Chapter 5

Regression

The Prediction Problem

How do we make predictions from data?

Every day, algorithms make predictions that affect your life:

- Spotify predicts which songs you'll like based on your listening history
- Insurance companies predict your risk based on demographics and behaviour
- Universities predict student success from high school grades
- Weather apps predict tomorrow's temperature from atmospheric data

Intended Learning Outcomes

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

- Understand the equation of a straight line
- Define the least squares regression line
- Calculate slope and intercept from summary statistics
- Make predictions and recognise when they are unreliable
- Interpret slope and intercept in context
- Calculate and interpret R^2
- Compute and analyse residuals
- Identify influential points
- Recognise lurking variables and ecological fallacy

PART 1

Review of Straight Lines

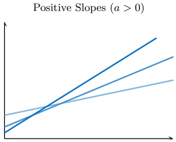
Quick Review: Straight Lines

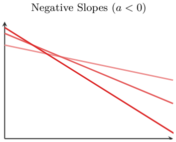
Straight Line

A **straight line** can be described by the equation

$$y = a + bx$$

where a is the y-**intercept** and b is the **slope**.





- The **slope** b tells you how much y changes when x increases by 1
- The **intercept** a is the value of y when $x = 0$

Example 5.1: Reviewing Lines

Given the points $(2, 5)$ and $(6, 13)$:

- Find the equation of the line passing through these points.
- What is the value of y when $x = 4$ on this line?

PART 2

Regression Lines

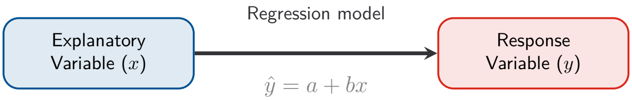
From Scatterplots to Prediction

Regression Line

A **regression line** is a straight line that describes how a response variable y changes as an explanatory variable x changes. It is used to **predict** the value of y based on the value of x.

New notation:

- $\hat{y}$ = predicted value of y
- $\hat{y} = a + bx$ = the regression equation

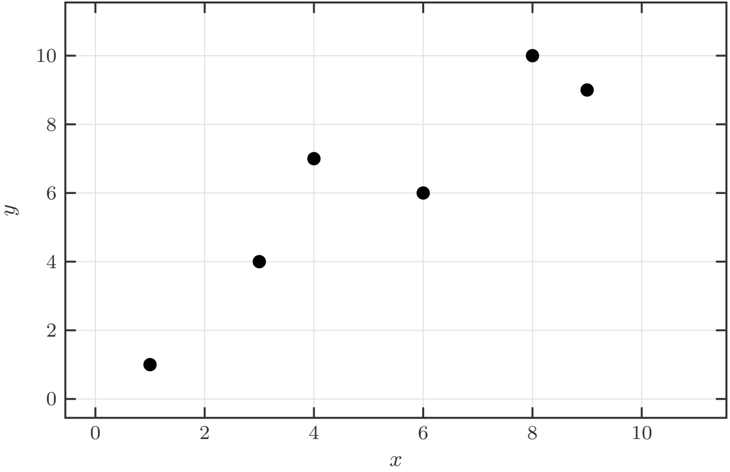


Notation Summary

We use $\hat{y}$ ("y-hat") to denote predicted values from a regression line.

Symbol	**Meaning**
x	Explanatory variable (input)
y	Response variable (observed output)
$\hat{y}$	Predicted value of y from the regression line
$\hat{y}_i$	Predicted value for the i-th observation: $\hat{y}_i = a + bx_i$
a	y-intercept of the regression line
b	Slope of the regression line

What should our line be?



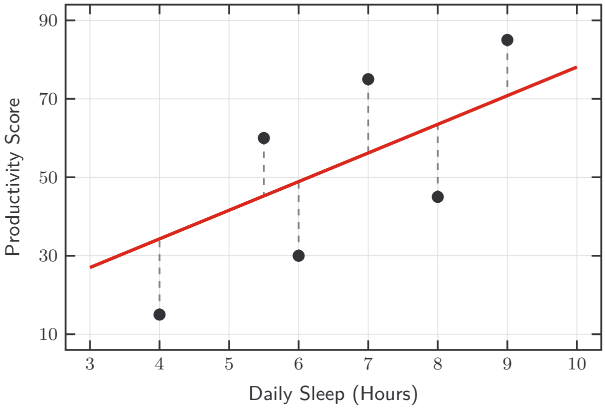
PART 3

The Least Squares Regression Line

Least Squares Regression Line

What is the least squares regression line?

The **least squares regression line** is the “best-fitting” straight line through a scatterplot.



