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

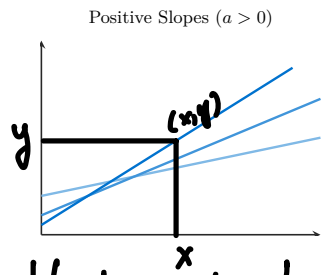
Straight Line

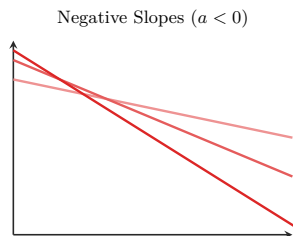
A **straight line** can be described by the equation

$$y = a + bx$$

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 $x = 0$





High school

$y = mx + b$

- b = intercept

y when $x = 0$

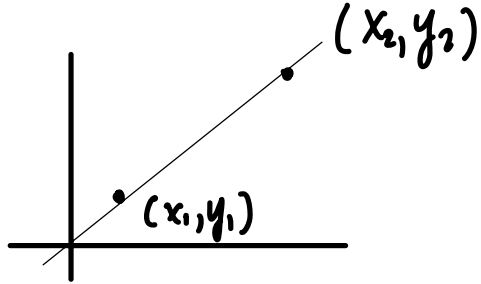
- m = slope

$\Delta y = m \Delta x$

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?



- Every 2 points determine a unique (straight) line
- We only need the intercept a and slope b to find the line

Note: generally if is easier to find slope first

$$b = \frac{\Delta y}{\Delta x} = \frac{13 - 5}{6 - 2} = \frac{8}{4} = 2 \quad \leftarrow \text{slope}$$

$(2, 5) = (x_1, y_1)$

a: Note that $y = a + 2x \Rightarrow 5 = a + 2(2) = a + 4$

$\Rightarrow a = 1$

- $\therefore y = 1 + 2x$
- $y = 1 + 2(4) = 9$

PART 2

Regression Lines

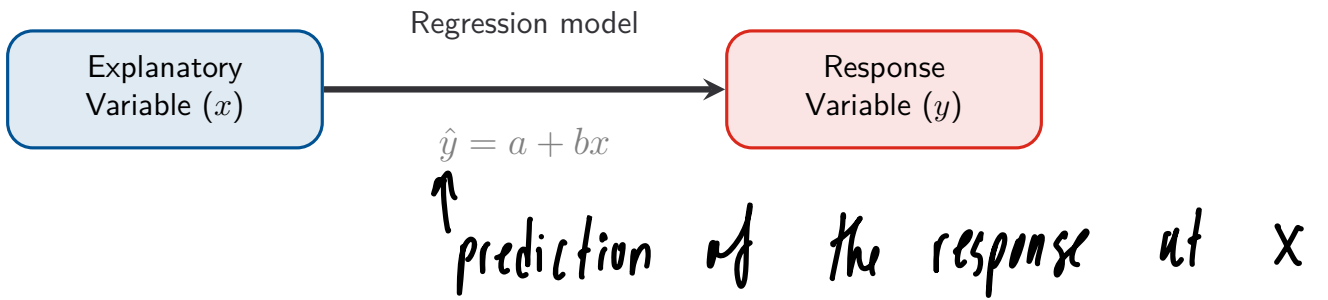
From Scatterplots to Prediction

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



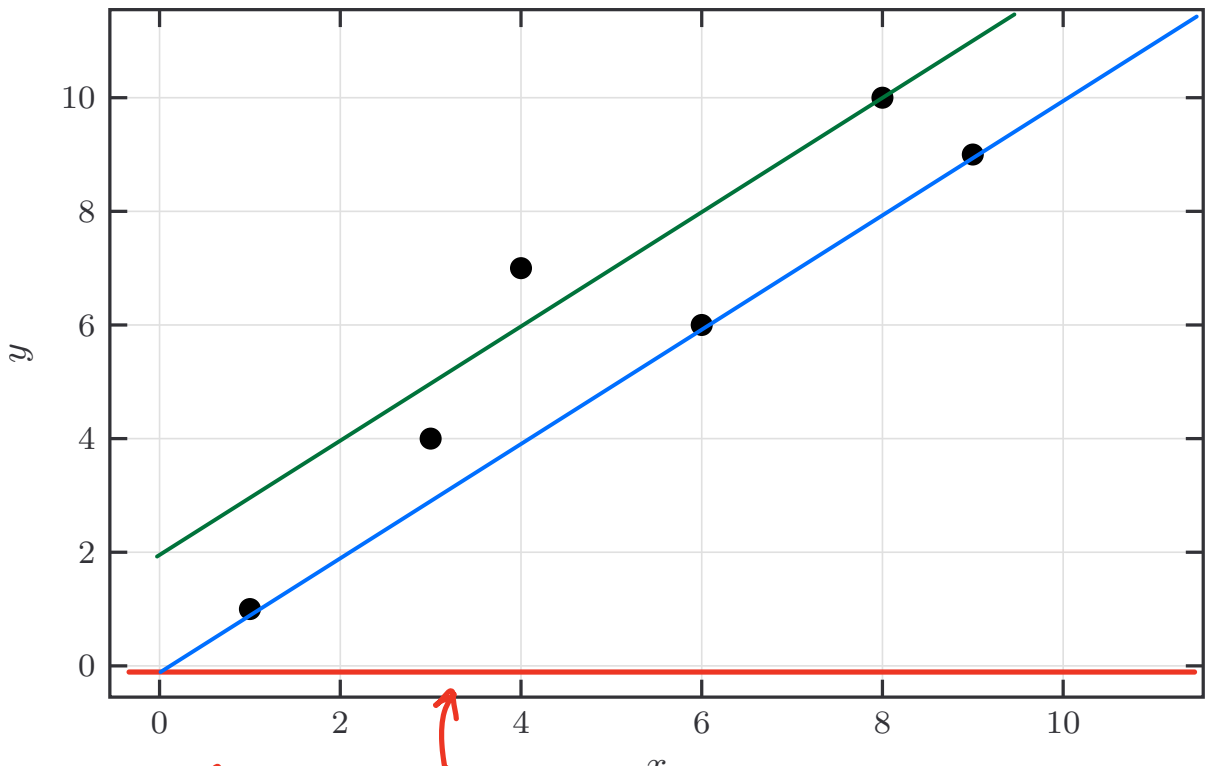
Notation Summary

Prediction according to our model

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

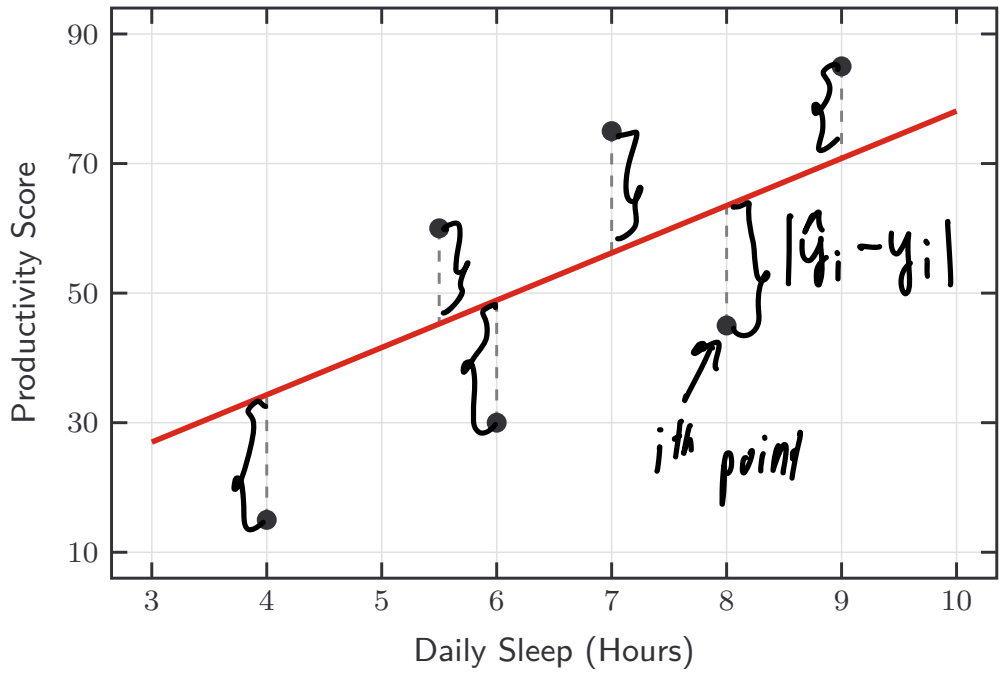
What is the least squares regression line?

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

Minimize the differences between $\hat{y}_i$ and y_i

Goal: Minimize $\sum_{i=1}^{n} |\hat{y}_i - y_i|$

super hard to minimize



Instead, we minimize the squared distances $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

The Least Squares Criterion

Least Squares Regression Line

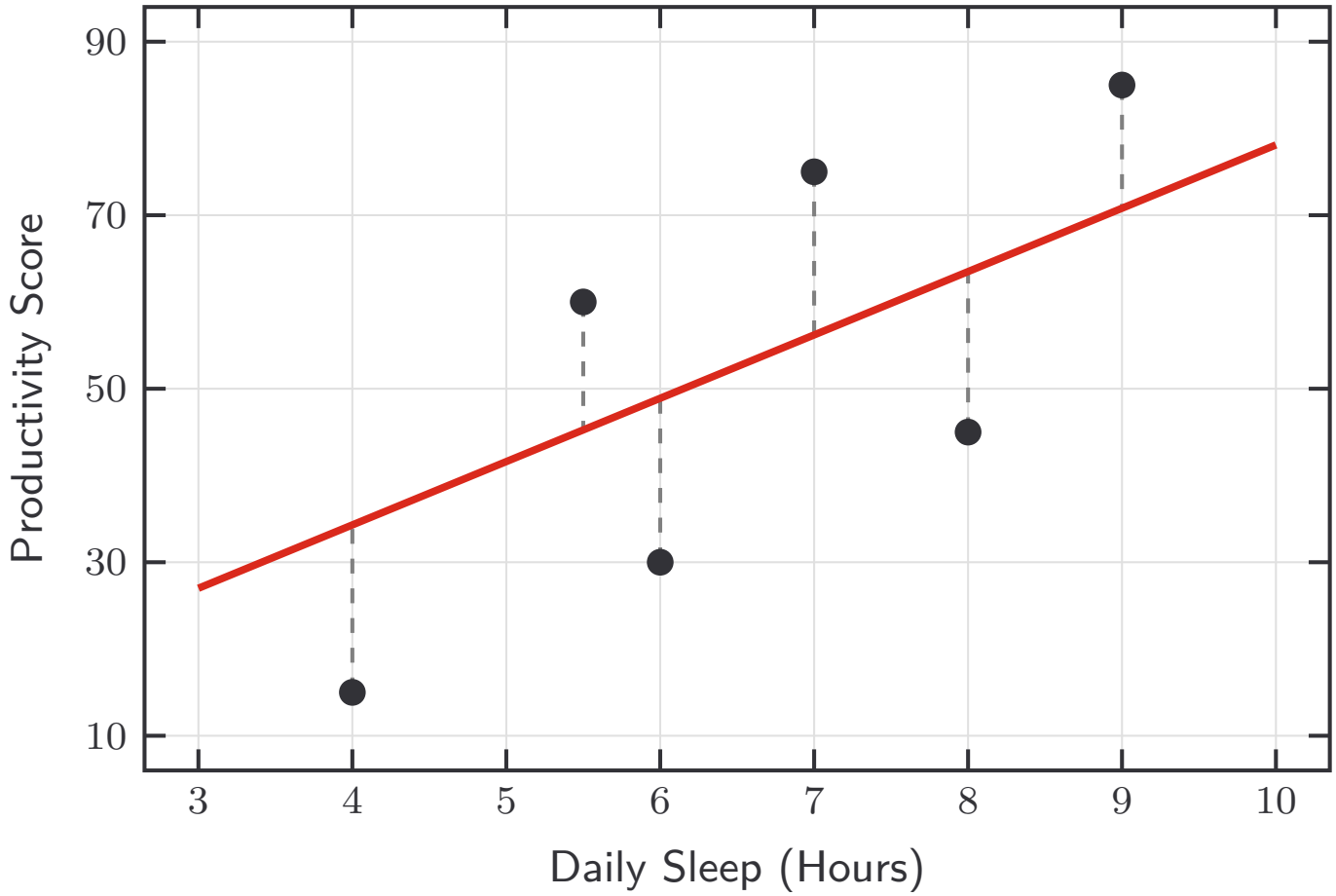
The **least squares regression line** is the line that minimizes the sum of the **squared vertical distances** between observed values and predicted values:

