Chapter 4

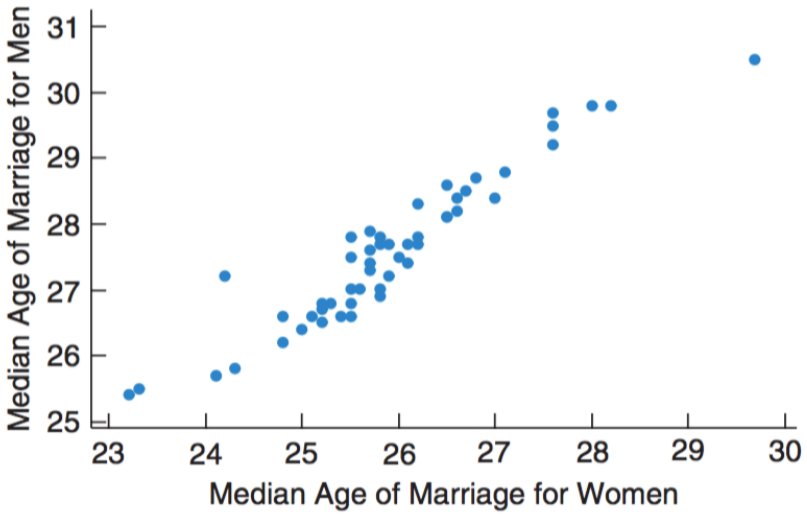
Regression Analysis:
Exploring Associations between Variables

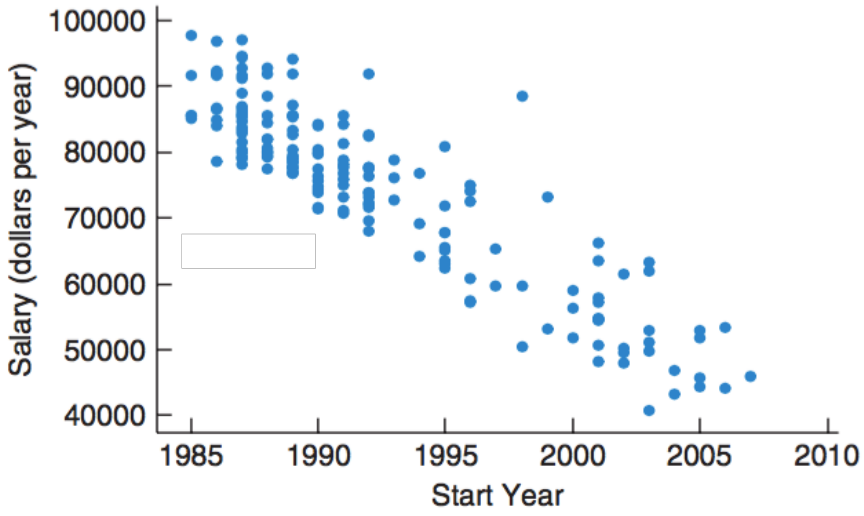
Scatterplots

- Scatterplots are the best way to start observing the relationship and the ideal way to picture associations between two quantitative variables.
- In a scatterplot, you can see patterns, trends, relationships, and even the occasional extraordinary value sitting apart from the others.
- The variable in the x-axis is called the explanatory (or independent) variable and the variable on the y-axis is called the response (or dependent) variable.
- When looking at scatterplots, we look for trend, shape, strength and unusual features.

Trend

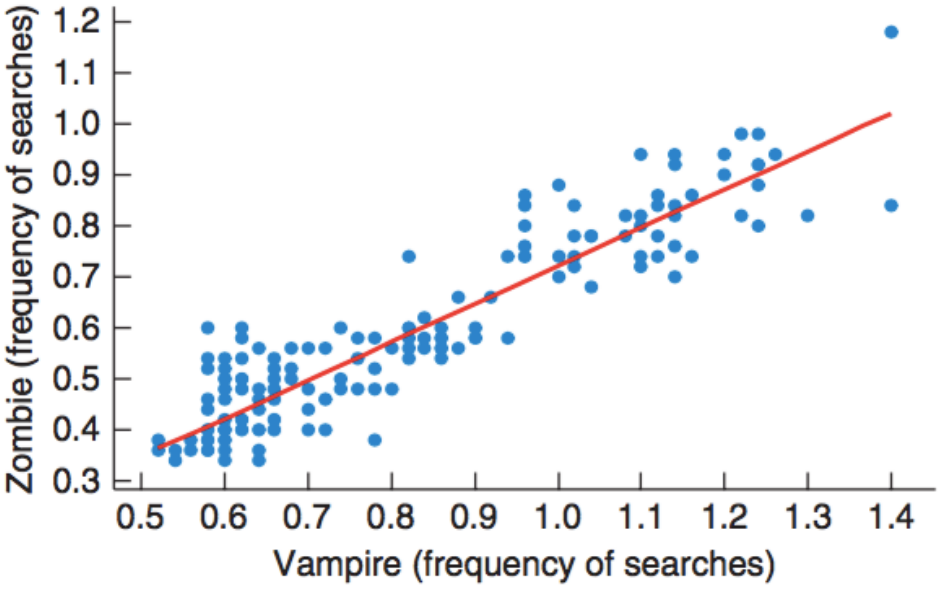
- If a trend is increasing or running from lower left to upper right (uphill) it has a positive direction (as x increases, y increases).
- If a trend is decreasing or running from upper left to lower right (downhill) it has a negative direction (as x increases, y decreases).

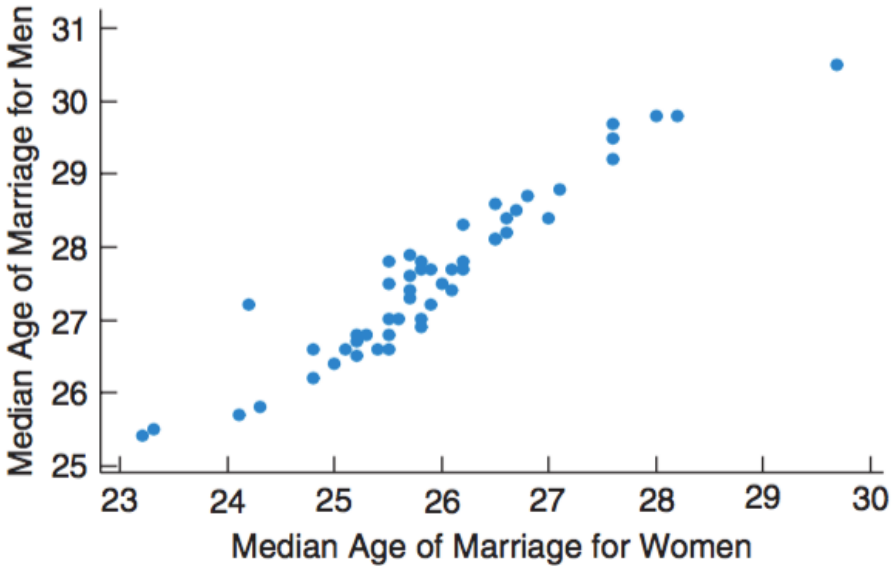




Shape

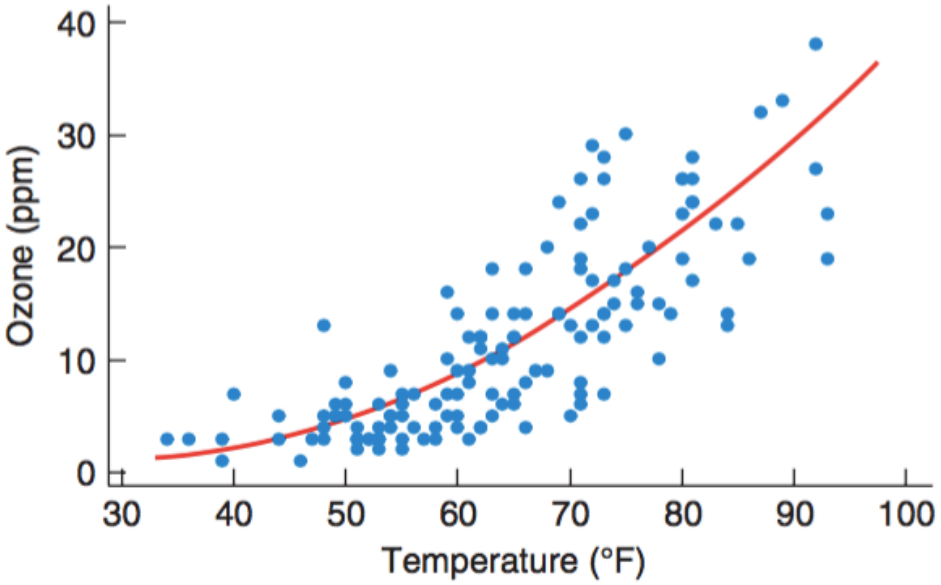
- If the points appear to follow a straight line with a negative or positive slope then we can assume that the two variables are linearly related or the trend is linear.

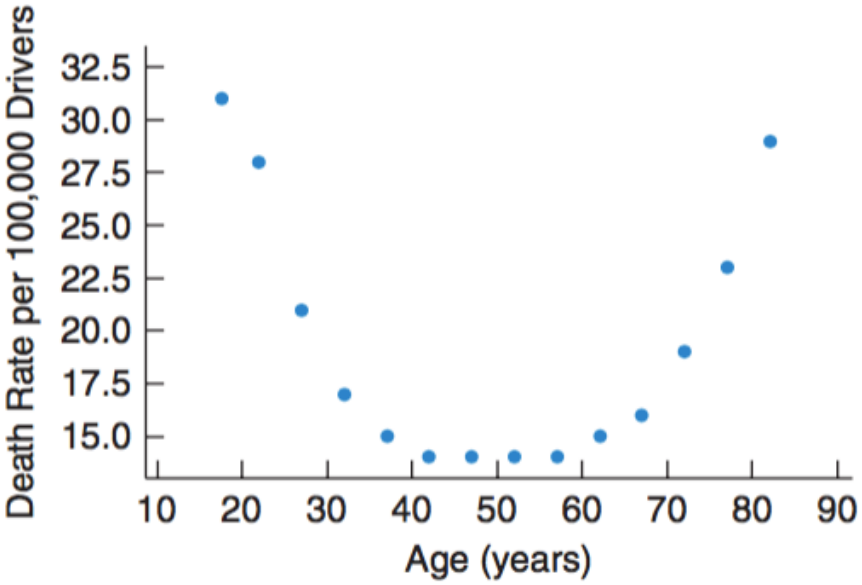




Shape

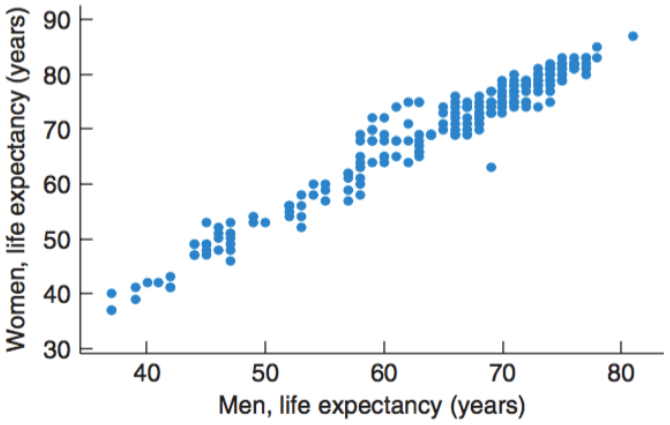
- If the points don't seem to follow a linear trend then the relationship is non-linear.