The Least Squares Criterion

Least Squares Regression Line

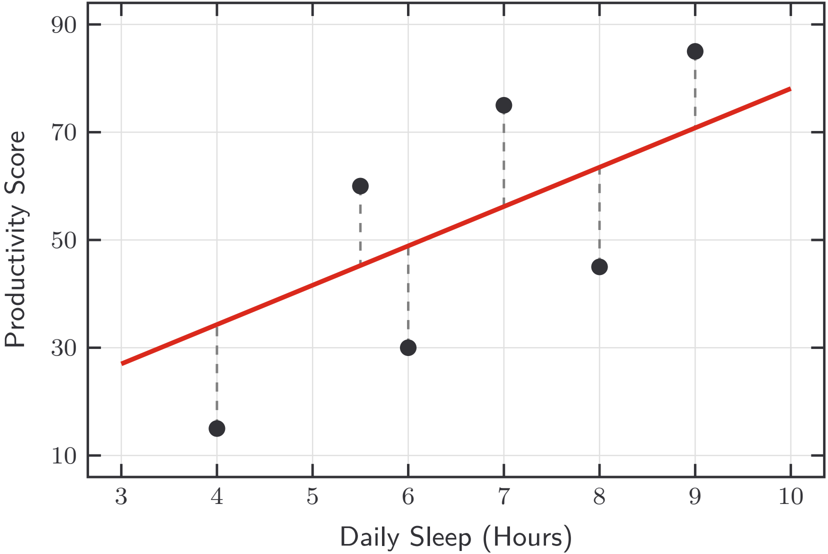
The **least squares regression line** is the line that minimizes the sum of the **squared vertical distances** between observed values and predicted values:

$$\text{Minimize} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Why square the residuals?

- Treats positive and negative errors equally (no cancellation)
- Penalizes large errors more heavily than small ones
- Yields clean mathematical formulas

Visualising the Least Squares Idea



The Formulas for the Least Squares Line

The least squares regression line is given by

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

where

$$b = r \cdot \frac{s_y}{s_x} \quad \text{(slope)}$$

$$a = \bar{y} - b\bar{x} \quad \text{(intercept)}$$

Example 5.2: Advertising and Daily Sales

Context: A small business tracks weekly ads and daily sales over 5 weeks. What is the least squares regression line of daily sales (y) on ads per week (x)?

Ads/week (x)	1	2	3	4	5
Sales ($100s) ($y$)	1	5	11	9	9

Summary statistics:

- $\bar{x} = 3$, $\bar{y} = 7$
- $s_x = 1.581$, $s_y = 4$

Slope:

$b =$

Intercept:

$a =$

Equation:

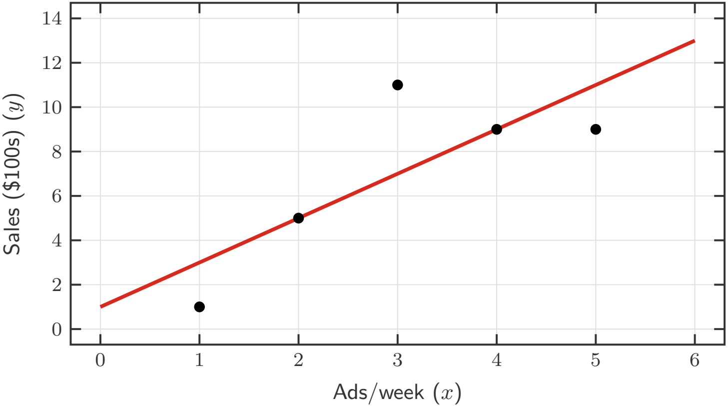
$\hat{y} =$

Phraseology

When we refer to the line predicting y from x, we can say any of the following:

- The regression of y on x
- The least squares regression line predicting y from x
- The least squares regression line with y as the response and x as the explanatory variable
- The line with y as the dependent variable and x as the independent variable
- The line modeling y as a function of x
- The line regressing y against x

Advertising and Daily Sales: Visualization



Example 5.3: Spotify Playlist Data

Context: Data from 200 songs released in 2024. Find the regression line to predict total streams (millions) from the number of playlists (thousands) a song appears on. Summary statistics:

- $\bar{x} = 18.5$ thousand playlists
- $\bar{y} = 142.3$ million streams
- $s_x = 12.4$, $s_y = 98.7$
- $r = 0.78$

Understanding the Slope Formula

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

What does this formula tell us?

- The **sign** of b matches the sign of r
  - Positive correlation $\Rightarrow$ positive slope
  - Negative correlation $\Rightarrow$ negative slope
- The **magnitude** depends on how spread out y is relative to x
  - If $s_y > s_x$: the slope is steeper than r
  - If $s_y < s_x$: the slope is flatter than r

Understanding the Intercept Formula

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

What does this formula tell us?

- The regression line **always passes through** the point $(\bar{x}, \bar{y})$
  - This is the "balance point" of the data
  - Substituting $x = \bar{x}$ gives $\hat{y} = a + b\bar{x} = \bar{y}$
- The intercept a adjusts the line vertically to ensure it passes through $(\bar{x}, \bar{y})$

PART 4

Interpreting the Coefficients

Interpreting the Slope

Interpretation of Slope

The **slope** b represents the **predicted change** in y for a **one-unit increase** in x.

"For each additional [unit of x], the predicted [response variable] changes by b [units of y]."

Advertising example: For $\hat{y} = 1 + 2x$:

Spotify example: For $\hat{y} = 27.4 + 6.21x$:

Interpreting the Intercept

Interpretation of Intercept

The **intercept** a represents the predicted value of y when $x = 0$.