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

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

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

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

calculate slope first

Example 5.2: Advertising and Daily Sales

Goal: Find $\hat{y} = a + bx$

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:

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

- $r = 0.791$

Slope:

$b = r \cdot \frac{s_y}{s_x} = 0.791\left(\frac{4}{1.581}\right) = 2$

Intercept:

$a = \bar{y} - b\bar{x} = 7 - 2(3) = 1$

Equation:

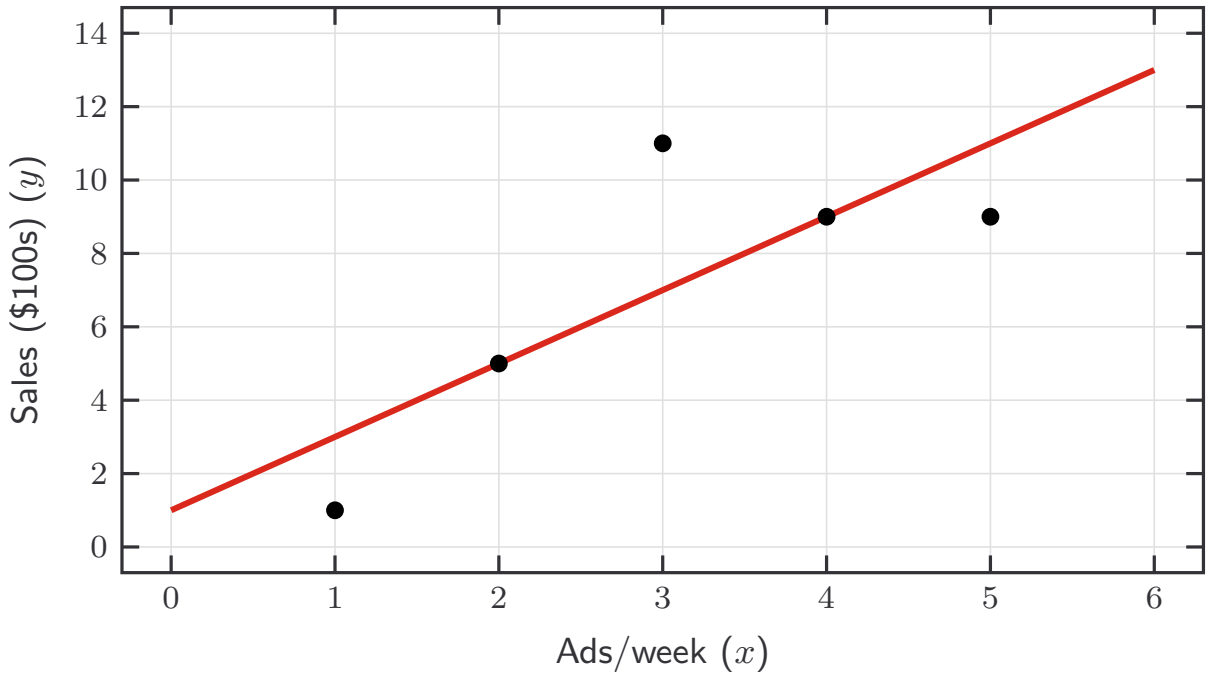
$\hat{y} = 1 + 2x$

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
- $s_x = 12.4$, $s_y = 98.7$
- $r = 0.78$

$\hat{y} = a + bx$, $b = r\frac{s_y}{s_x} = (0.78)\frac{(98.7)}{(12.4)} = 6.21$

$a = \bar{y} - b\bar{x} = 142.3 - 6.21(18.5)$

$= 27.4$

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

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

Aside: $b\Delta x = \Delta \hat{y}$

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

For each additional ad, the expected weekly sales increase by 2 (100's of dollars).

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

For each additional 1000 playlists that include a song, the predicted number of additional streams is 6.21 million.

. $\hat{y} = 1 - 3.41x$

For each additional unit increase in [Explanatory variable], the expected [response variable] decreases by 3.41.

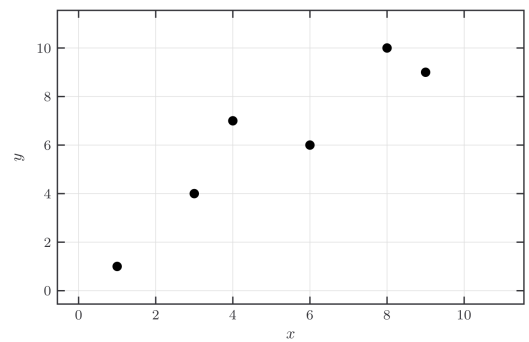
Recall

- Linear models
  - used for inference
  - used for prediction

- Form of linear model:

$\hat{y} = a + bx$, ① $b = r \frac{S_y}{S_x}$ ← corr(x,y)

② $a = \bar{y} - b\bar{x}$



21. Suppose the least squares regression line for predicting costs as a function of revenue is $\hat{y} = 3.5 + 0.60x$

What is the interpretation of slope?

21. Suppose the least squares regression line for predicting costs as a function of revenue is

$\hat{y} = 3.5 + 0.60x$

What is the interpretation of slope?

For each unit increase of [x] revenue, the [y] cost increases by 0.60 [b] units.

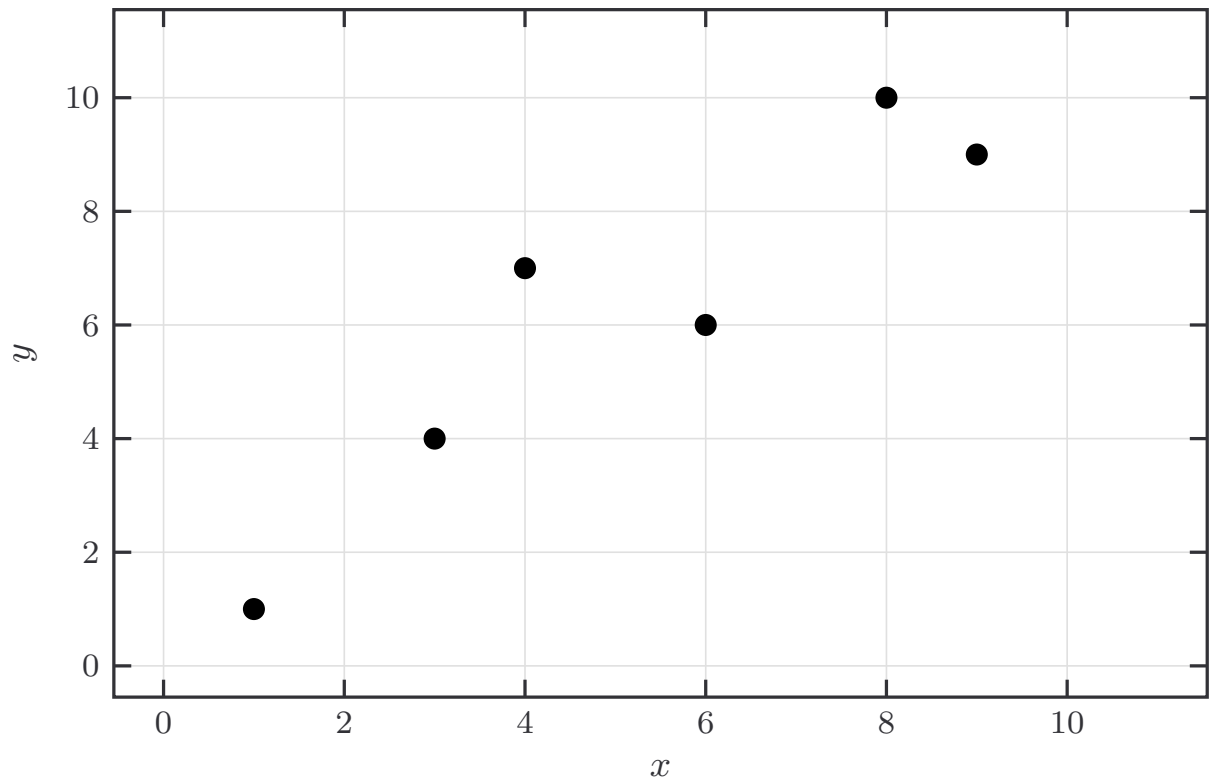
$\Delta y = b\Delta x \longrightarrow \hat{y}_0 = 3.5 + 0.60x_0$

$(y_1 - y_0) = b(x_1 - x_0)$

$\Delta y = \hat{y}_1 - \hat{y}_0 = 3.5 + 0.60(x_0 + 1) - (3.5 + 0.60x_0)$

$= 0.60 \cdot 1$

What should our line be?



Interpreting the Intercept

Recall, mathematically: a = value of y when $x=0$

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.

$\hat{y} = 27.4 + 6.21x \xrightarrow{x=0} \hat{y} = 27.4$

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.

Slope: Each additional square foot increase results in $(0.12 \times 1000) = \$120$ increase in expected (or predicted) house price.

Intercept: As there are no house with 0 sq. ft. it does not make sense to interpret.

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?