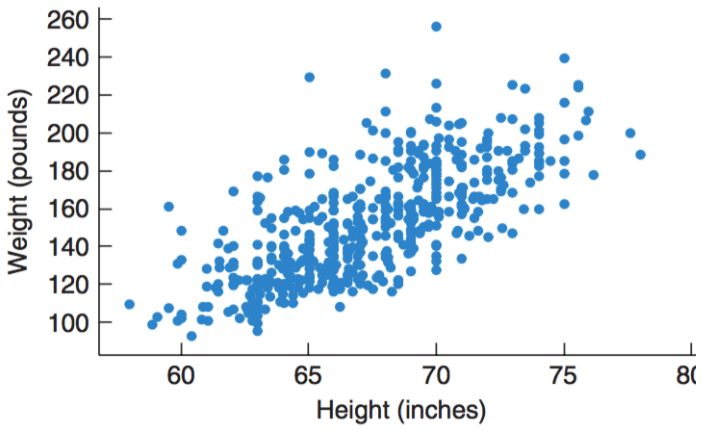


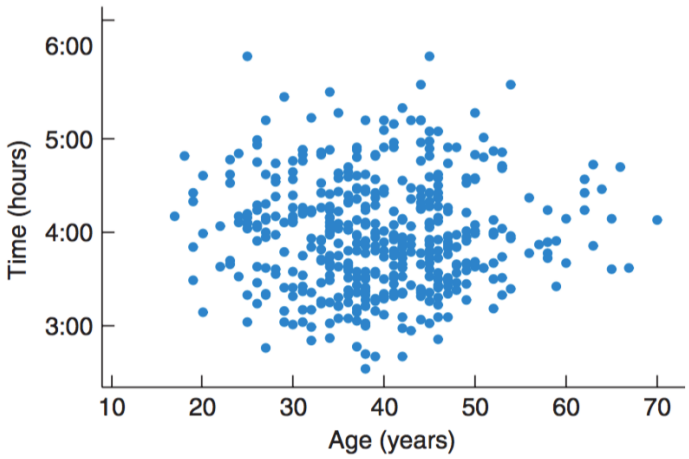


Strength

- If there doesn't appear to be a lot of scatter, there is a strong relationship between the two variables.
- If there appears to be some scatter, there is a weak relationship between the two variables.
- If there appears to be a lot of scatter, there is no relationship between the two variables.

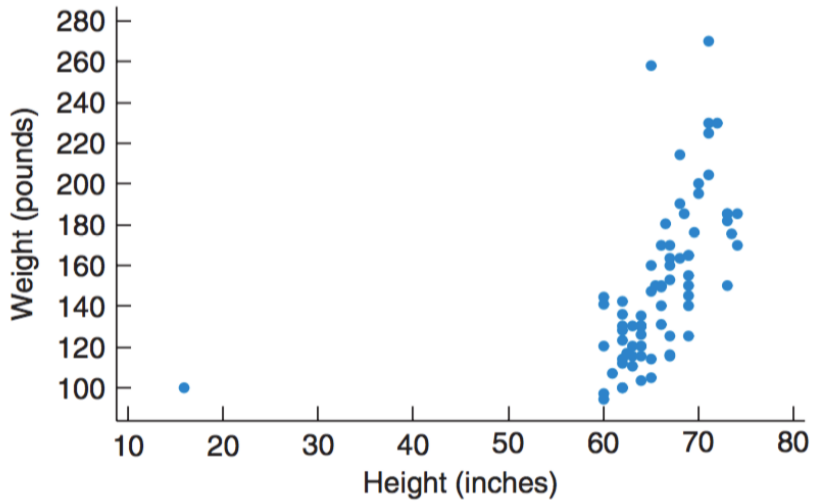






Unusual Features

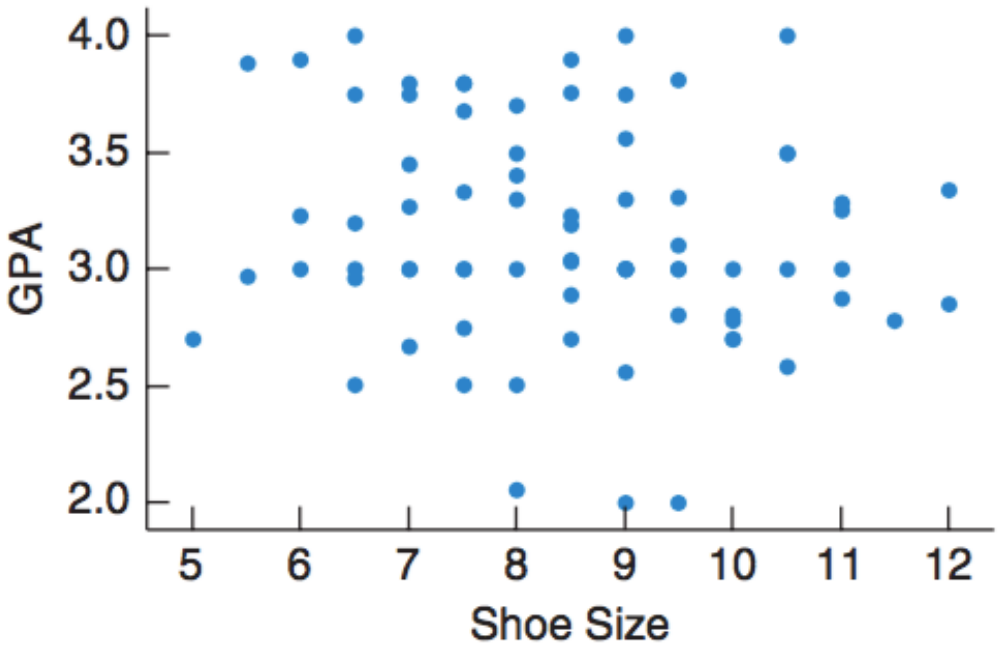
- Look for the unexpected; what you never thought to look for might be an interesting feature.
- One example of such a surprise is an outlier standing away from the overall pattern of the scatterplot.
- You should also look for clusters or subgroups.



Example

Clicker!

The figure shows a scatterplot of shoe size and GPA for some college students. Does it show an increasing trend, a decreasing trend or no trend?



Answer:

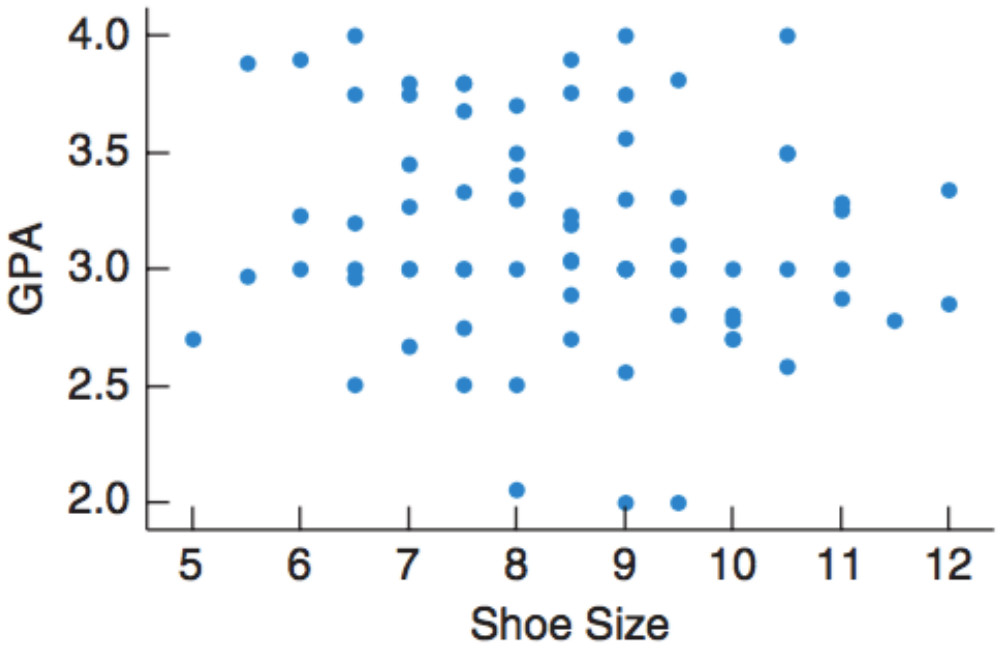
A. Increasing trend

B. Decreasing trend

C. No trend

Example

The figure shows a scatterplot of shoe size and GPA for some college students. Does it show an increasing trend, a decreasing trend or no trend?



Answer:

A. Increasing trend

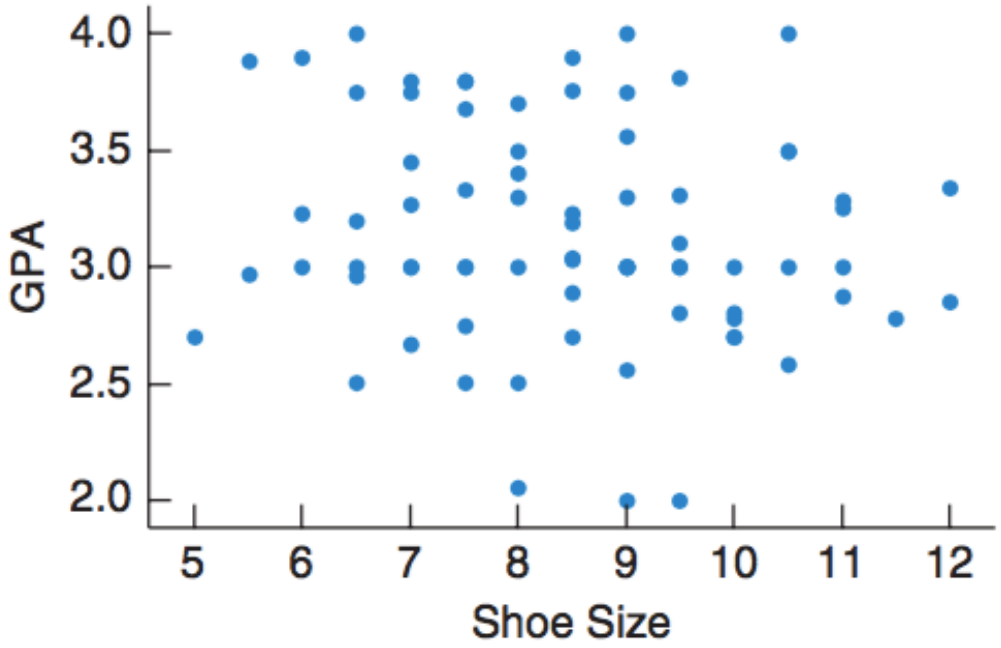
B. Decreasing trend

C. No trend

Example

Clicker!

The figure shows a scatterplot of shoe size and GPA for some college students. Is there a strong relationship?