"When [explanatory variable] is zero, the predicted [response variable] is a [units]."

Spotify example: "A song on zero playlists would get 27.4 million streams."

⚠ **Caution:** This interpretation is only meaningful if $x = 0$ makes sense in context.

Example 5.4: Interpreting a House Price Model

Context: A model predicts house prices ($000s) from size (sq ft):

$$\hat{y} = 45 + 0.12x$$

Interpret the slope and intercept.

PART 5

Predictions

How Do We Use the Regression Line?

To make predictions of the response variable y based on new values of the explanatory variable x.

Advertising example: $\hat{y} = 1 + 2x$
What is the expected daily sales if the business runs 7 ads per week?

Making Predictions

Making Predictions with Regression

1. Compute the regression equation $\hat{y} = a + bx$ based on the data
2. Substitute the given x-value into the equation
3. Calculate $\hat{y}$ (the predicted value)

Example 5.3: Spotify Playlist Data (Continued)

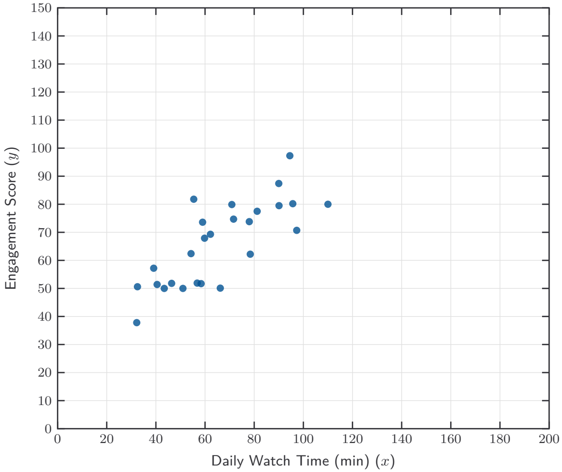
Context: Data from 200 songs released in 2024. We want to predict total streams (millions) from the number of playlists (thousands) a song appears on.
Recall that the regression equation is:

$$\hat{y} = 27.4 + 6.21x$$

a) Find the predicted number of total streams for a song on 10,000 playlists.
b) Find the predicted number of total streams for a song on 1,344 playlists.

Example 5.5: Netflix Watch Time and Engagement

Context: Netflix analyzes subscriber data to predict monthly engagement score (0–100) from average daily watch time (minutes).

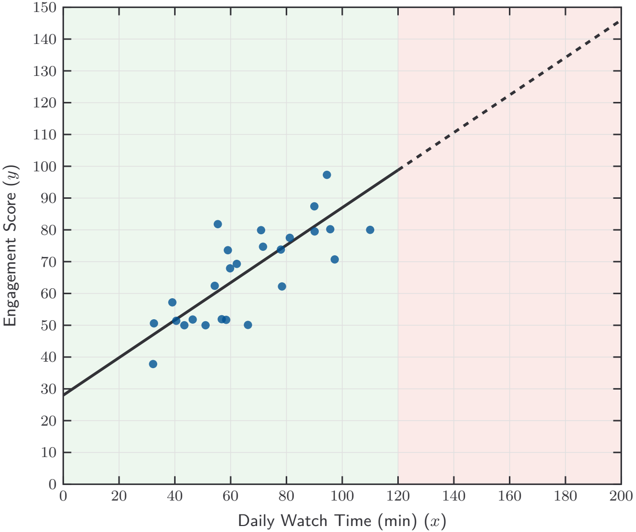


Example 5.5: Netflix Watch Time and Engagement (Continued)

Context: Netflix analyzes subscriber data to predict monthly engagement score (0–100) from average daily watch time (minutes). Regression line: $\hat{y} = 28 + 0.59x$

Predict the engagement score for
(a) a subscriber watching 60 min/day, and
(b) a subscriber watching 180 min/day.

Example 5.5: Netflix Watch Time and Engagement



Example 5.6: Children's Height and Age

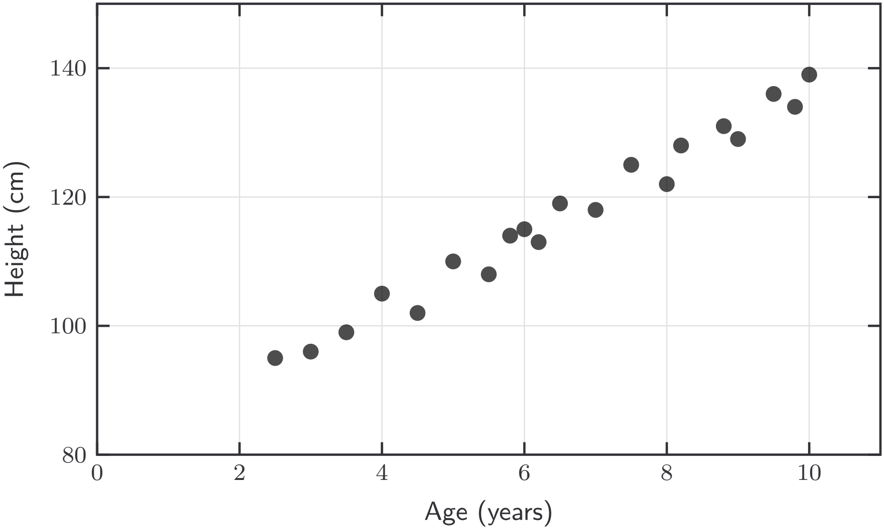
Context: A pediatrician models the relationship between children's height (cm) and age (years) for patients aged 2–10 years. Data from 180 children shows: $\bar{x} = 6.2$ years, $\bar{y} = 115.4$ cm, $s_x = 2.4$, $s_y = 14.8$, $r = 0.92$.