$x = 7$

LM: $\hat{y} = 1 + 2x \longrightarrow$ Compute $1 + 2(7) = 1 + 14 = 15$

The expected sales of 7 ads per week is 15

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.

a) $x = 10{,}000$

$\hat{y} = 27.4 + 6.21x$
$= 27.4 + 6.21(10000)$

Correction: $\hat{y} = 27.4 + 6.21(10)$
$= 89.5$ million streams

Need to convert to the correct units

b) $x = 1344$

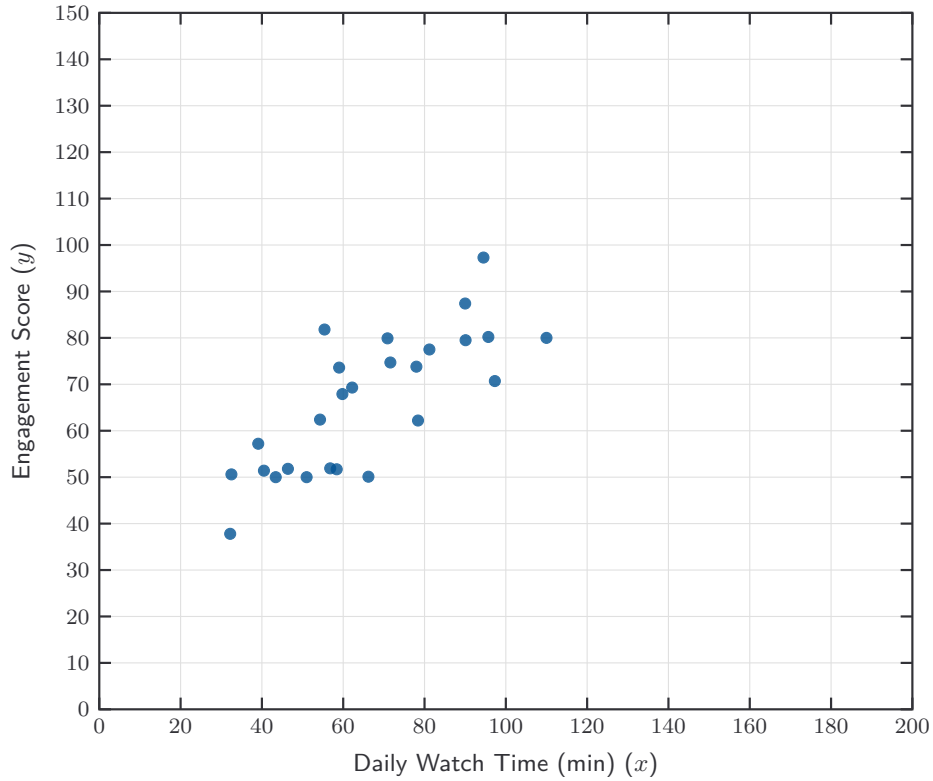
$\hat{y} = 27.4 + 6.21(1344)$
$=$

$\hat{y} = 27.4 + 6.21(1.344)$
$= 37.75$ million streams

Thanks Nicole! ☺

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.

LM: $\hat{y} = 28 + 0.59x$

a) $x = 60$

$\hat{y} = 28 + 0.59(60)$

$= 63.4$

b) $x = 180$

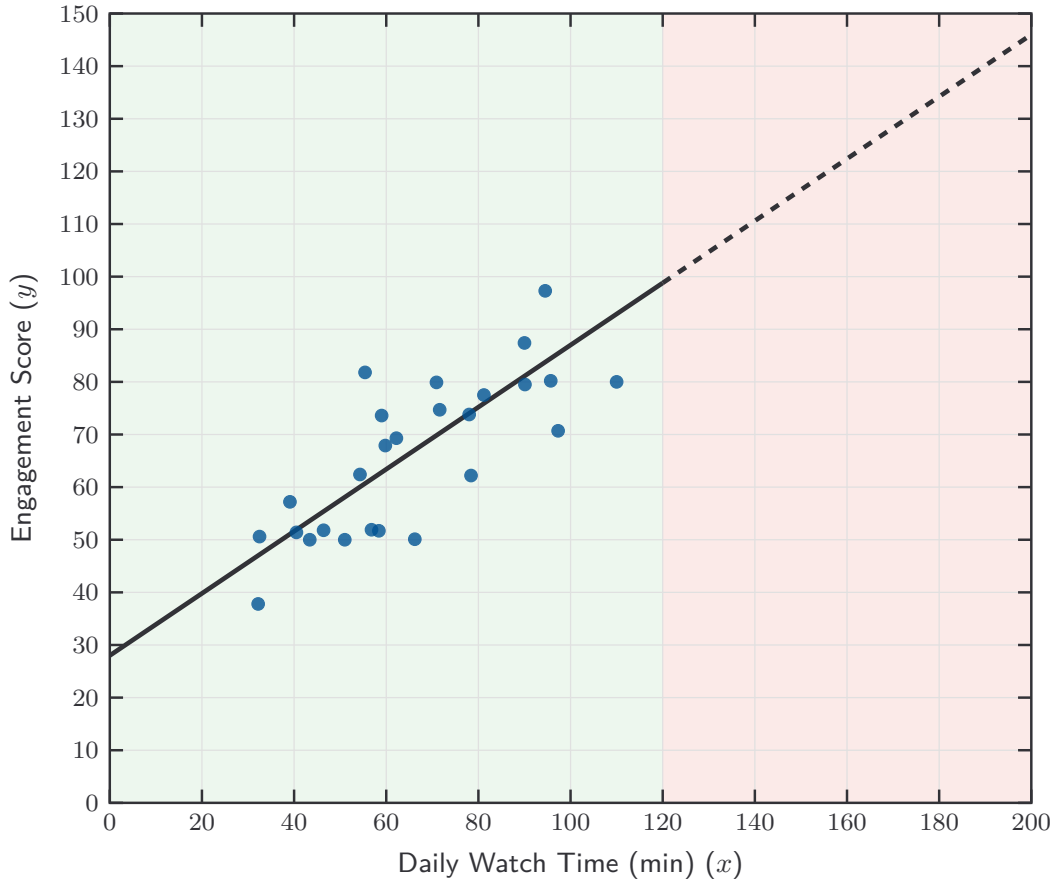
$\hat{y} = 28 + 0.59(180)$

$= 134.2$

inadmissible

The value of y must be in $(0, 100)$.

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.

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

$b = r\frac{s_y}{s_x} = 0.92\frac{(14.8)}{2.4} = 5.67$ $a = \bar{y} - b\bar{x}$

$= 115.4 - (5.67)(6.2)$

$= 80.22$

$= 80.22 + 5.67x$

a) $\hat{y} = 80.22 + 5.67(3) = 97.2$

b) $\hat{y} = 80.22 + 5.67(7) = 119.9$

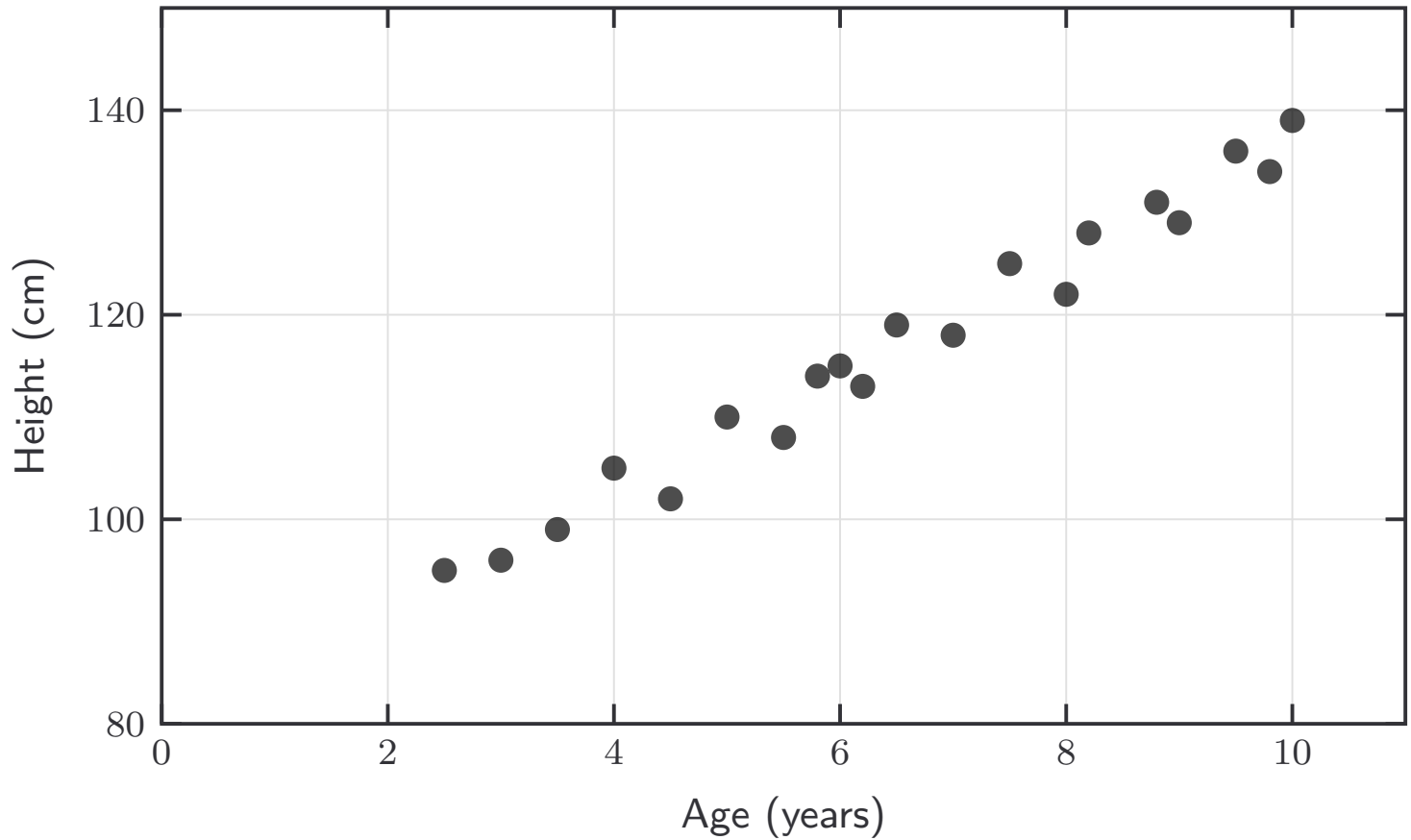
c) $\hat{y} = 80.22 + 5.67(33)$

$= 267.3$ cm

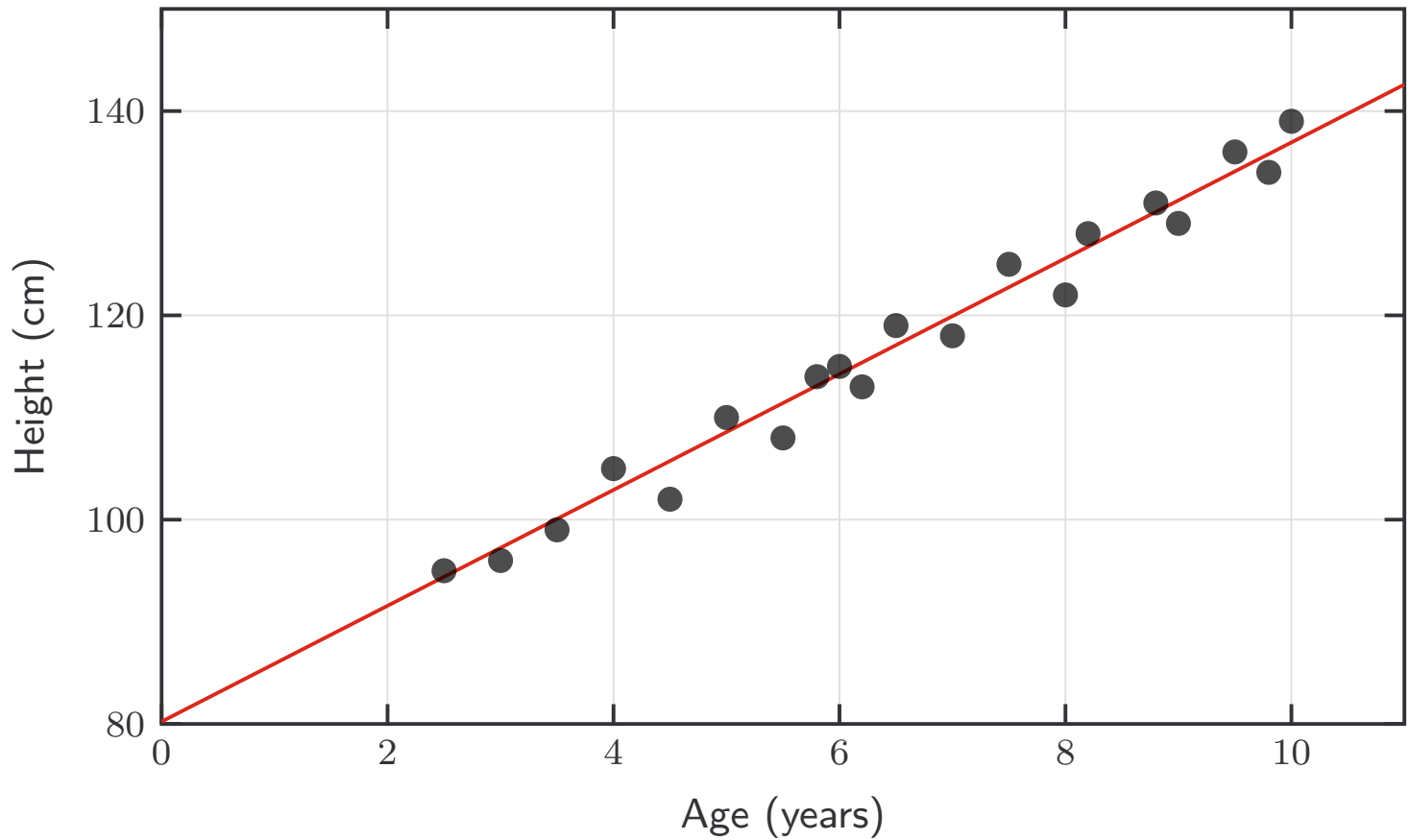
Absurd.

