling the line — the fit is misleading
- Dataset IV has one **leverage point** that single-handedly determines the slope

9.3 The Lesson

Summary statistics and regression coefficients are not enough.

Always visualize your data before fitting a model.

Note:

- Dataset II has a curved relationship — a linear model is wrong
- Dataset III has one outlier pulling the line — the fit is misleading
- Dataset IV has one **leverage point** that single-handedly determines the slope

This is why we'll learn **residual diagnostics** next

9.3 The Lesson

tools for checking whether a linear model is appropriate.

10. How Good Is Our Model?

10.1 Partitioning Variance

For each data point, the total deviation from $\overline{y}$ can be split:

$$\underbrace{y_i - \overline{y}}_{\text{total}} = \underbrace{\hat{y}_i - \overline{y}}_{\text{explained}} + \underbrace{y_i - \hat{y}_i}_{\text{residual}}$$

10.1 Partitioning Variance

For each data point, the total deviation from $\overline{y}$ can be split:

$$\underbrace{y_i - \overline{y}}_{\text{total}} = \underbrace{\hat{y}_i - \overline{y}}_{\text{explained}} + \underbrace{y_i - \hat{y}_i}_{\text{residual}}$$

Squaring and summing over all data points:

$$\underbrace{\sum (y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum (\hat{y}_i - \overline{y})^2}_{\text{SSR}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{SSE}}$$

10.1 Partitioning Variance

For each data point, the total deviation from $\overline{y}$ can be split:

$$\underbrace{y_i - \overline{y}}_{\text{total}} = \underbrace{\hat{y}_i - \overline{y}}_{\text{explained}} + \underbrace{y_i - \hat{y}_i}_{\text{residual}}$$

Squaring and summing over all data points:

$$\underbrace{\sum (y_i - \overline{y})^2}_{\text{SST}} = \underbrace{\sum (\hat{y}_i - \overline{y})^2}_{\text{SSR}} + \underbrace{\sum (y_i - \hat{y}_i)^2}_{\text{SSE}}$$

SST = total variation in Y

SSR = variation explained by the regression

SSE = unexplained variation (residuals)

10.2 R^2: The Coefficient of Determination

Definition: R-squared

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

The proportion of variance in Y explained by the regression on X.

10.2 R^2: The Coefficient of Determination

10.2 R^2: The Coefficient of Determination

Definition: R-squared

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

The proportion of variance in Y explained by the regression on X.

- $R^2 = 1$: perfect fit (all points on the line)
- $R^2 = 0$: the line explains nothing (slope is zero)
- Typical values depend on the domain

10.2 R^2: The Coefficient of Determination

10.2 R^2: The Coefficient of Determination

Definition: R-squared

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

The proportion of variance in Y explained by the regression on X.

- $R^2 = 1$: perfect fit (all points on the line)
- $R^2 = 0$: the line explains nothing (slope is zero)
- Typical values depend on the domain

10.2 R^2: The Coefficient of Determination

Example: Study Hours

Our study hours regression had $R^2 \approx 0.95$ — studying explains most of the variation in scores.

The baseball height-weight regression (Matloff) had $R^2 = 0.28$ — height explains only 28% of weight variation.

10.3 Connection to Correlation

R^2 is literally the **square of the correlation coefficient**:

$$R^2 = r^2$$

10.3 Connection to Correlation

R^2 is literally the **square of the correlation coefficient**:

$$R^2 = r^2$$

This follows from the variance decomposition (Matloff Ch. 19):

$$\text{Var}(Y) = \text{Var}[E(Y|X)] + \text{Var}[Y - E(Y|X)]$$

10.3 Connection to Correlation

R^2 is literally the **square of the correlation coefficient**:

$$R^2 = r^2$$

This follows from the variance decomposition (Matloff Ch. 19):

$$\text{Var}(Y) = \text{Var}[E(Y|X)] + \text{Var}[Y - E(Y|X)]$$

Dividing both sides by $\text{Var}(Y)$:

$$1 = \underbrace{\frac{\text{Var}[E(Y|X)]}{\text{Var}(Y)}}_{R^2} + \underbrace{\frac{\text{Var}[Y - E(Y|X)]}{\text{Var}(Y)}}_{1-R^2}$$

10.3 Connection to Correlation

R^2 is literally the **square of the correlation coefficient**:

$$R^2 = r^2$$

This follows from the variance decomposition (Matloff Ch. 19):

$$\text{Var}(Y) = \text{Var}[E(Y|X)] + \text{Var}[Y - E(Y|X)]$$

Dividing both sides by $\text{Var}(Y)$:

$$1 = \underbrace{\frac{\text{Var}[E(Y|X)]}{\text{Var}(Y)}}_{R^2} + \underbrace{\frac{\text{Var}[Y - E(Y|X)]}{\text{Var}(Y)}}_{1-R^2}$$

Note: $R^2 = r^2 = 0.95$ means 95% of the variance in exam scores is "explained" by study hours. The remaining 5% is unexplained noise.

10.4 Recap

Today we covered:

- **Regression function:** $m(t) = E[Y|X = t]$ models the mean of Y given X
- **Simple linear model:** $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- **Least squares:** $\hat{\beta}_1 = \mathrm{Cov}(X, Y)/\ \mathrm{Var}(X), \quad \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$
- **Residuals:** always plot your data (Anscombe's Quartet)
- $R^2 = r^2$ = proportion of variance explained by the model