(a) Predict the height of a 3-year-old.

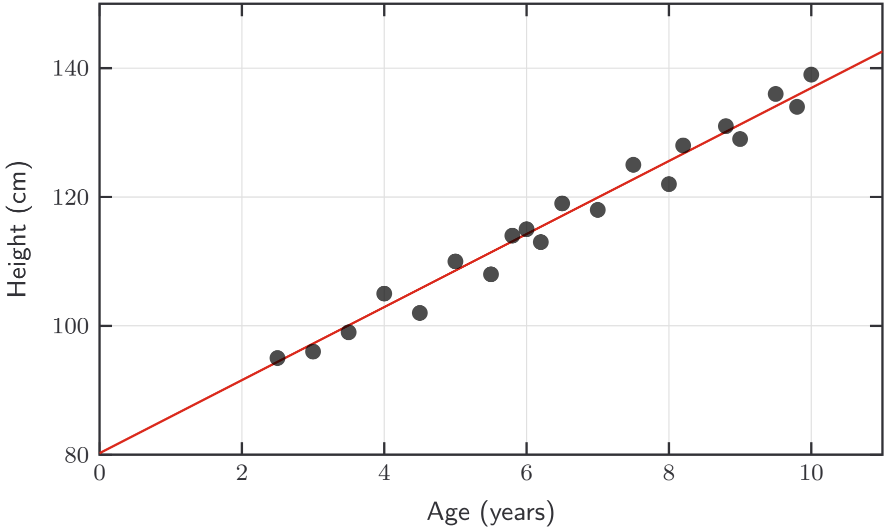
(b) Predict the height of a 7-year-old child.

(c) Predict the height of a 33-year-old.

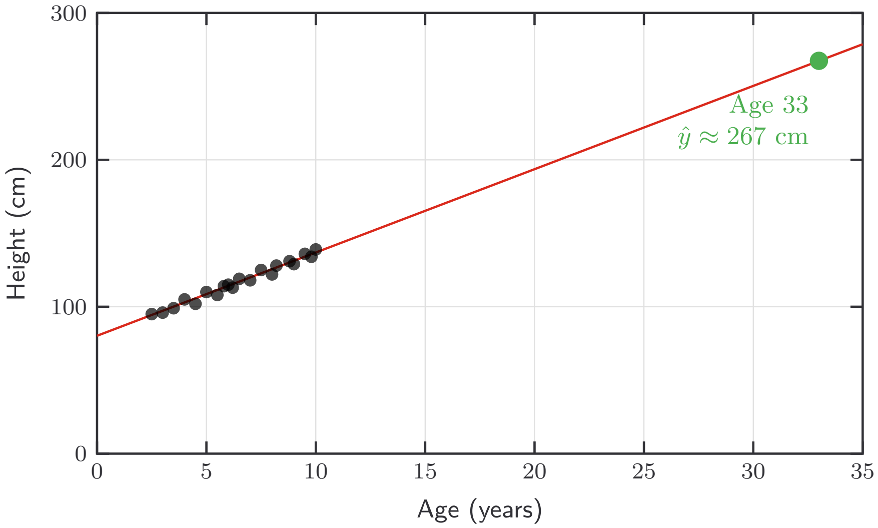
Example 5.6: Children's Height and Age



Example 5.6: Children's Height and Age



Example 5.6: Children's Height and Age



Limitations of Predictions

Predictions are not trustworthy if

- They are based on **extrapolation** (outside the data range)
- The predicted values are **inadmissible** in context (e.g., negative prices)

PART 6

Diagnostics

The Coefficient of Determination

Coefficient of Determination

The **coefficient of determination** R^2 is the **squared correlation** between x and y:

$$R^2 = r^2$$

The value R^2 always lies between 0 and 1.

The coefficient of determination R^2 represents the **proportion of variation** in the response variable y that is **explained** by the explanatory variable x using the regression line.

Key Point: Higher R^2 means a better fit of the regression line to the data.