The problem: 33 year olds are not children. (except me)

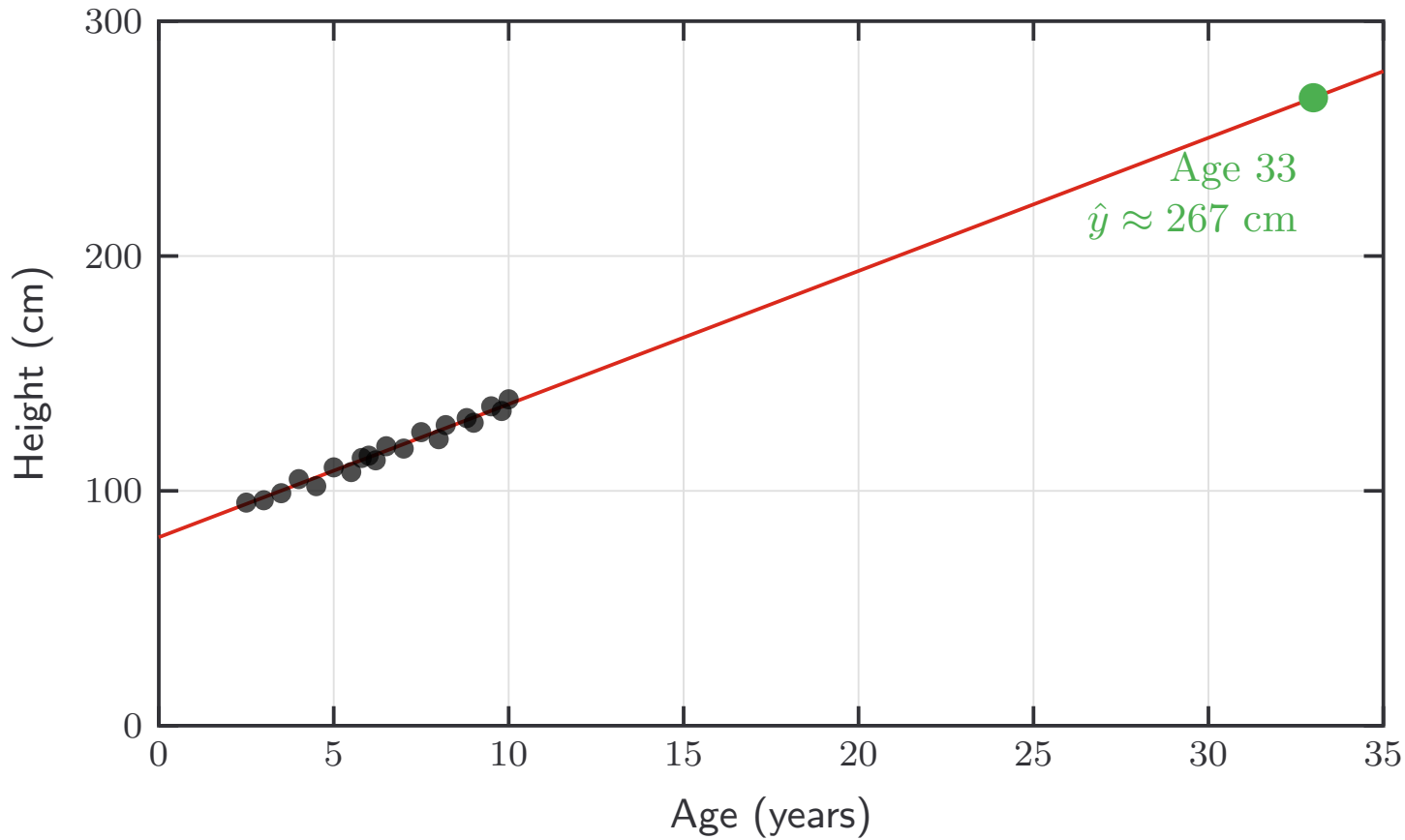
Example 5.6: Children's Height and Age



Example 5.6: Children's Height and Age



Example 5.6: Children's Height and Age



Limitations of Predictions

Is the explanatory variable close to the range of data we trained the model with?

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)

Does response make sense?

Does the explanatory variable make sense?

PART 6

Diagnostics

The Coefficient of Determination

Coefficient of Determination

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

1. $R^2 = r^2$

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

2. 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?

$R^2 =$ $(0.791)^2 = 0.626$

Interpretation: The proportion of total variation in sales which is explained by ads is 0.626.

[y] [x]

Spotify example: We found $r = 0.78$.

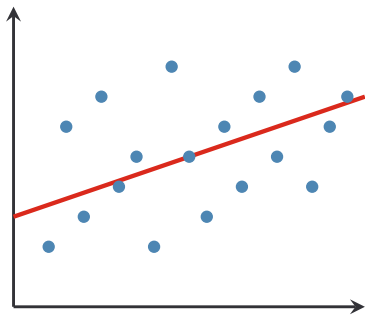
What is R^2?

$R^2 =$ $(0.78)^2 = 0.608$

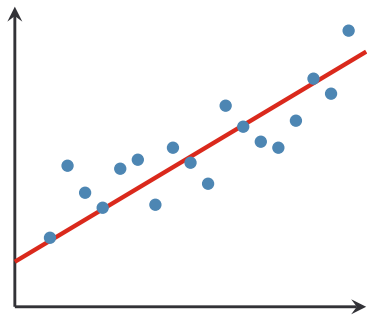
Interpretation: The proportion of total variation in song streams which is explained by ads is 0.626.

Visualising Different R^2 Values

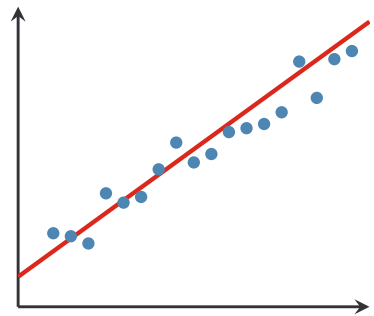
$R^2 \approx 0.13$



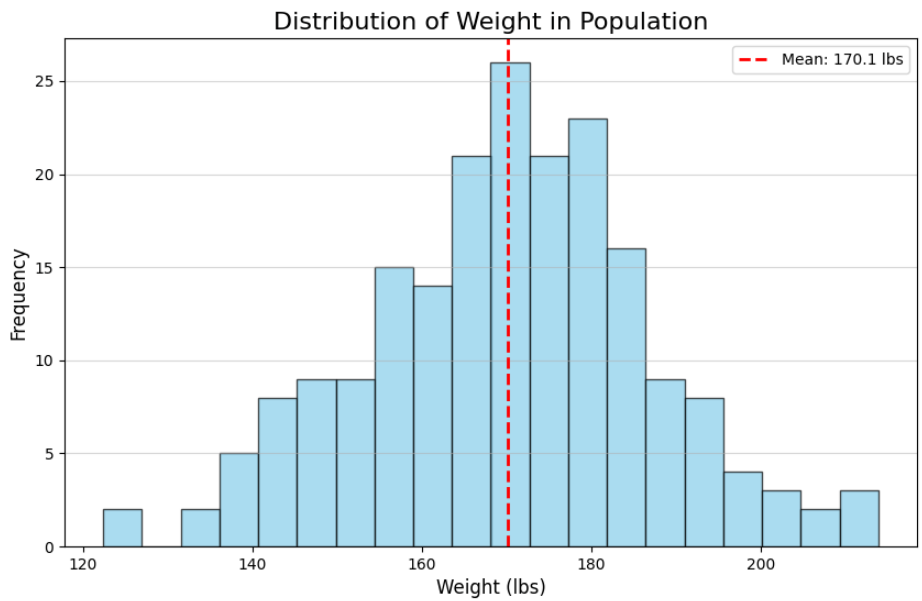
$R^2 \approx 0.65$

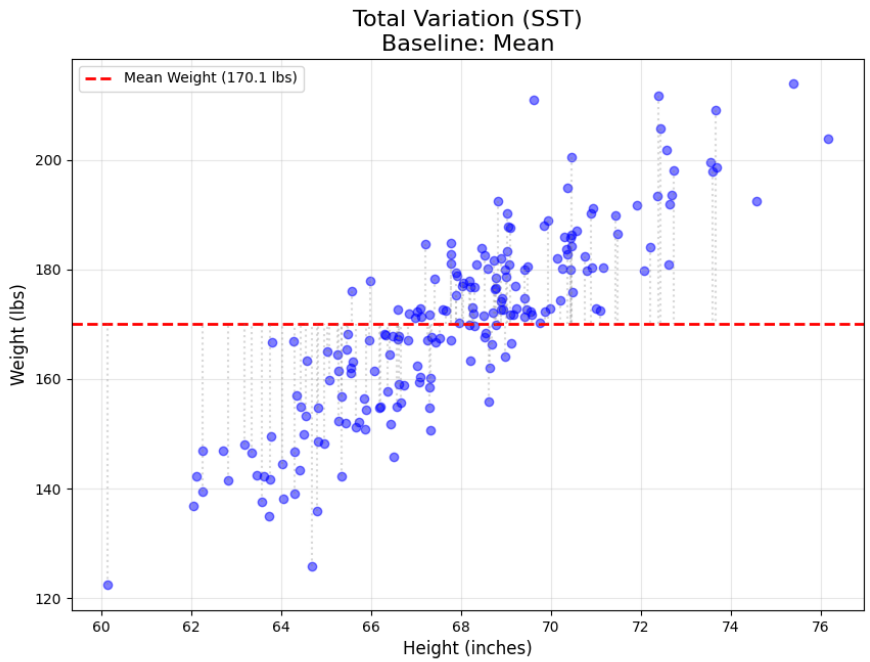


$R^2 \approx 0.90$



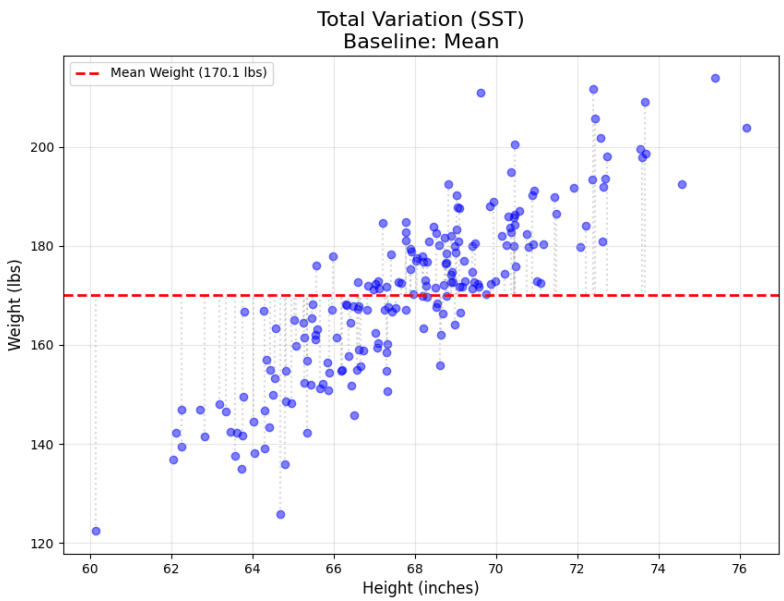
Intuition: Higher R^2 means points cluster more tightly around the line.