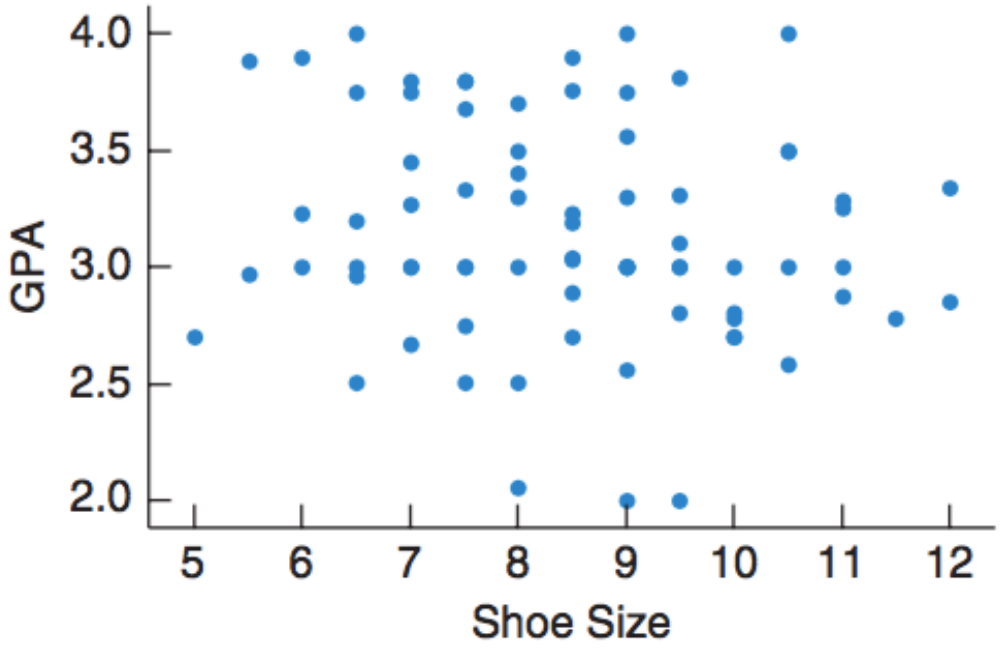
Answer:

A. Yes

B. No

Example

The figure shows a scatterplot of shoe size and GPA for some college students. Is there a strong relationship?



Answer:

A. Yes

B. No

Variables

- The explanatory variable (independent variable) is the variable that it is the predictor.
    - It goes on the x-axis.
- The response variable (dependent variable) is the variable of interest.
    - It is the outcome variable.
    - It goes on the y-axis.

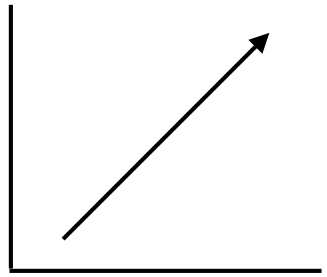
Correlation Coefficient

- The correlation coefficient (r) gives us a numerical measurement of the strength of the linear relationship between two numerical variables.
- It is also called the Pearson correlation coefficient.
- Correlation treats the two variables (x and y) symmetrically: The correlation of x and y is the same as the correlation of y and x.
- Correlation only makes sense if the trend is linear.

Correlation Properties

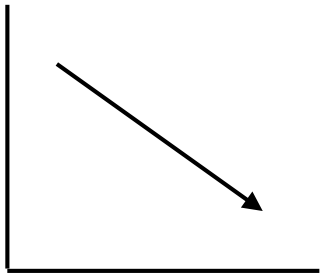
- Correlation is always between -1 and +1.
    - Correlation can be exactly equal to -1 or +1, but these values are unusual in real data because this means that all the data falls exactly on a single straight line.
    - A correlation near zero corresponds to a weak linear association.
- The sign of a correlation coefficient gives the direction of the association.
- Correlation has no units.

Positive Correlation



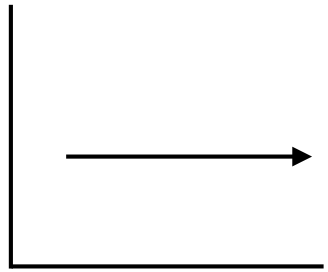
- We have positive correlation when r is greater than 0 ($r > 0$).
- The closer the correlation is to 1, the stronger the association between the two variables.
- The closer the correlation is to 0, the weaker the association between the two variables.

Negative Correlation



- We have negative correlation when r is less than 0 ($r < 0$).
- The closer the correlation is to -1, the stronger the association between the two variables.
- The closer the correlation is to 0, the weaker the association between the two variables.

No Correlation

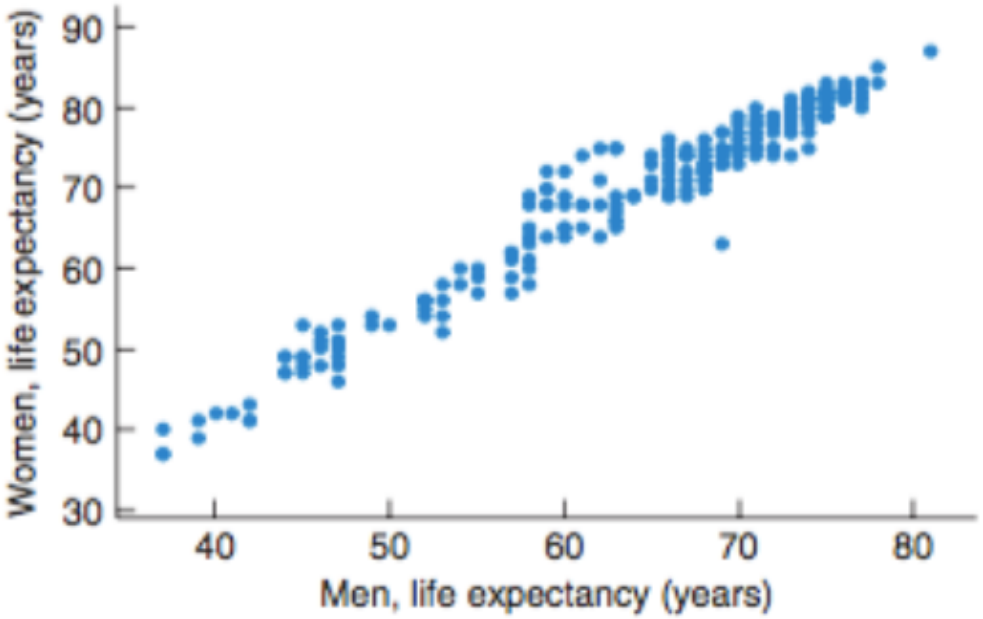


- We have zero or no correlation when r is equal to 0 (r = 0).
- There is no relationship or association between the two variables.The x-axis variable has no affect on the y-axis variable.
- Points are randomly scattered on the grid. Even though a line can be drawn through these points, this type of scatterplot still shows no correlation.

Example

Clicker!

Is there a positive, negative or no correlation?



Answer:

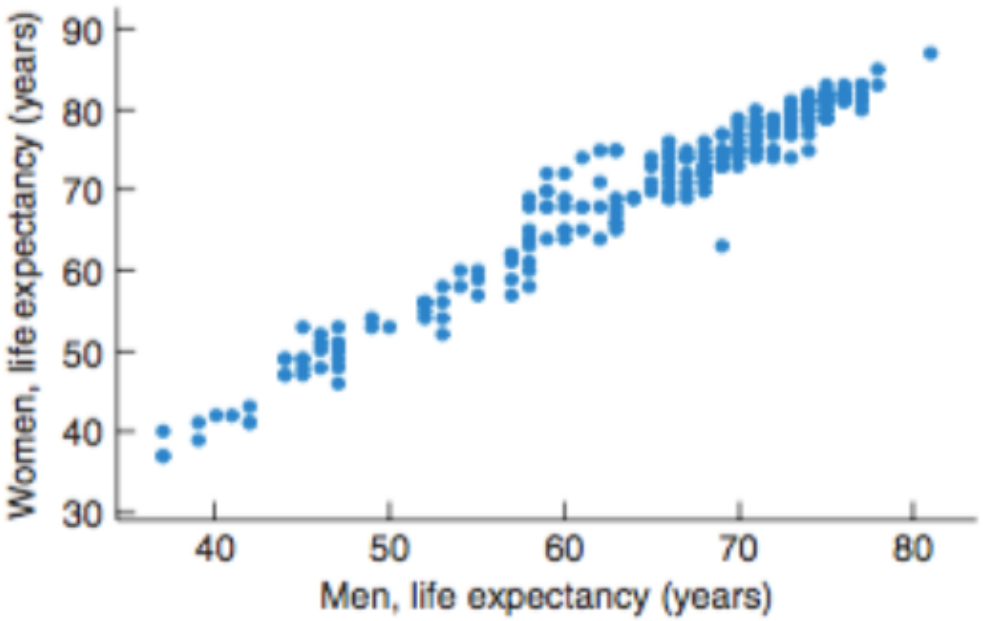
A. Positive

B. Negative

C. No

Example

Is there a positive, negative or no correlation?



Answer:

A. Positive

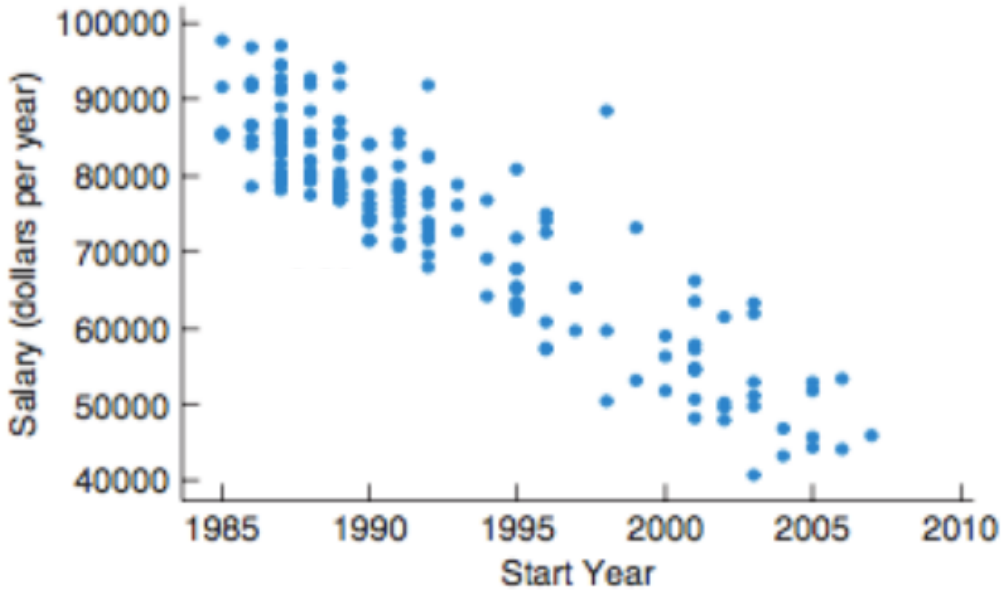
B. Negative

C. No

Example

Clicker!

Is there a positive, negative or no correlation?