Example 5.7: Computing R^2

Advertising example: We found $r = 0.791$. What is R^2?

$$R^2 =$$

Interpretation:

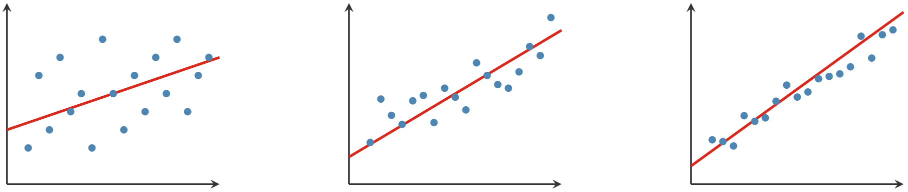
Spotify example: We found $r = 0.78$. What is R^2?

$$R^2 =$$

Interpretation:

Visualising Different R^2 Values

$R^2 \approx 0.13$ $R^2 \approx 0.65$ $R^2 \approx 0.90$



Intuition: Higher R^2 means points cluster more tightly around the line.

Residuals

Residual

A **residual** is the difference between an observed value and its predicted value from the regression line:

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

Interpretation:

- **Positive residual**: Actual value is **above** the line (model **underestimates**)
- **Negative residual**: Actual value is **below** the line (model **overestimates**)
- **Zero residual**: Perfect prediction (point lies on line)

Key Point: Residuals tell you how wrong each prediction is, and in which direction.

Example 5.8: Computing Residuals

Context: A bookstore tracks the number of hours open per day (x) and daily revenue in hundreds of dollars (y) over 5 days. The regression line is $\hat{y} = 2 + 3x$.

Hours (x)	Revenue (y)	Predicted ($\hat{y}$)	Residual (e_i)
4	15		
6	19		
8	26		
10	31		
12	39		

Sum of Residuals is Zero

Note

The least squares method finds the line that minimizes $\sum(y_i - \hat{y}_i)^2$. Calculus shows that for this minimum, the sum of residuals must be zero:

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

This means the average residual is also zero, which implies:

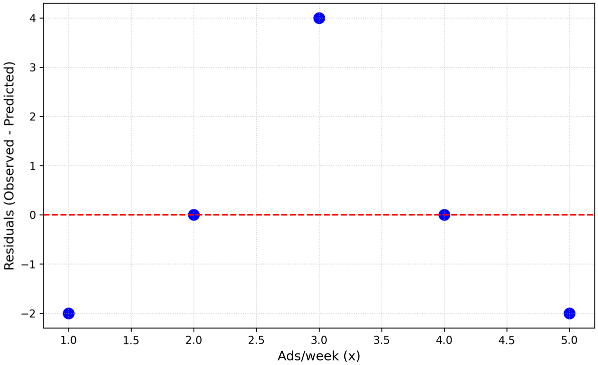
$$\bar{y} = \bar{\hat{y}} = a + b\bar{x}$$

Key Point: This is why the regression line **always** passes through $(\bar{x}, \bar{y})$.

Residual Plots: Checking Your Model

Residual Plot

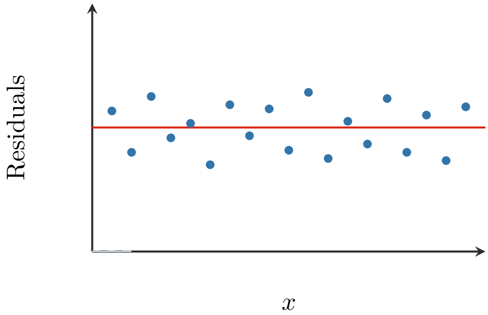
A **residual plot** shows residuals (y-axis) versus (fitted) predicted values or explanatory variable (x-axis).



Residual Plots

If the model is reasonable, then the residual plots would show:

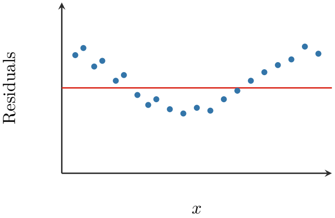
- No systematic pattern
- Points randomly distributed around zero
- Constant spread across all x values

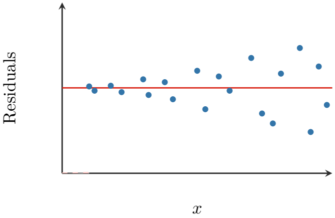


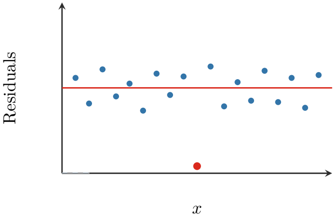
Residual Plots

Otherwise, we might see:

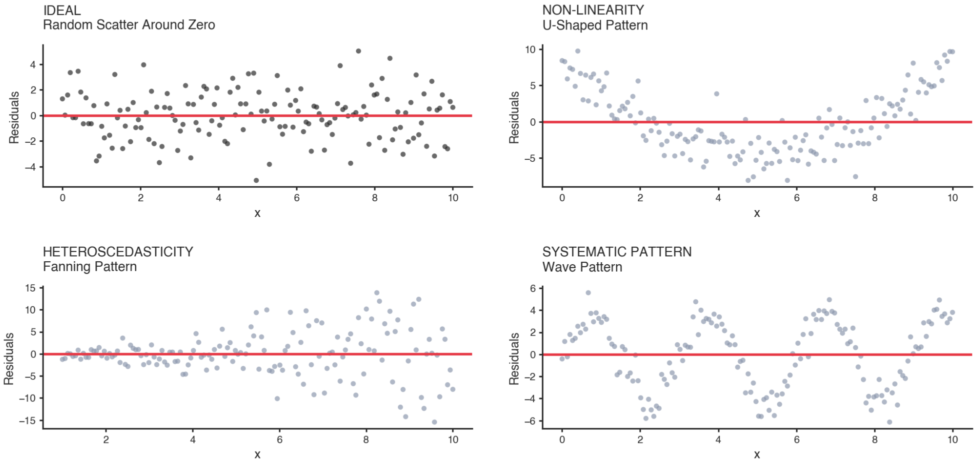
- Curved patterns (non-linearity)
- Funnel shapes (non-constant variance)
- Large residuals (potential outliers)







Residual Plots



PART 7

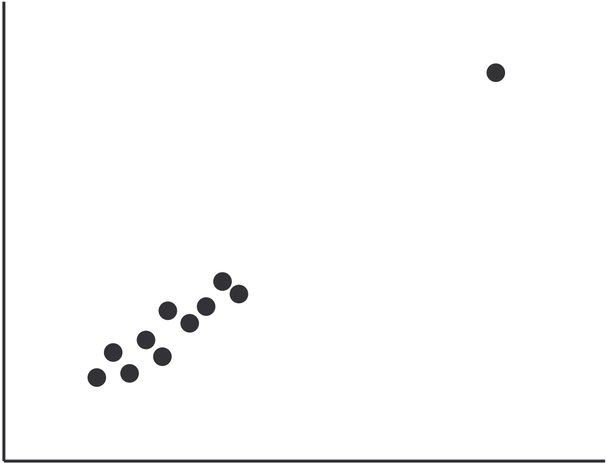
Outliers and Regression

Outliers and Regression

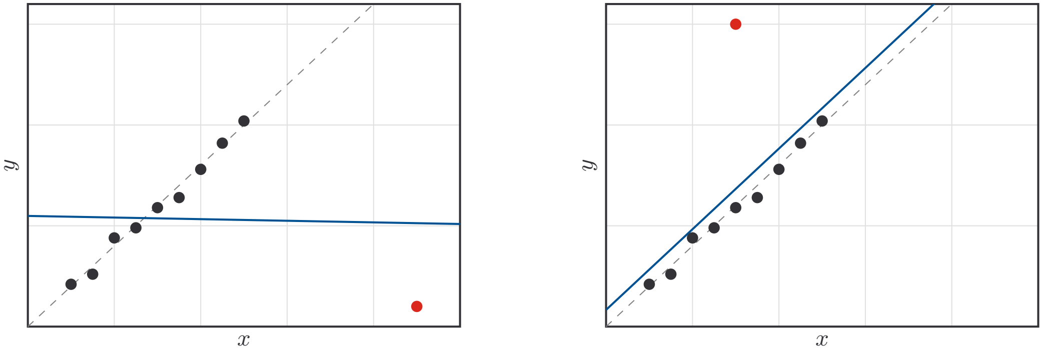
Certain types of outliers can substantially affect regression models in two ways:

- Position: They can pull the regression line toward themselves, changing both slope and intercept
- Strength: They can weaken the correlation (reduce R^2), making predictions less reliable

Impact on Regression Model



Types of Influential Points



What Makes an Outlier Influential?

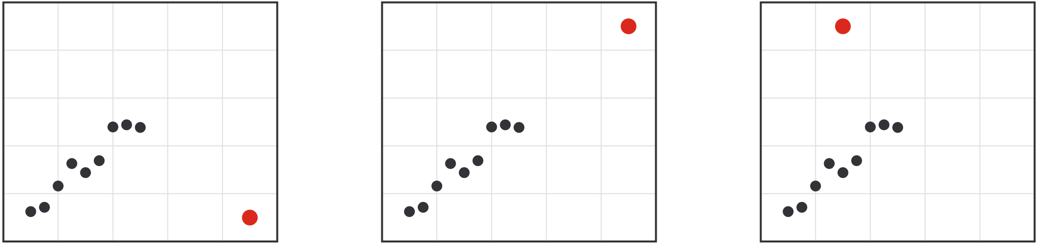
Influential Point

An **influential point** is an observation that, if removed, would substantially change the regression line (slope, intercept, or both).

Generally speaking, this is a point that has an extreme value of x with a value of y that deviates from the regression line.

Example 5.9: Identifying Influential Points

For each plot, indicate if the highlighted point is influential.



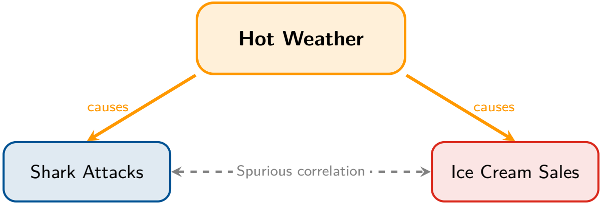
PART 8

Pitfalls

Lurking Variables

Lurking Variable

A **lurking variable** (or **confounding variable**) is an unobserved variable that influences **both** the explanatory and response variables, potentially creating a spurious association.