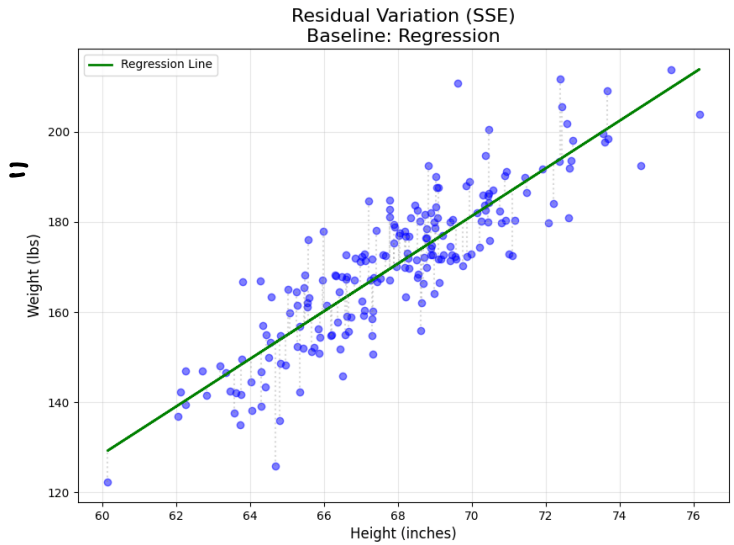

Distribution of Weight in Population
Mean: 170.1 lbs
Frequency
Weight (lbs)
0
5
10
15
20
25
120
140
160
180
200

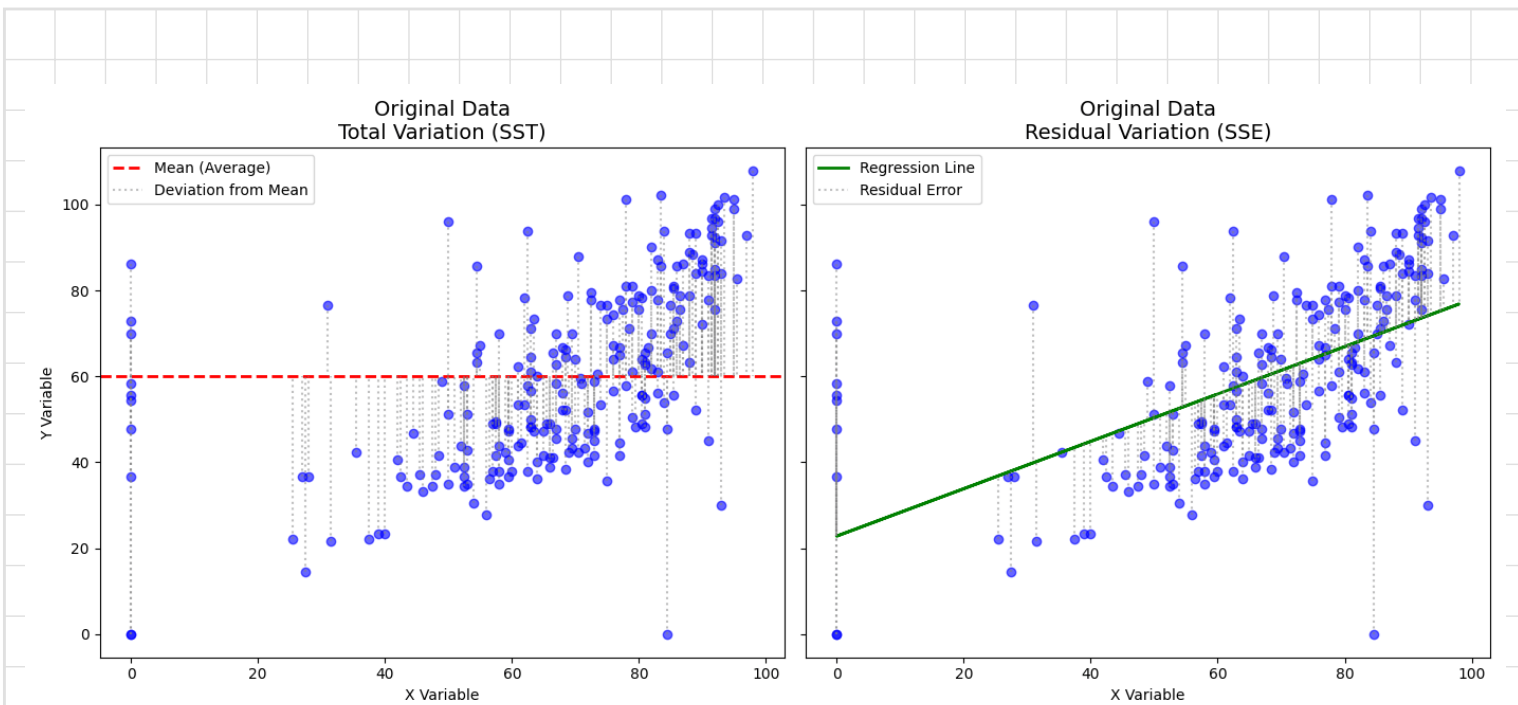


This uses the model

$\hat{y} = \bar{y}$


Total Variation (SST)
Baseline: Mean
Mean Weight (170.1 lbs)
Weight (lbs)
Height (inches)
120
140
160
180
200
60
62
64
66
68
70
72
74
76


Residual Variation (SSE)
Baseline: Regression
Regression Line
Weight (lbs)
Height (inches)
120
140
160
180
200
60
62
64
66
68
70
72
74
76


Original Data
Total Variation (SST)
Mean (Average)
Deviation from Mean
Y Variable
X Variable
Original Data
Residual Variation (SSE)
Regression Line
Residual Error
X Variable

Last time

- Interpretation of slope and intercept
  → "value of y when $x = 0$"
- Predictions

To predict at x_0 using $\hat{y} = a + bx$, substitute: predicted value $= a + b x_0$

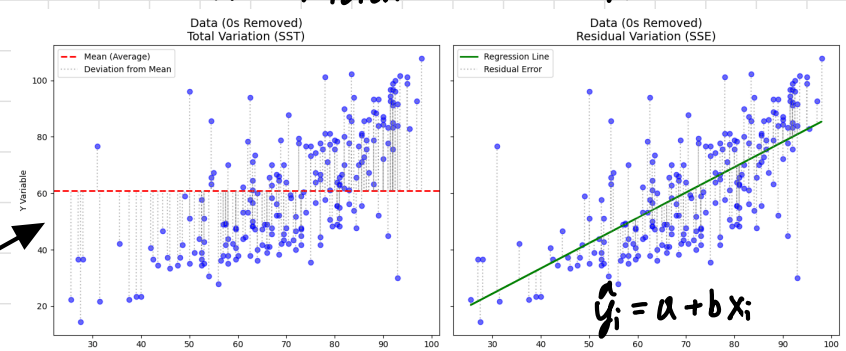
- Extrapolation (is x_0 close to the range of the data)
- Inadmissible values (check that x_0 and predicted value are plausible)

X = Midterm Y = Final

- Coefficient of Determination

$R^2 = r^2 = \text{Corr}(x, y)^2$

Use the model $\hat{y}_i = \bar{y}$



Residuals

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

Residual

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

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

$$\hat{y}_i = a + bx_i$$

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

Fact: $\sum_{i=1}^{n} e_i = 0$ for any LS LM.

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 $\hat{y} = 2 + 3x$.

$Y = X\beta + \varepsilon$

Hours (x)	Revenue (y)	Predicted ($\hat{y}$)	Residual (e_i)
4	15	$2+3(4) = 14$	$y_i - \hat{y}_i = 15 - 14 = 1$
6	19	$2+3(6) = 20$	$y_i - \hat{y}_i = 19 - 20 = -1$
8	26	26	0
10	31	32	-1
12	39	38	$+1$

Adding up residuals: $1+(-1)+(0)+(-1)+1 = 0$

Sum of Residuals is Zero

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:

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

Centroid

$\bar{y} = a + b\bar{x}$ ⟶ $(\bar{x}, \bar{y})$ lies on the LS LM

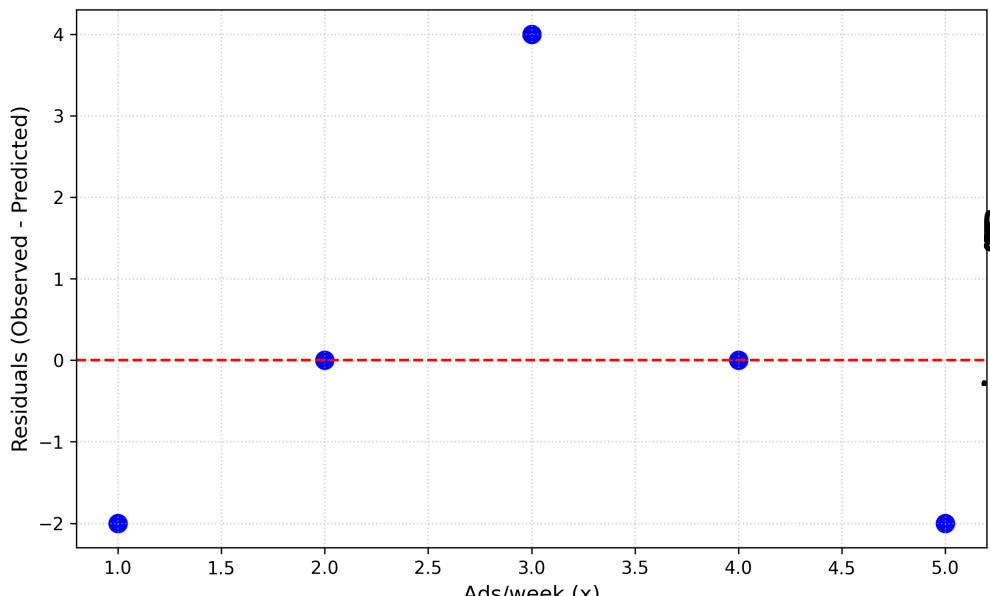
Key Point: This is why the regression line **always** passes through $(\bar{x}, \bar{y})$.