Answer:

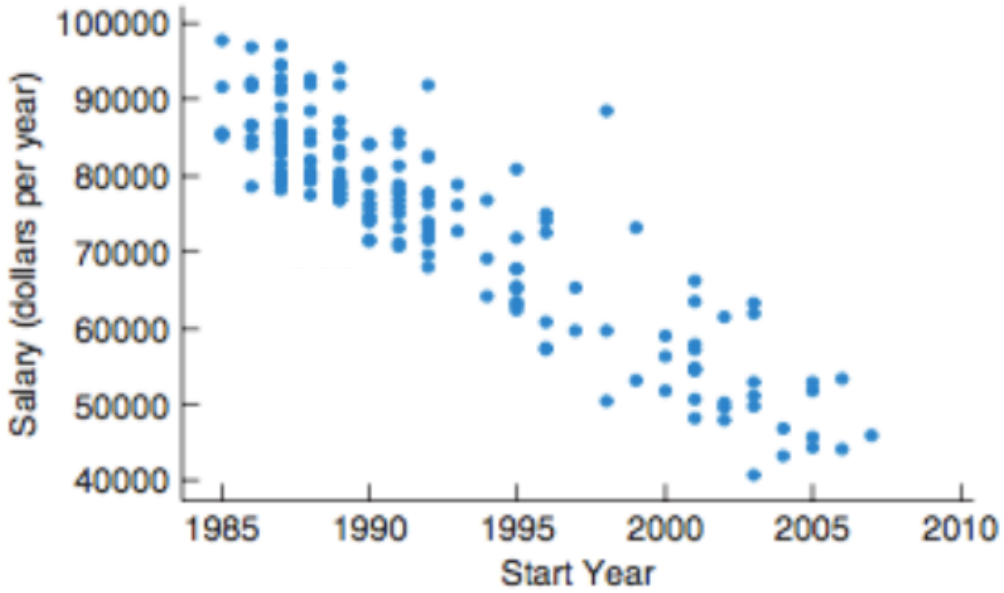
A. Positive

B. Negative

C. No

Example

Is there a positive, negative or no correlation?



Answer:

A. Positive

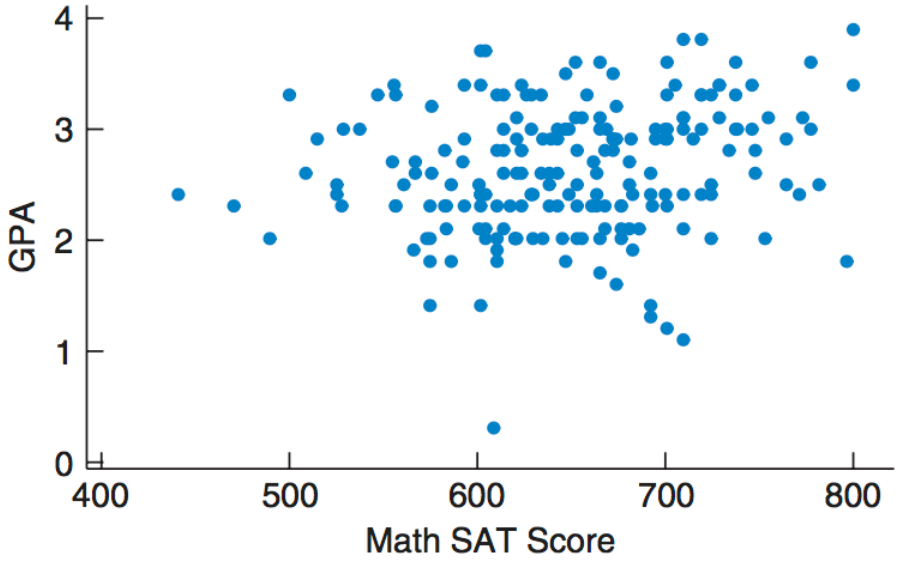
B. Negative

C. No

Example

Clicker!

Is there a positive, negative or no correlation?



Answer:

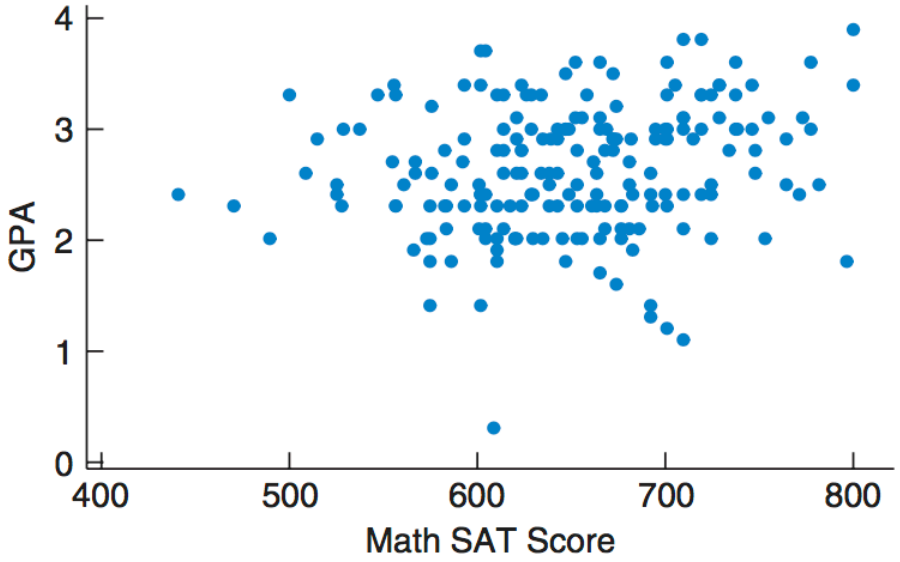
A. Positive

B. Negative

C. No

Example

Is there a positive, negative or no correlation?

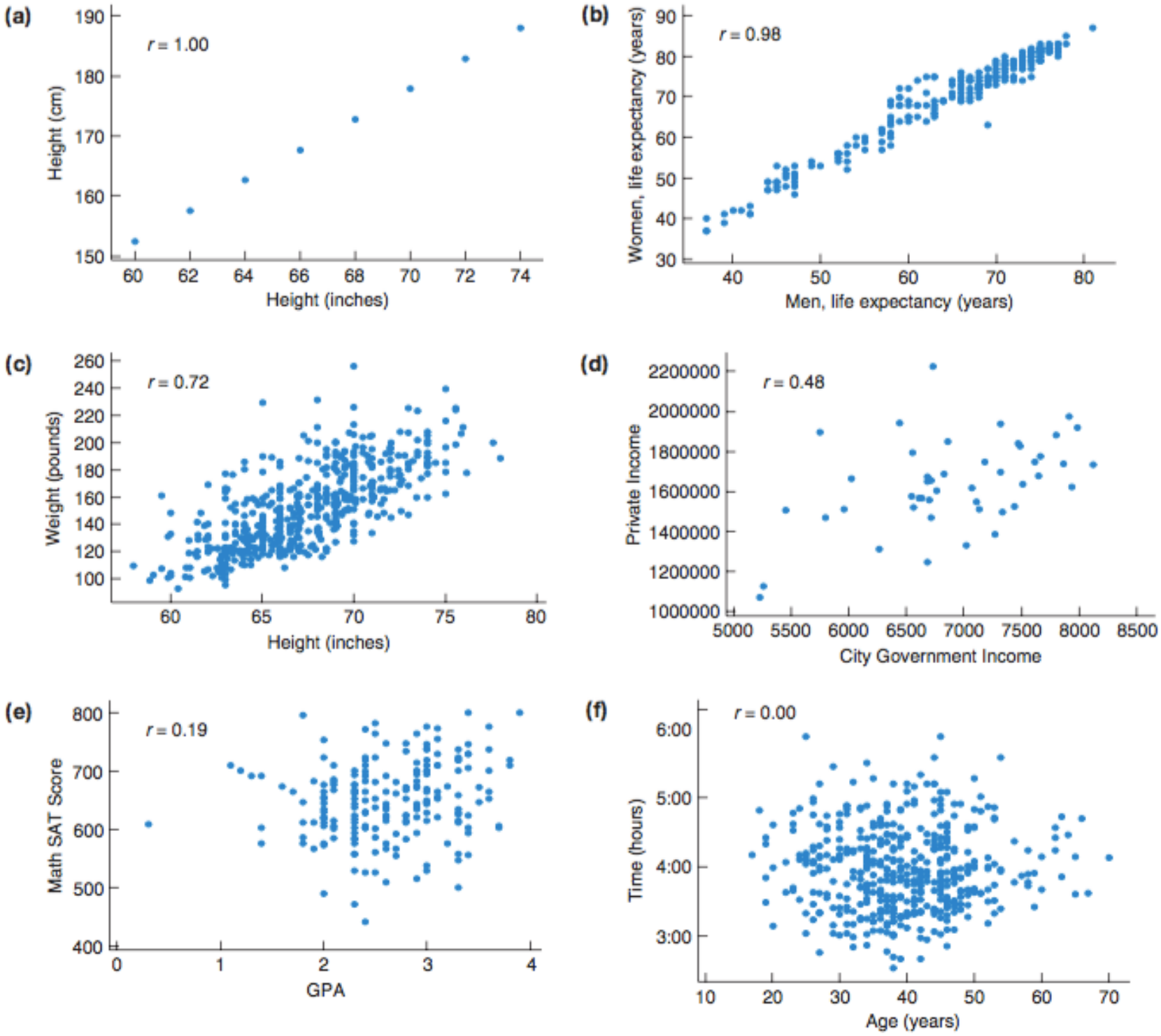


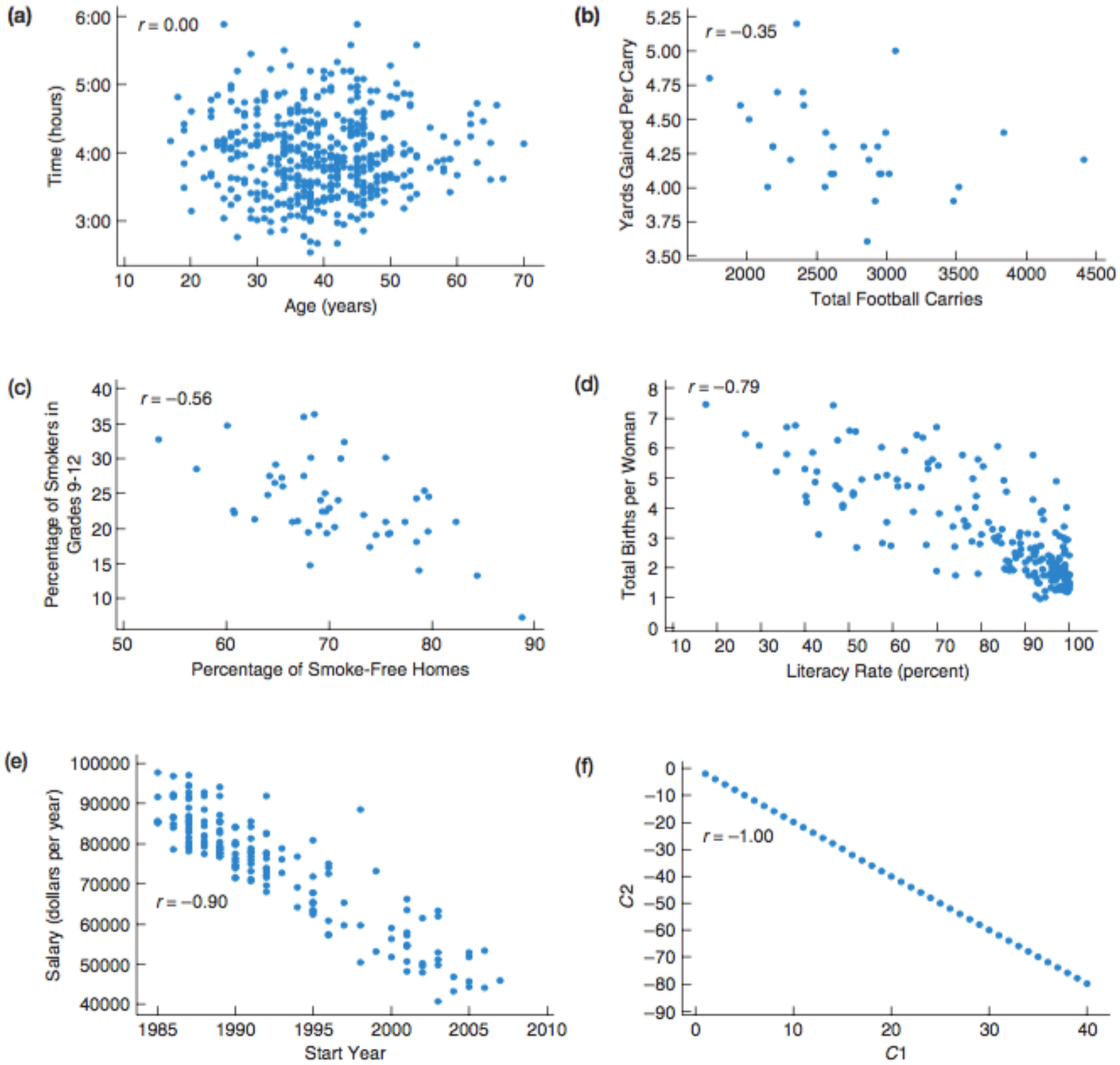
Answer:

A. Positive

B. Negative

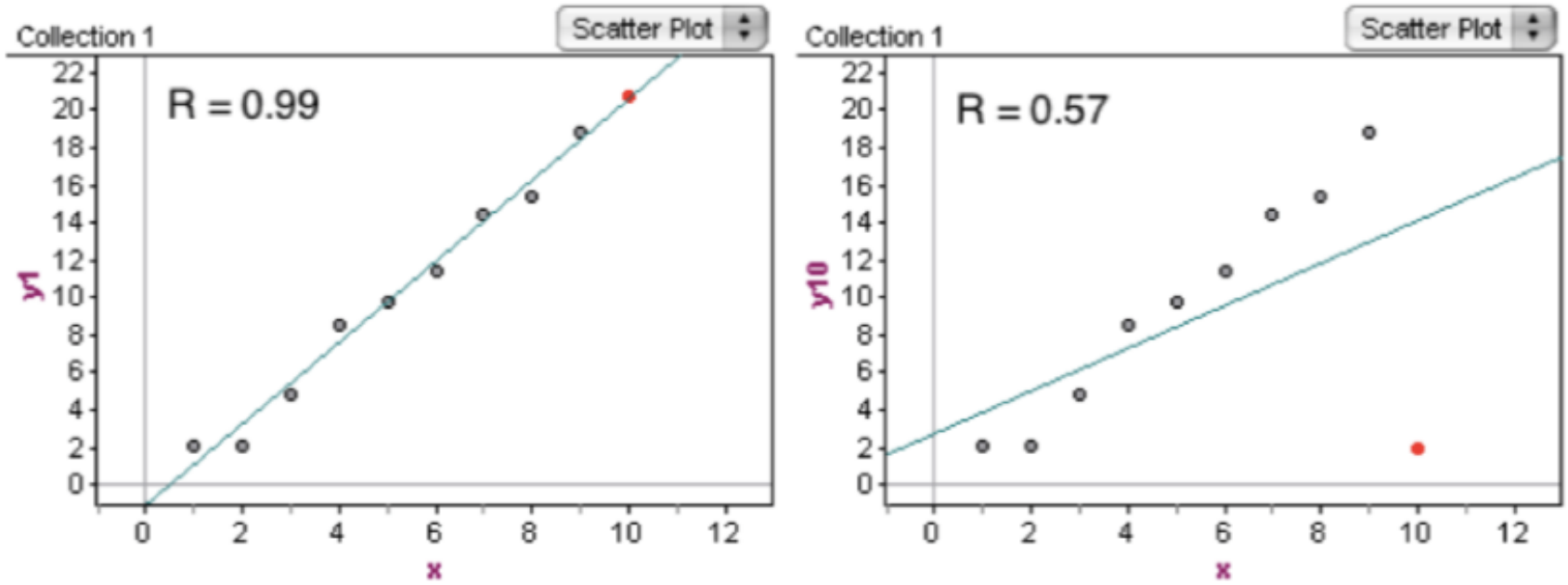
C. No


(a)
r = 1.00
Height (cm)
190
180
170
160
150
60
62
64
66
68
70
72
74
Height (inches)
(b)
r = 0.98
Women, life expectancy (years)
90
80
70
60
50
40
30
40
50
60
70
80
Men, life expectancy (years)
(c)
r = 0.72
Weight (pounds)
260
240
220
200
180
160
140
120
100
60
65
70
75
80
Height (inches)
(d)
r = 0.48
Private Income
2200000
2000000
1800000
1600000
1400000
1200000
1000000
5000
5500
6000
6500
7000
7500
8000
8500
City Government Income
(e)
r = 0.19
Math SAT Score
800
700
600
500
400
0
1
2
3
4
GPA
(f)
r = 0.00
Time (hours)
6:00
5:00
4:00
3:00
10
20
30
40
50
60
70
Age (years)


(a)
r = 0.00
Time (hours)
6:00
5:00
4:00
3:00
10
20
30
40
50
60
70
Age (years)
(b)
r = −0.35
Yards Gained Per Carry
5.25
5.00
4.75
4.50
4.25
4.00
3.75
3.50
2000
2500
3000
3500
4000
4500
Total Football Carries
(c)
r = −0.56
Percentage of Smokers in Grades 9-12
40
35
30
25
20
15
10
50
60
70
80
90
Percentage of Smoke-Free Homes
(d)
r = −0.79
Total Births per Woman
8
7
6
5
4
3
2
1
0
10
20
30
40
50
60
70
80
90
100
Literacy Rate (percent)
(e)
r = −0.90
Salary (dollars per year)
100000
90000
80000
70000
60000
50000
40000
1985
1990
1995
2000
2005
2010
Start Year
(f)
r = −1.00
C2
0
−10
−20
−30
−40
−50
−60
−70
−80
−90
0
10
20
30
40
C1

Correlation and Outliers

- Correlation is sensitive to outliers.
- Depending on your data, a single extreme point can change the correlation by a lot.



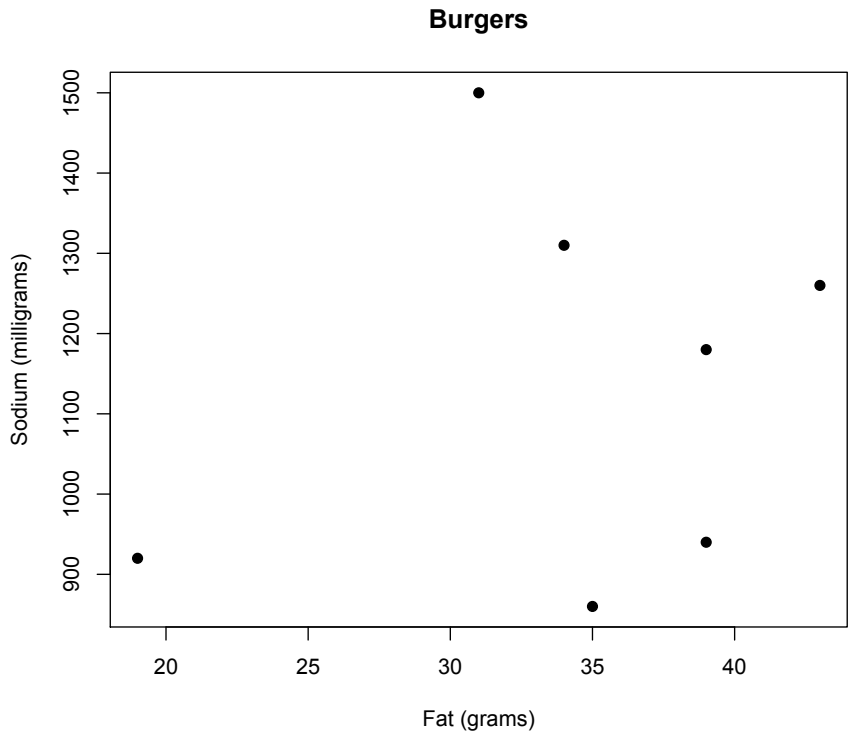
Correlation Does Not Mean Causation

- As ice cream sales increase so do shark attacks. Does that mean an increase in ice cream sales causes an increase in shark attacks?
- A high correlation between two variables does not imply causation.
- Do not conclude that a cause-and-effect relationship between two variables exists just because there is a strong correlation.

Example

- Fast food is often considered unhealthy because much of it is both high in fat and sodium. But are the two related?
- There does not appear to be a relationship between sodium and fat content in burgers. The correlation of 0.199 shows a weak relationship between the two variables.

Burgers	Fat (grams)	Sodium (milligrams)
1	19	920
2	31	1500
3	34	1310
4	35	860
5	39	1180
6	39	940
7	43	1260



Correlation

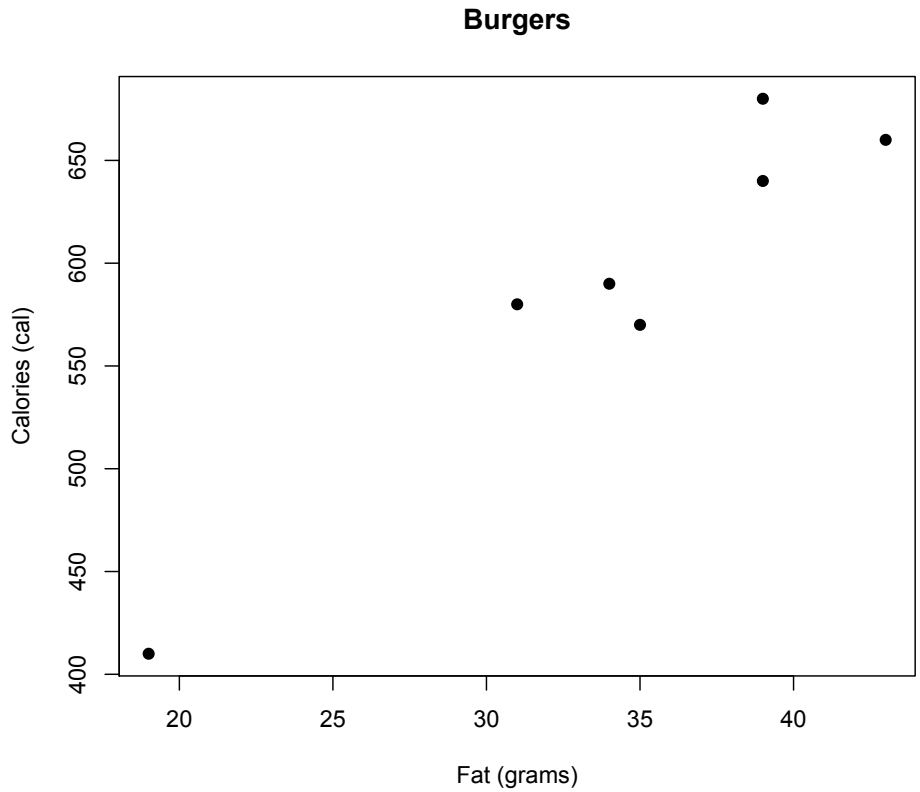
	Fat
Sodium	0.199

Example

Clicker!