Correlation Does Not Imply Causation

⚠ **Caution:** A strong relationship between x and y does **not** mean that x causes y. There may be lurking variables creating a spurious association.

Famous examples of spurious correlations:

- Ice cream sales and drowning deaths (*lurking: summer weather*)
- Shoe size and reading ability in children (*lurking: age*)
- Number of firefighters and fire damage (*lurking: size of fire*)
- Nicolas Cage films and swimming pool drownings (*lurking: nothing, just coincidence!*)

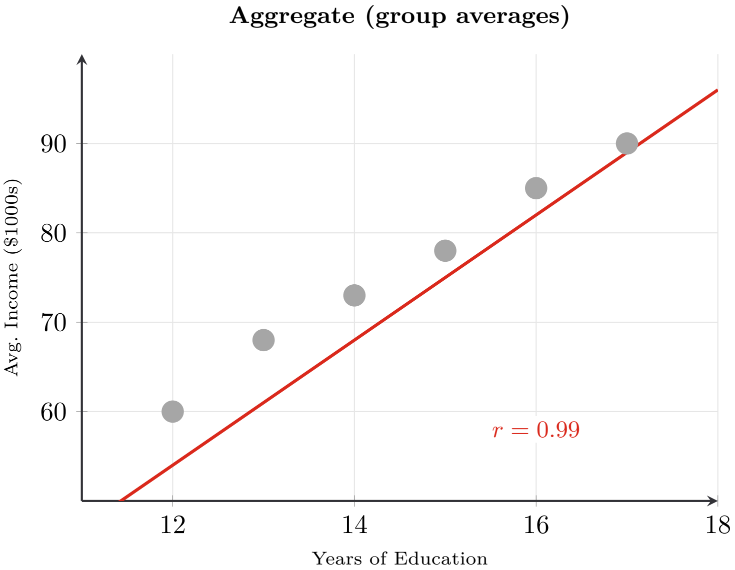
The Ecological Fallacy

Ecological Fallacy

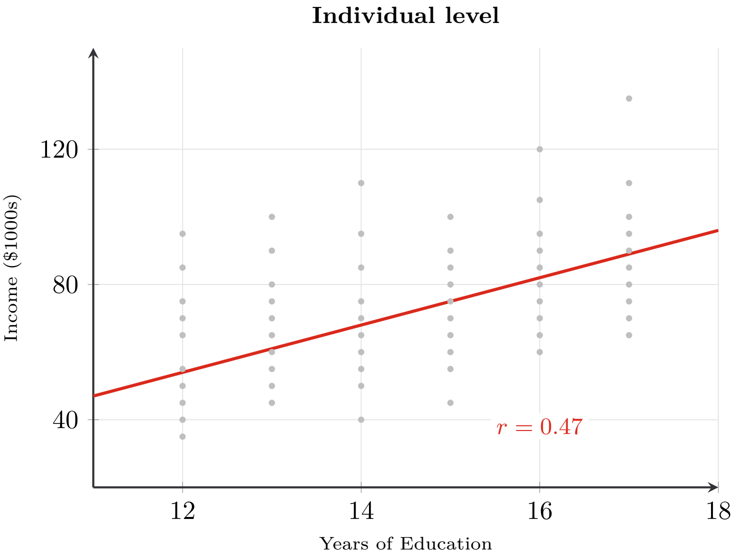
The **ecological fallacy** occurs when inferences about **individual** behaviour are incorrectly drawn from **aggregate** (group-level) data.

This is a special case of Simpson's paradox, which we will discuss in the next chapter.

Example 5.10: Education and Income



Example 5.10: Education and Income (Continued)



Chapter Summary: Core Formulas

Least Squares Regression Line

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

where:

- **Slope:** $b = r \cdot \frac{s_y}{s_x}$
- **Intercept:** $a = \bar{y} - b\bar{x}$

Measuring Model Quality

- **Residual:** $e_i = y_i - \hat{y}_i$
- **Coefficient of determination:** $R^2 = r^2$

Chapter Summary: Interpretation

Understanding the Coefficients

Slope b**:** Change in predicted y for each one-unit increase in x

- "For each additional [unit of x], we predict y to change by b [units]."

Intercept a**:** Predicted value of y when $x = 0$

- Only meaningful if $x = 0$ makes sense in context
- Often just a mathematical anchor for the line

Coefficient of determination R^2**:** Proportion of variation in y explained by x

- "$R^2 \times 100\%$ of the variation in [response] is explained by [explanatory variable]."