$b = r\frac{S_y}{S_x}$ $\quad a = \bar{y} - b\bar{x} \longleftrightarrow \bar{y} = a + b\bar{x}$

Residual Plots: Checking Your Model

Residual Plot

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



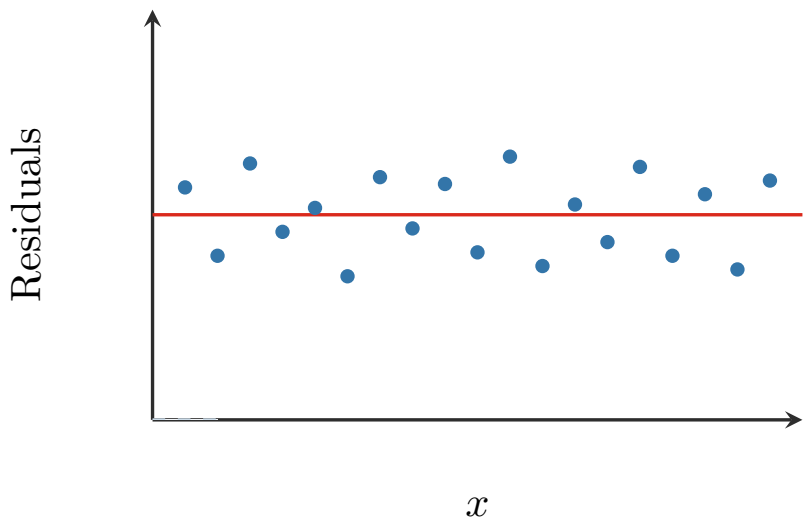
(Or, x-axis can be the "fitted values" AKA "predicted values" $\hat{y}_i = a + bx_i$)

explanatory variable

Residual Plots

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

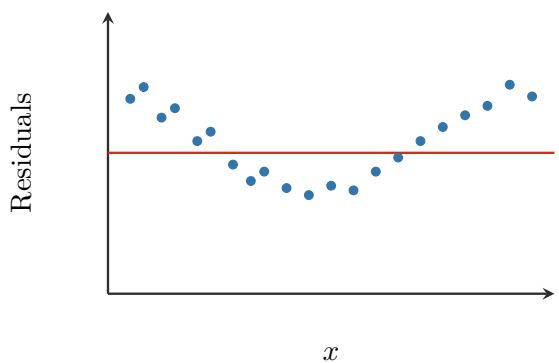
- No systematic pattern
- Points randomly distributed around zero
- Constant spread across all x values

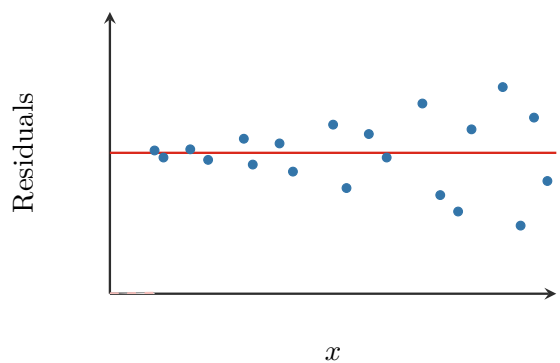


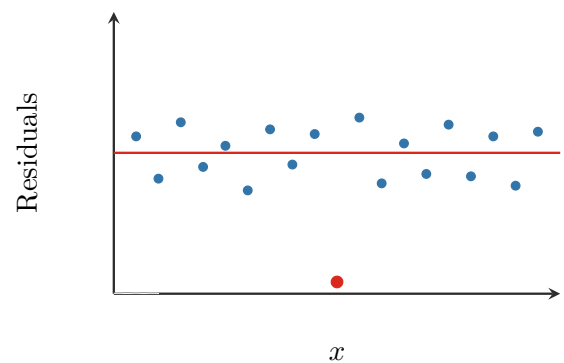
Residual Plots

Otherwise, we might see:

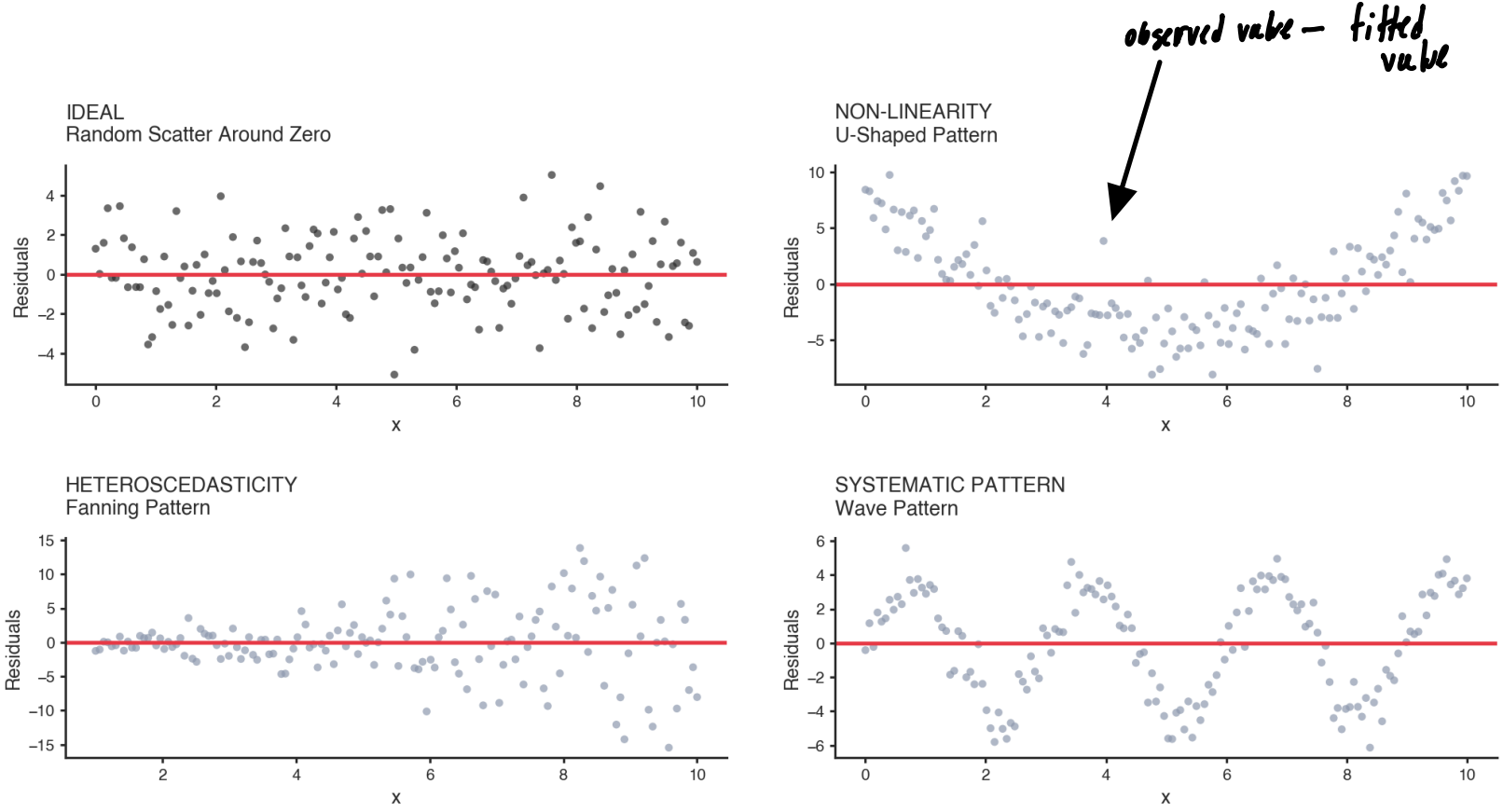
- Curved patterns (non-linearity)
- Funnel shapes (non-constant variance)
- Large residuals (potential outliers)







Residual Plots



PART 7

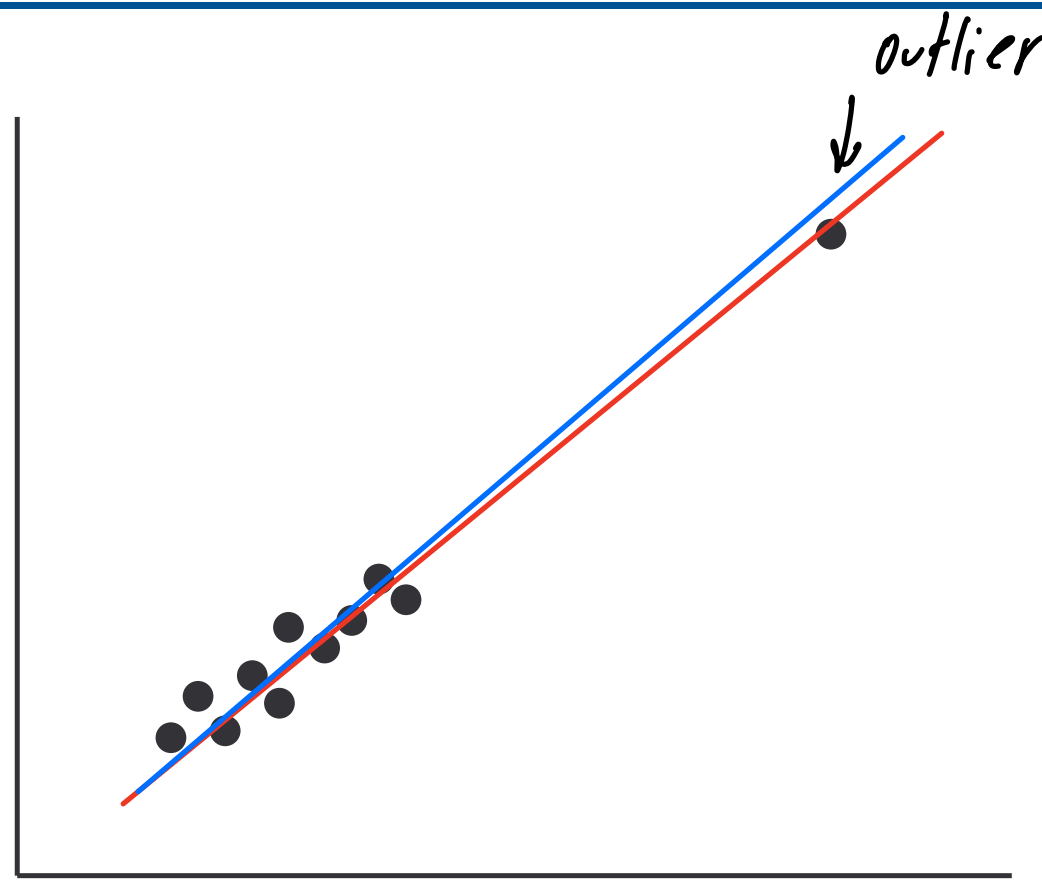
Outliers and Regression

Outliers and Regression

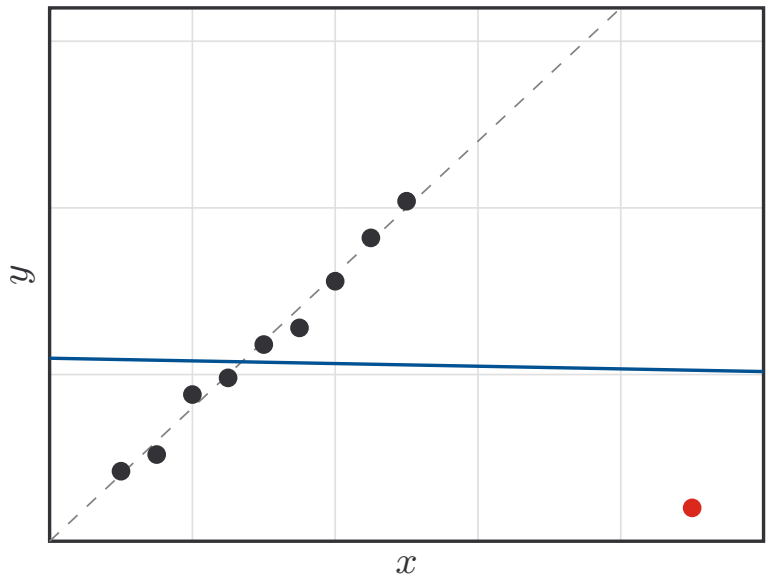
Certain types of outliers can substantially affect regression models in two ways:

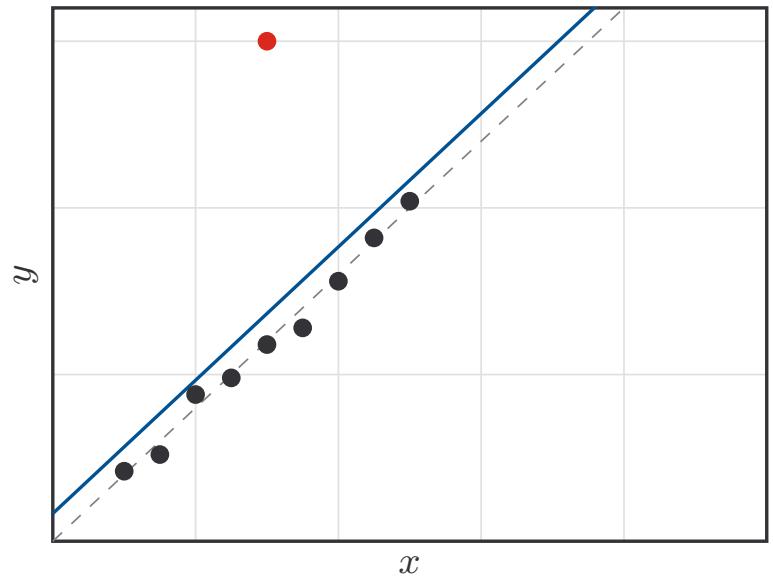
- Position: They can pull the regression line toward themselves, changing both slope and intercept
- Strength: They can weaken the correlation (reduce R^2), making predictions less reliable

Impact on Regression Model



Types of Influential Points





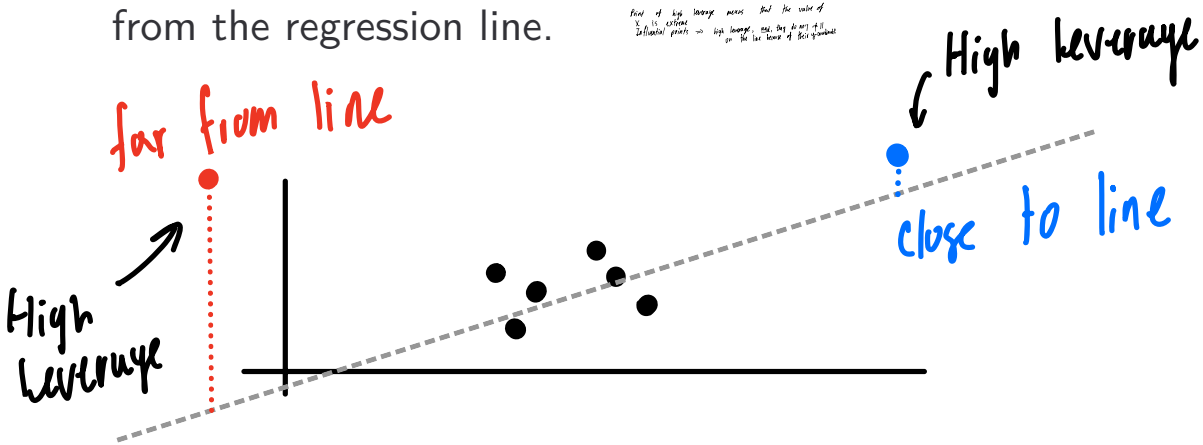
What Makes an Outlier Influential?

A data point (x_i, y_i) has high leverage if its x-coordinate is extreme

Influential Point

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

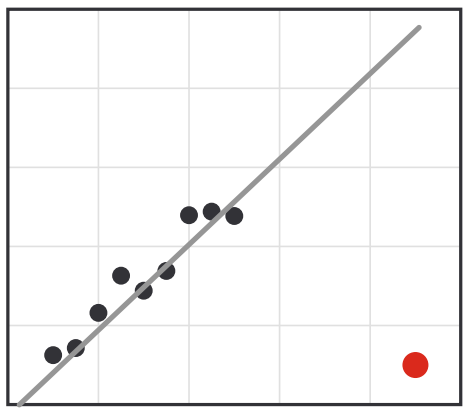
Generally speaking, this is a point that has an extreme value of x with a value of y that deviates from the regression line.



The blue and red points both have high leverage. However, the red point is influential while the blue is not.

Example 5.9: Identifying Influential Points

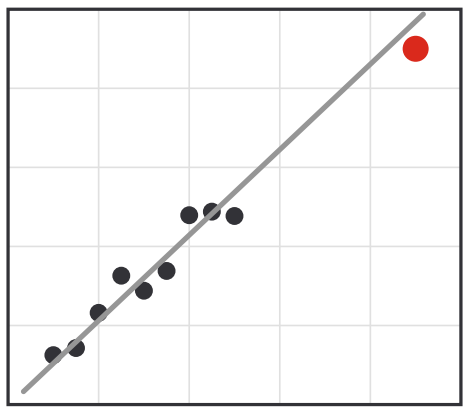
For each plot, indicate if the highlighted point is influential.