- What is the relationship between fat and calories?
  - A. Strong, negative relationship
  - B. Strong, positive relationship
  - C. Weak, negative relationship
  - D. Weak, positive relationship

Burgers	Fat (grams)	Calories (cal)
1	19	410
2	31	580
3	34	590
4	35	570
5	39	640
6	39	680
7	43	660



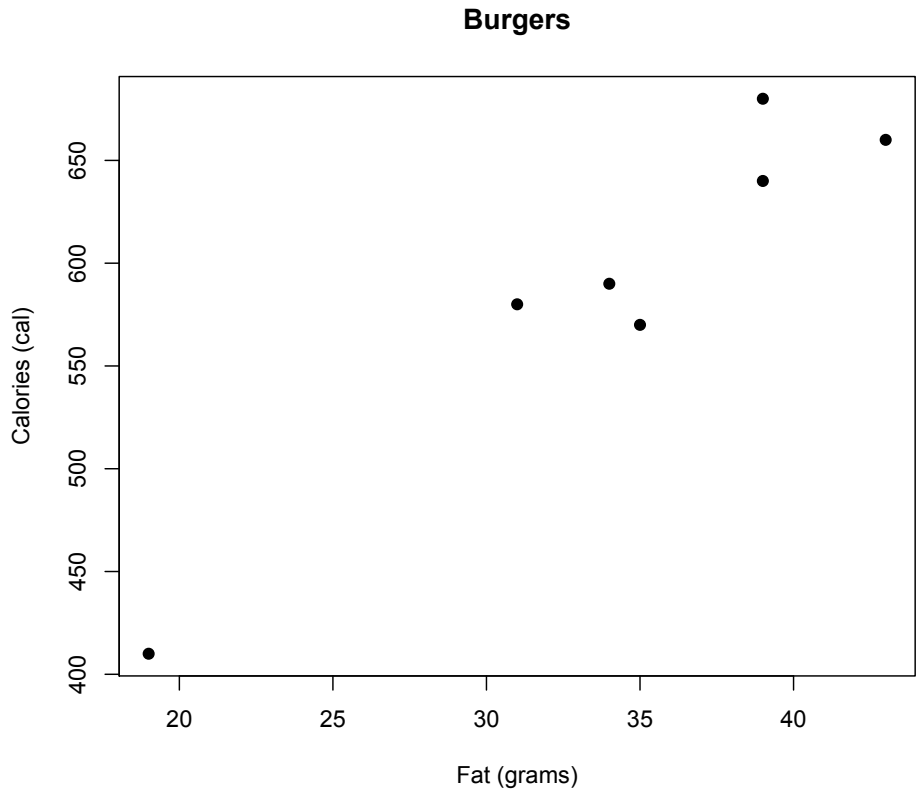
Correlation

	Fat
Calories	0.961

Example

- What about fat and calories?
- There appears to be a strong, positive linear relationship between fat and calories. The correlation of 0.962 supports this conclusion. But there appears to be an outlier at 410 calories and 19 grams of fat.

Burgers	Fat (grams)	Calories (cal)
1	19	410
2	31	580
3	34	590
4	35	570
5	39	640
6	39	680
7	43	660



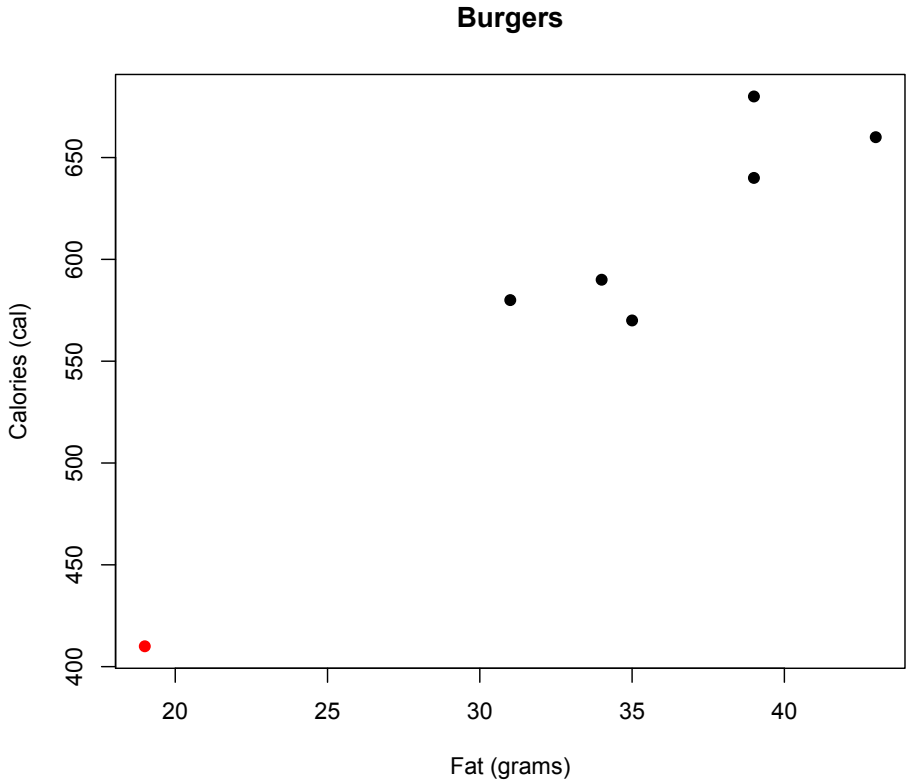
Correlation

	Fat
Calories	0.961

Example

- If we remove the outlier, the correlation is 0.837. Even without the outlier at 410 calories and 19 grams of fat the correlation is strong.
- Hence, we can conclude fat and calories are positively associated.

Burgers	Fat (grams)	Calories (cal)
1	19	410
2	31	580
3	34	590
4	35	570
5	39	640
6	39	680
7	43	660



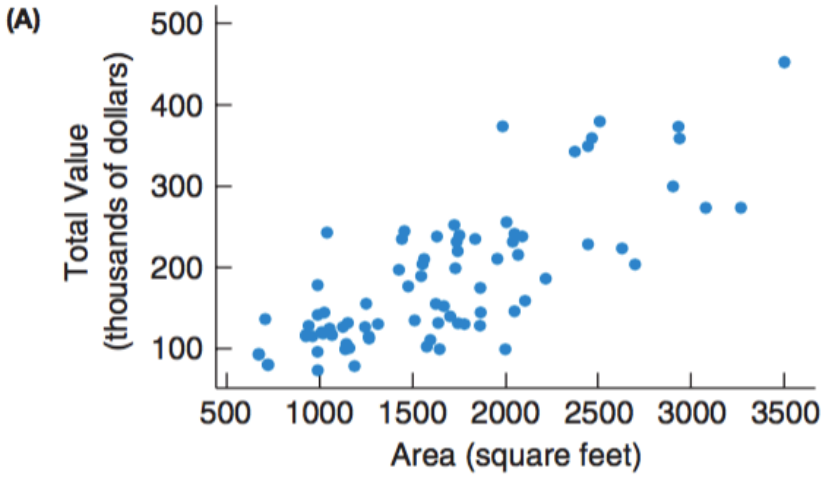
Correlation

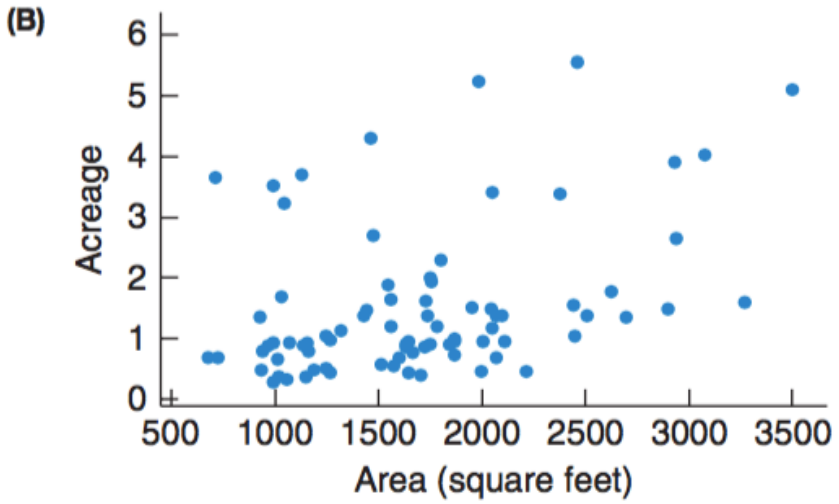
	Fat
Calories	0.837

Practice

Clicker!

- Match the plots to the following correlations: 0.767, 0.299, -0.980



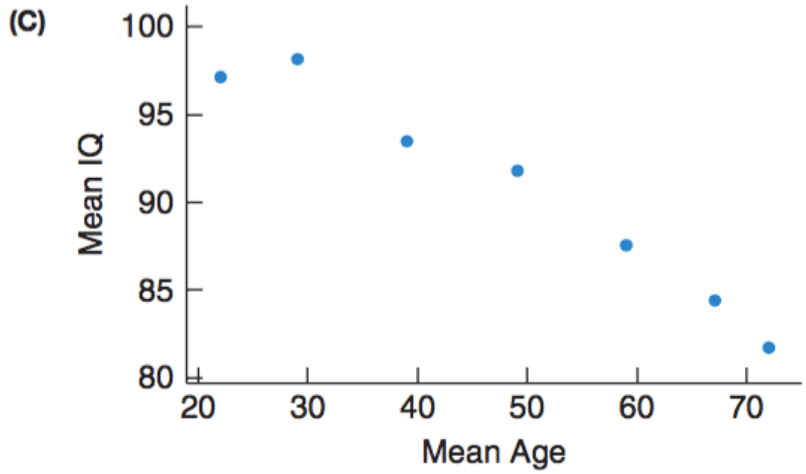


Answer:

A. (A) 0.299, (B) 0.767, (C) -0.980

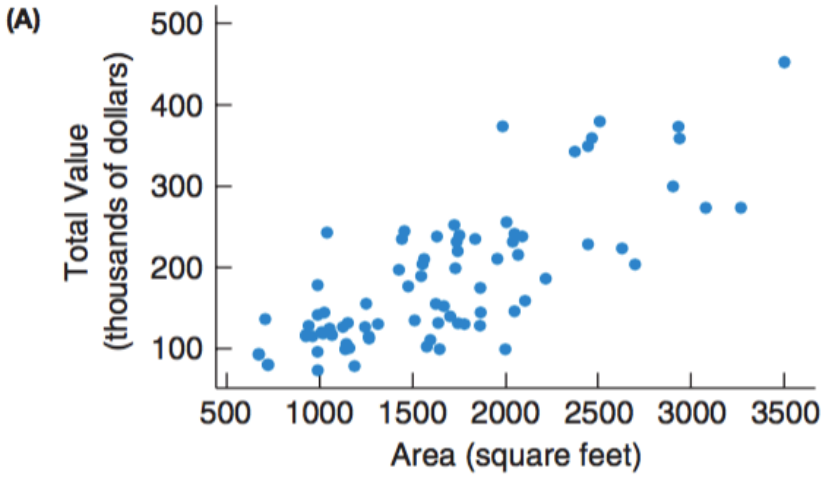
B. (A) 0.767, (B) 0.299, (C) -0.980

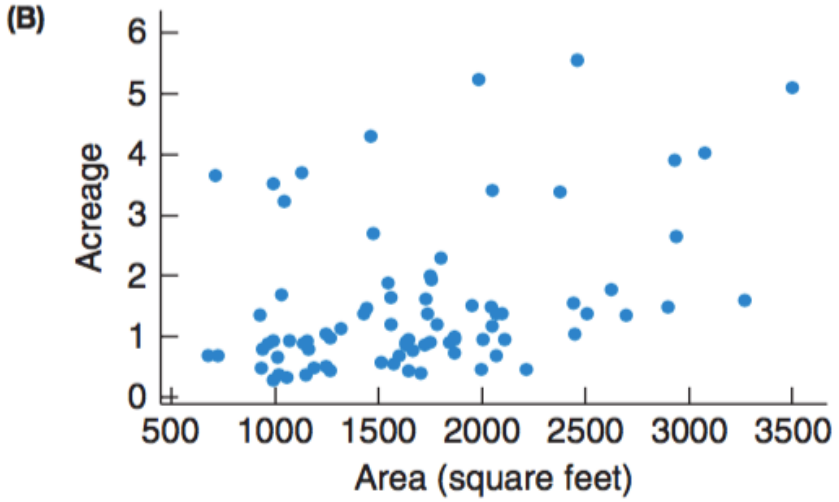
C. (A) -0.980, (B) 0.767, (C) 0.299



Practice

- Match the plots to the following correlations: 0.767, 0.299, -0.980



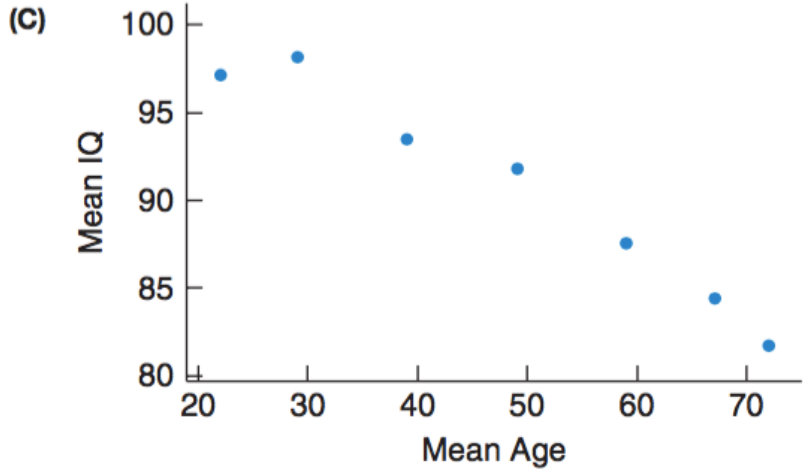


Answer:

A. (A) 0.299, (B) 0.767, (C) -0.980

B. (A) 0.767, (B) 0.299, (C) -0.980

C. (A) -0.980, (B) 0.767, (C) 0.299



Finding the Correlation Coefficient

- The correlation coefficient is determined through technology.
- We need to convert each observation to a z-score.

$$r = \frac{\sum z_x z_y}{n-1}$$

- z_x is the z-score for the x-variable, z_y is the z-score for the y-variable, n is the sample size.

Understanding the Correlation Coefficient

- Changing the order of the variables does not change the correlation coefficient (r).
- Adding a constant or multiplying by a positive constant does not affect r.
- The correlation coefficient is unitless.
- Must have a linear trend.

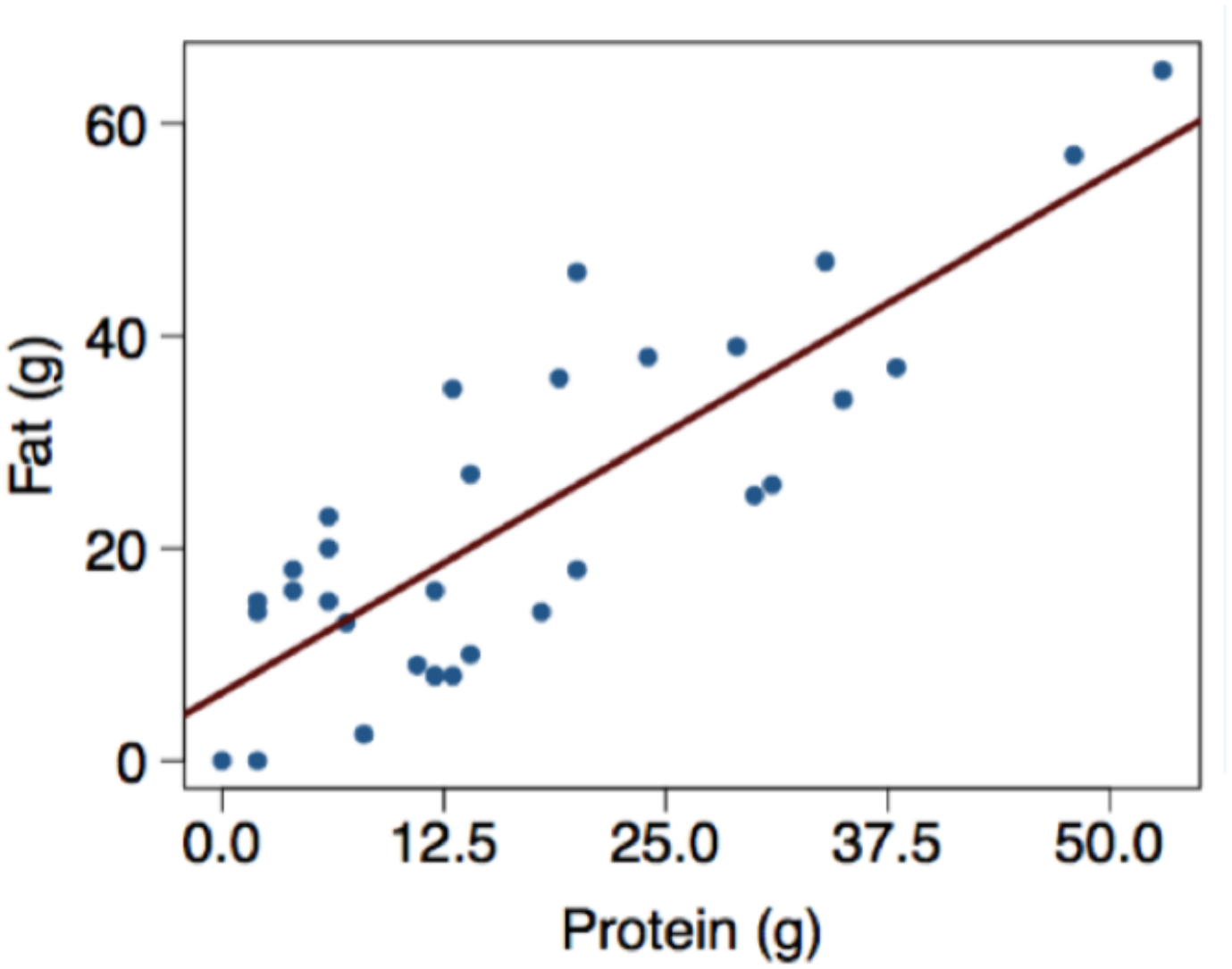
Linear Model

- Correlation says “there seems to be a linear association between these two variables” but it doesn’t tell us what that association is.
- We can say more about the linear relationship between two quantitative variables with a model.
- A model simplifies reality to help us understand underlying patterns and relationships.

Linear Model

- The linear model is just an equation of a straight line through the data.
    - The points in the scatterplot don't all line up, but a straight line can summarize the general pattern.
    - The linear model can help us understand how the explanatory (independent) and response (dependent) variables are associated.

Linear Model



Regression Line

- The regression line is a tool for making predictions about future observations.
- It is a useful method for summarizing a linear relationship.
- It is given by an equation for a straight line.

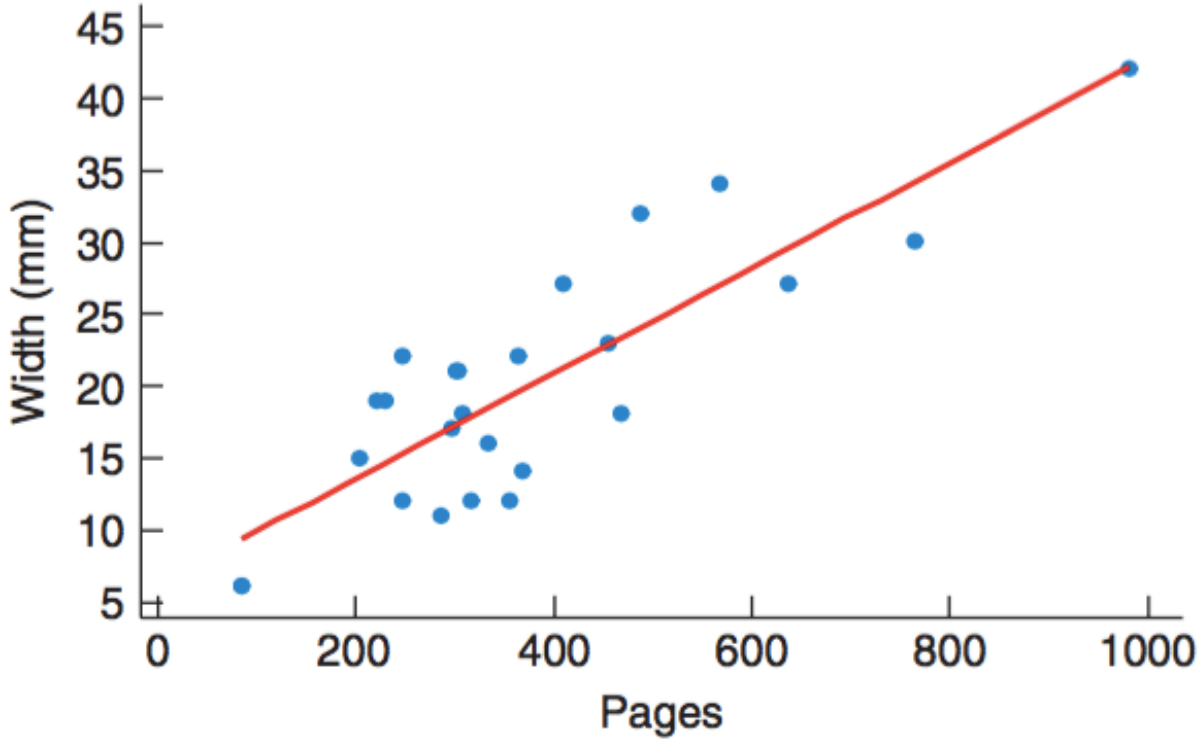
$$y = a + bx$$

where y is the y-variable, x is the x-variable, a is the intercept and b is the slope.

Visualizing the Regression Line

- Can we know how wide a book is based on the number of pages in the book?
- The equation for the regression line is

$$\text{Predicted Width} = 6.22 + 0.0366 \text{ Pages}$$

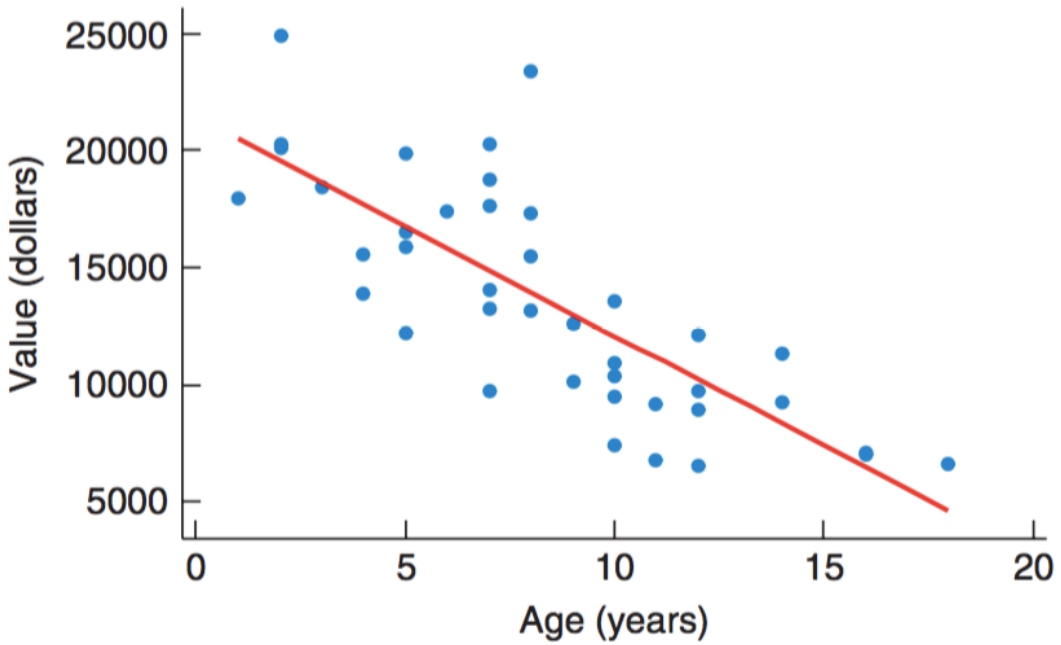


The regression line summaries the relationship between the width of the book and the number of pages for a small sample of books.