Chapter Summary: Making Predictions

How to Predict

1. Calculate regression equation $\hat{y} = a + bx$
2. Substitute the given value of x
3. Calculate $\hat{y}$ (the predicted value)

When Predictions Fail

- **Extrapolation:** Predicting outside the range of observed x values
- **Inadmissible predictions:** Results that are impossible or nonsensical in context
  - Examples: negative prices, scores above 100%, heights of 300 cm

Chapter Summary: Model Diagnostics

Checking the Model

Residual plots should show:

- No systematic patterns
- Random scatter around zero
- Constant spread across all x values

Watch out for:

- Curved patterns (suggests non-linear relationship)
- Funnel shapes (non-constant variance)
- Large isolated residuals (potential outliers)

Influential points:

- extreme x values that deviate from the trend in y
- dramatically change slope and intercept

Chapter Summary: Common Pitfalls

Correlation $\neq$ Causation

- A strong relationship between x and y does not mean x causes y
- **Lurking variables** can create spurious associations
- Example: Ice cream sales and drowning deaths (lurking: summer weather)

Ecological Fallacy

- Group-level patterns may not hold at the individual level
- Example: Countries with higher average education may have higher average income, but within countries the pattern could reverse

Additional Practice Problems

Example 5.11: Fast Food Spending

Context: A study of 50 college students examined the relationship between weekly fast food spending ($) and GPA. Summary statistics:

- $\bar{x} = 15$ dollars, $\bar{y} = 3.2$ GPA
- $s_x = 8.5$, $s_y = 0.6$
- $r = -0.68$

a) Calculate the slope and intercept of the regression line.
b) Write the regression equation $\hat{y} = a + bx$.
c) Interpret the slope in context.
d) Predict the GPA of a student who spends $25 per week on fast food.

Example 5.12: Sleep and Test Performance

Context: A teacher tracks student sleep hours (night before exam) and exam scores (0–100) for 40 students. Summary statistics:

- $\bar{x} = 7$ hours, $\bar{y} = 72$ points
- $s_x = 1.8$, $s_y = 12.5$
- $r = 0.82$

a) Find the regression equation.
b) A student slept 9 hours. Predict their exam score.
c) Calculate and interpret R^2.
d) Is this prediction reliable? Why or why not?

Example 5.13: Temperature and Coffee Sales

Context: A coffee shop records daily temperature (°C) and coffee sales (cups sold) over 60 days. The regression line is:

$$\hat{y} = 280 - 8.5x$$

a) Interpret the slope.
b) Interpret the intercept. Does it make practical sense?
c) Predict sales when it's 20°C.
d) Predict sales when it's -5°C. Is this a reliable prediction? Why?

Example 5.14: Computing Residuals

Context: Using the fast food/GPA regression line $\hat{y} = 3.8 - 0.04x$, compute residuals for four students:

Student	Spending (x)	GPA (y)	Predicted ($\hat{y}$)	Residual (e_i)
A	10	3.5		
B	20	3.0		
C	5	3.6		
D	30	2.8		

Which student's GPA does the model overestimate? Underestimate?

Example 5.15: Putting It Together

Context: A researcher studying 100 athletes measures training hours per week (x) and performance score (0–100) (y). Results:

- Regression equation: $\hat{y} = 40 + 2.1x$
- $R^2 = 0.71$
- Data range: 5–25 training hours
- Three athletes have residuals of -8, $+5$, and $+12$

a) Interpret $R^2 = 0.71$ in context.

b) Which athlete's performance does the model most severely overestimate?

c) Predict the score for an athlete training 15 hours/week.

d) Predict the score for an athlete training 40 hours/week.

Example 5.16: Fitness App Data

Context: A fitness app tracked 50 users over 12 weeks to understand the relationship between weekly workout sessions and weight loss (kg).

	Sessions/week	Weight loss (kg)
Mean	$\bar{x} = 3.50$	$\bar{y} = 4.20$
Standard deviation	$s_x = 1.41$	$s_y = 1.68$

The regression equation is $\hat{y} = 0.63 + 1.02x$. Find R^2 for this linear model.

Example 5.17: Height and Shoe Size (Part 1)

Context: A researcher studies the relationship between height (in inches) and shoe size (US men's sizing) among college students. After collecting data, they obtain the following least squares regression line:

$$\text{Height} = 50 + 2.5 \times (\text{Shoe Size})$$

(a) Suppose the residual for a particular observation is 1 inch and the observed height is 74 inches. What is the shoe size of this student? Show your work.

Example 5.17: Height and Shoe Size (Part 2)

Context: Continuing with the regression line:

$$\text{Height} = 50 + 2.5 \times (\text{Shoe Size})$$

(b) If the mean height of the sampled students is 66.67 inches, what is the mean shoe size? Show your work.

Example 5.17: Height and Shoe Size (Part 3)

Context: Using the same data, the researchers found that the least squares regression line for predicting shoe size from height is:

$$\text{Shoe Size} = -3.2 + 0.16 \times (\text{Height})$$

(c) Calculate the correlation coefficient (r) between height and shoe size. Show your work.

Hint: Consider the relationship between the two slopes and the correlation coefficient.

Example 5.17: Height and Shoe Size (Part 4) 🌶

(d) If we measure height in centimeters instead of inches (where 1 inch = 2.54 cm), would the correlation coefficient change? Would the slope of the regression line change? Explain.

Consider both:

- How does a linear transformation affect correlation?
- How does a linear transformation affect the slope?