Influential

High leverage

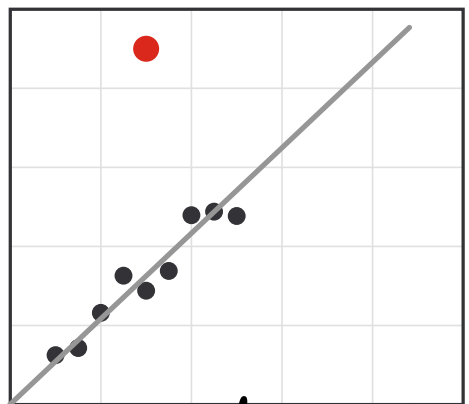
Point not on trend



Not influential

High leverage

Point lies on trend



Not influential

Low leverage

Point not on trend

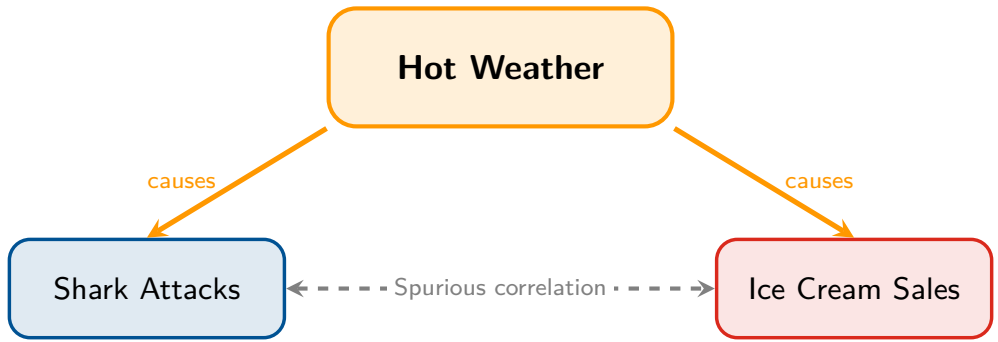
PART 8

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

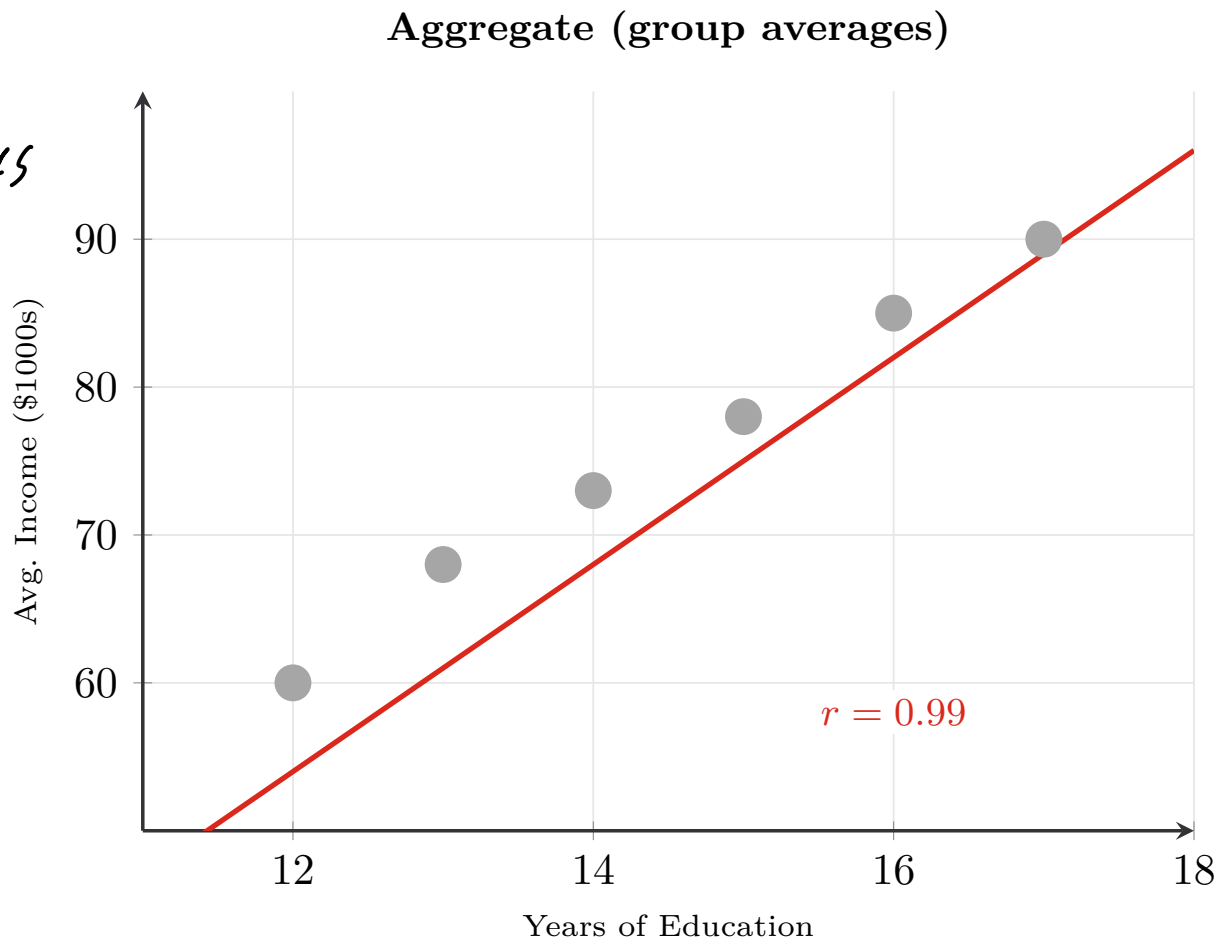
Ecological Fallacy

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

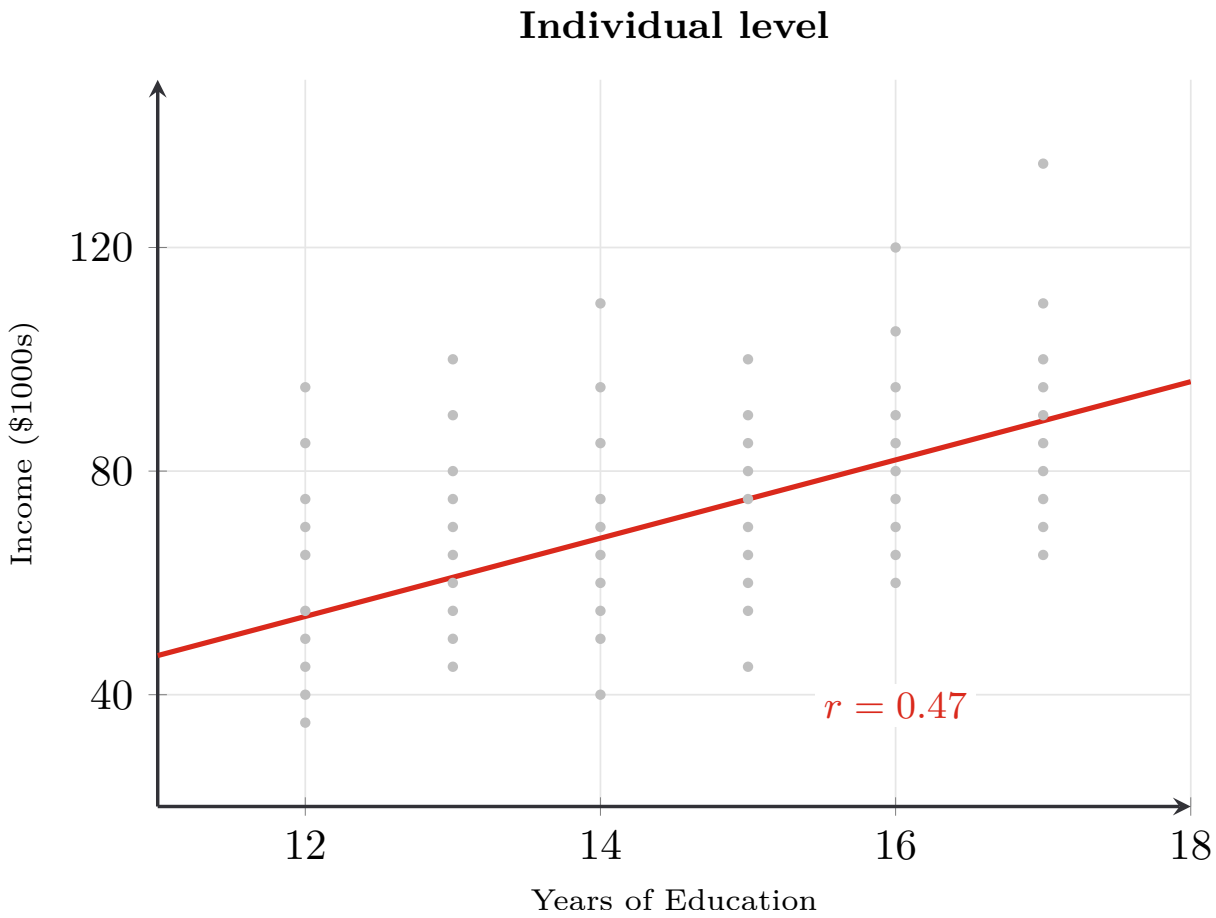
This is a special case of Simpson's paradox, which we will discuss in the next chapter.

Example 5.10: Education and Income

This obfuscates the variation within the groups.



Example 5.10: Education and Income (Continued)



Chapter Summary: Core Formulas

Least Squares Regression Line

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

where:

- **Slope:** $b = r \cdot \dfrac{s_y}{s_x}$
- **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:

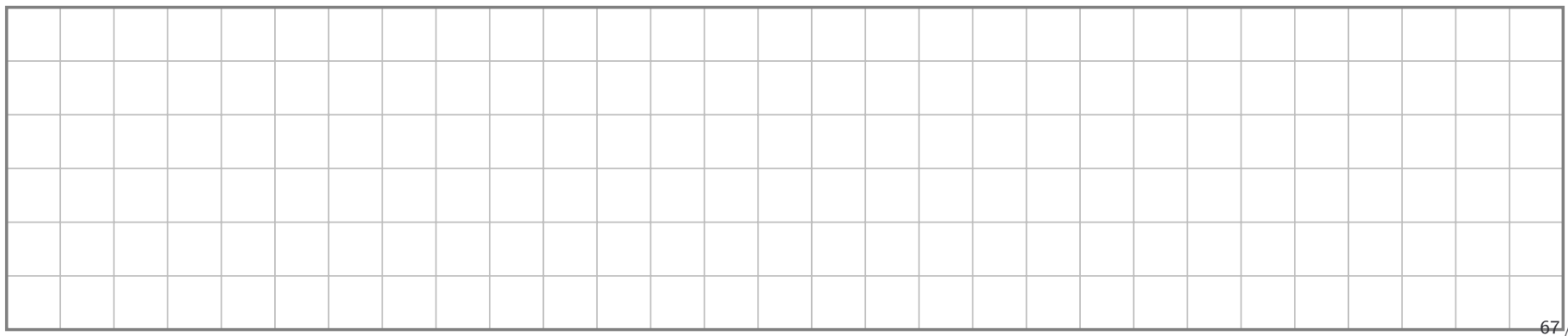
- $\bar{x} = 15$ dollars, $\bar{y} = 3.2$ GPA
- $s_x = 8.5$, $s_y = 0.6$
- $r = -0.68$

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?

a) $b = r\frac{s_y}{s_x}$, $a = \bar{y} - b\bar{x}$

$b = 0.82\left(\frac{12.5}{1.8}\right)$, $a = 72 - 5.69(7)$

$= 5.69$ $= 32.13$

$\Rightarrow \hat{y} = 32.13 + 5.69x$

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?

b) $x = 9$ $\quad \hat{y} = 32.13 + 5.69x$

$\Rightarrow \quad \hat{y} = 32.13 + 5.69(9) = 83.34$

d) Yes, this is reliable as the sleep quantity is close to the center of data, is admissible and the predicted value is plausible.

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?

c) $R^2 = \text{corr}(x,y)^2 = r^2 = (0.82)^2 = 0.67$

The LS LM regressing exam score on sleep explains 67% of the total variation in exam scores.

Example 5.13: Temperature and Coffee Sales

(x) (y)

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?

c) $\hat{y} = 280 - 8.5(20) = 110$

d) $\hat{y} = 280 - 8.5(-5) = 322.5$.
This prediction may not be reliable if -5°C is outside the range of observed temps. or is not a feasible temperature (for the coffee shop location).

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:

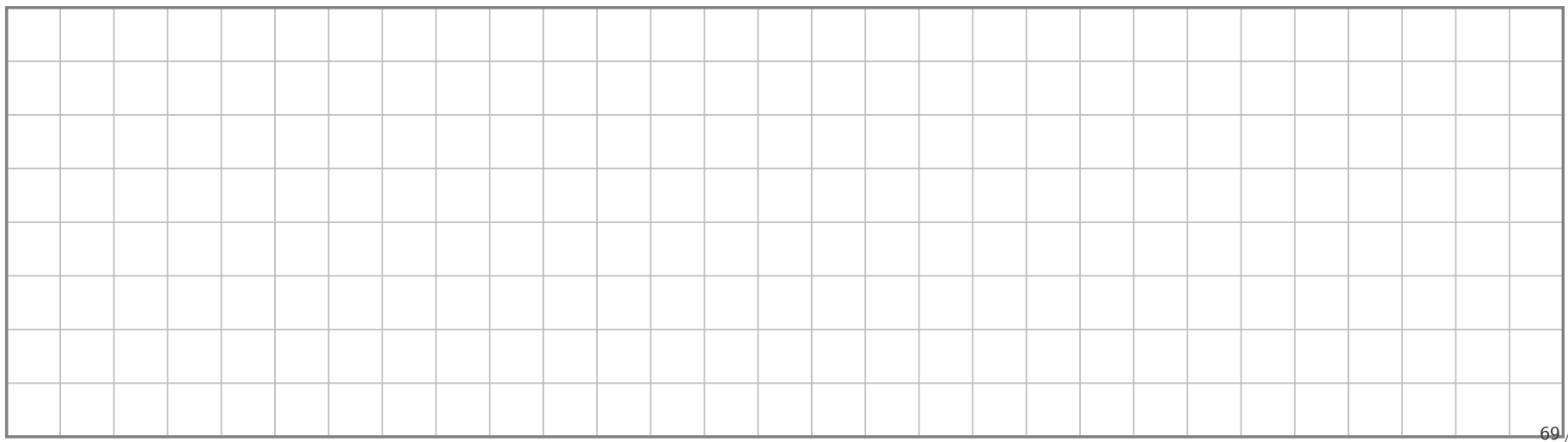
$$\hat{y} = 280 - 8.5x$$

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.

b) The athlete with residual -8.

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.

$b = r \cdot \frac{s_y}{s_x} \Rightarrow \frac{b \cdot s_x}{s_y} = r \Rightarrow R^2 = \left(\frac{b \cdot s_x}{s_y}\right)^2 = \left(\frac{1.02 \cdot 1.41}{1.68}\right)^2 \approx 0.73$

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.

We know that $e_i = 1$, $y_i = 74$. We want x_i.

Idea: e_i, y_i allows us to find $\hat{y}_i$

Knowing $\hat{y}_i$ allows us to find x_i

$$\hat{y}_i = y_i - e_i = 73$$

$$\hat{y}_i = a + bx_i$$

$$\Rightarrow 73 = 50 + 2.5x_i$$

$$23 = 2.5x_i$$

$$\Rightarrow x_i = \frac{23}{2.5} = 9.2$$

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.

(As sum of residuals is 0) we know that LS linear model crosses through $(\bar{x}, \bar{y})$.
Therefore, the predicted value of $\bar{x}$ coincides with $\bar{y}$ and:

$$66.66 = 50 + 2.5(\bar{x})$$
$$16.66 = 2.5\,\bar{x}$$
$$\Rightarrow \quad \bar{x} = 6.66$$

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.

Note: $\hat{y} = a + bx$ where $b = \text{corr}(x,y)\frac{s_y}{s_x} = r\frac{s_y}{s_x}$

Similarly, when we fit the inverse model:

$\hat{x} = A + By$ where $B = \text{corr}(y,x)\frac{s_x}{s_y} = \text{corr}(x,y)\frac{s_x}{s_y} = r\frac{s_x}{s_y}$

If we multiply the slopes, the std's cancel: $bB = \left(r\frac{s_y}{s_x}\right)\left(r\frac{s_x}{s_y}\right) = r^2$

So, a square root of bB is the correlation. Which one? The positive one as $b > 0$ where $s_x, s_y > 0 \Rightarrow r > 0$. Therefore, $r = \sqrt{2.5 \cdot 0.16} = \sqrt{0.4} \doteq 0.63$

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?

- Correlation is invariant to change of units.
- Slope would change as follows: for regressing height onto shoe size, our new slope would be

$$b' = r \cdot \frac{s_y}{s_x'} = r \frac{s_y}{s_x 2.54} = \frac{b}{2.54}$$

For the reverse relationship, the new slope would be

$$B' = r \frac{s_x'}{s_y} = r(2.54)\frac{s_x}{s_y} = 2.54\,B.$$