Regression in Context

- The regression line can be used to make predictions.
- Suppose you have a 10 year old car and you want to estimate how much it is worth.
- Age and Value have a linear relationship and the equation of the line is

$$\text{Predicted Value} = 21375 - 1215\,\text{Age}$$



Regression in Context

- We can use this equation to predict approximately how much a 10 year old car is worth:

$$\begin{aligned}\text{Predicted Value} &= 21375 - 1215 \times 10 \\ &= 21375 - 12150 \\ &= 9225\end{aligned}$$

- The regression line predicts that a 10 year old car will be valued at about $9225.
- Of course there are more factors to consider than just the age.

Example

Clicker!

A college instructor with far too many books on his shelf is wondering whether he has room for one more. He has about 20 mm of space left on his shelf and he can tell from the online bookstore that the book he wants has 630 pages. The regression line is

$$\text{Predicted Width} = 6.22 + 0.0366\ \text{Pages}$$

What's the predicted width?

Answer:

A. 17043

B. 29.28

C. 20

Example

Will the book fit on his shelf?

Answer: No the book will not fit.

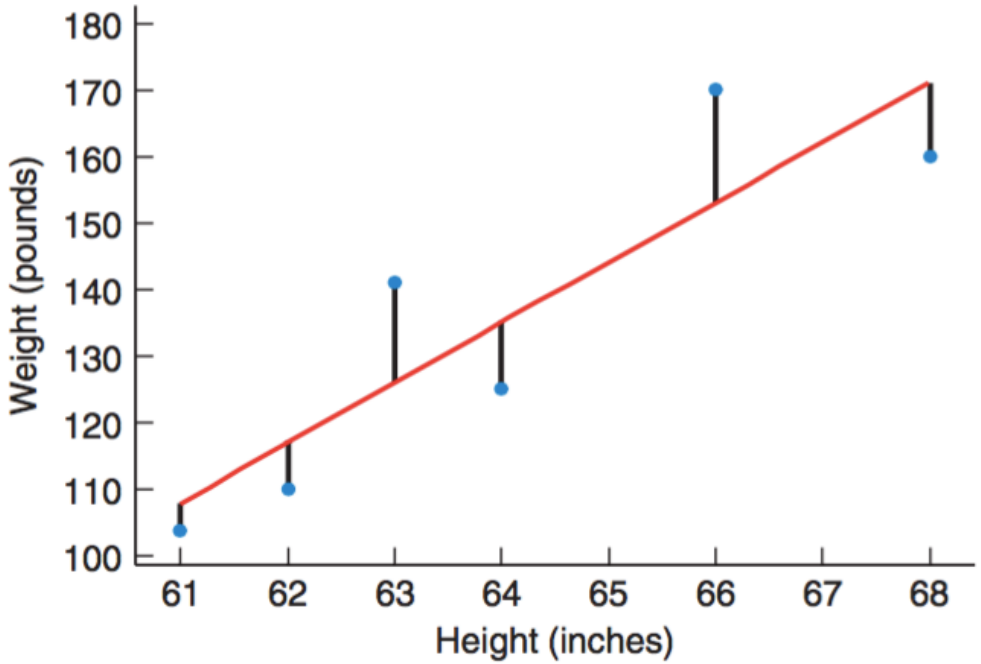
$$\begin{aligned}\text{Predicted Width} &= 6.22 + 0.0366 \times 630 \\ &= 6.22 + 23.068 \\ &= 29.278\text{mm}\end{aligned}$$

Finding the Regression Line

- To find the regression line we use technology.
- To understand how technology finds the regression line, try drawing a line through a scatterplot to best capture the linear trend.
- The regression line comes closest to most of the points in the scatterplot.
- Sometimes, it is called the “best fit” line because it provides the best fit to the data.

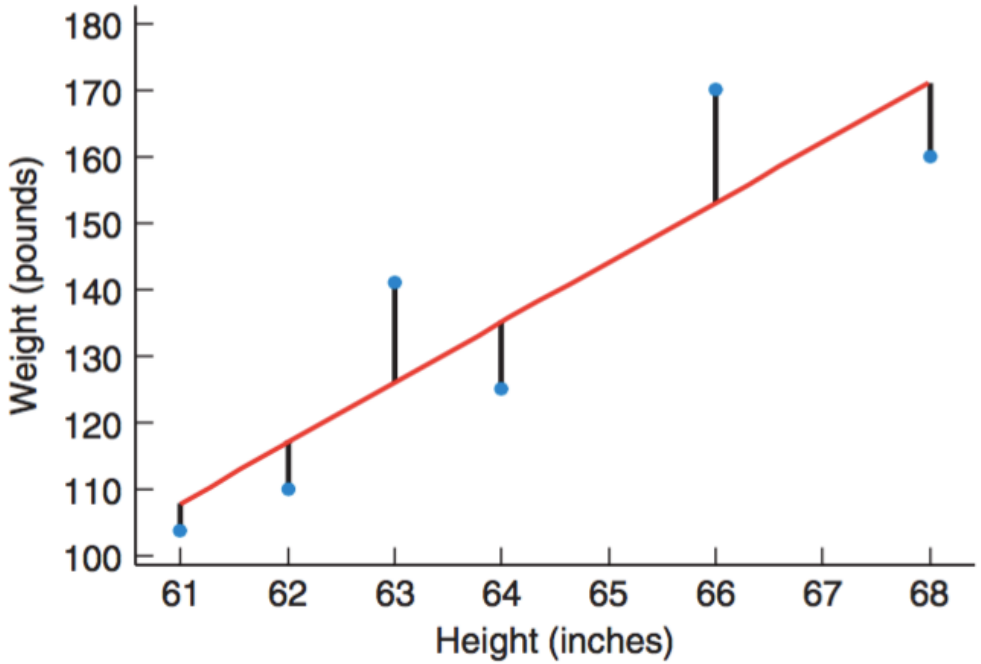
Finding the Regression Line

- On average, the sum of the squares of the vertical distances between the points or observed y-values (y) and the value predicted by the line ($\hat{y}$) is the smallest for the regression line. $(y - \hat{y})^2$
- Also called the least squares line.

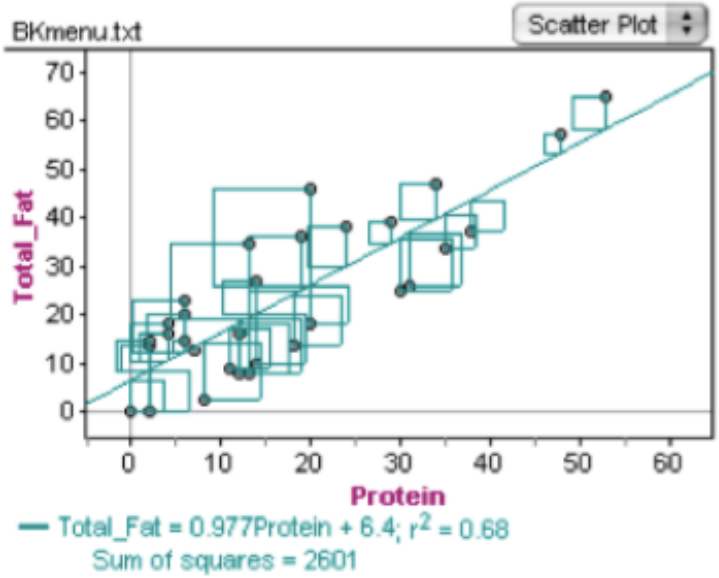


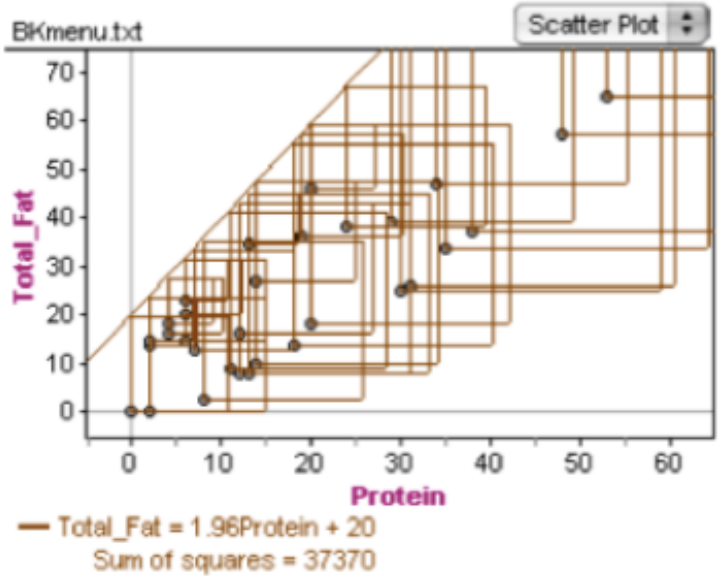
Residuals

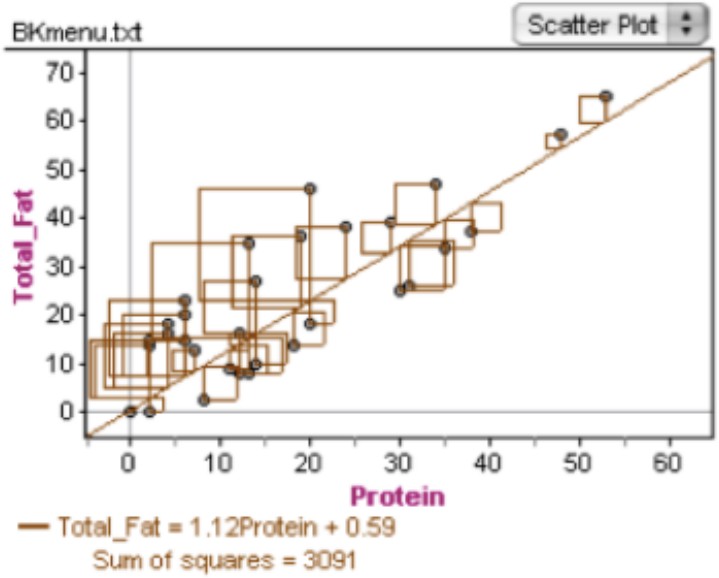
- $\hat{y}$ represents the **predicted value** or the estimate made from a model.
- **Residual** is the difference between the observed value and its associated predicted value.
- Residual = observed value - predicted value = $y - \hat{y}$

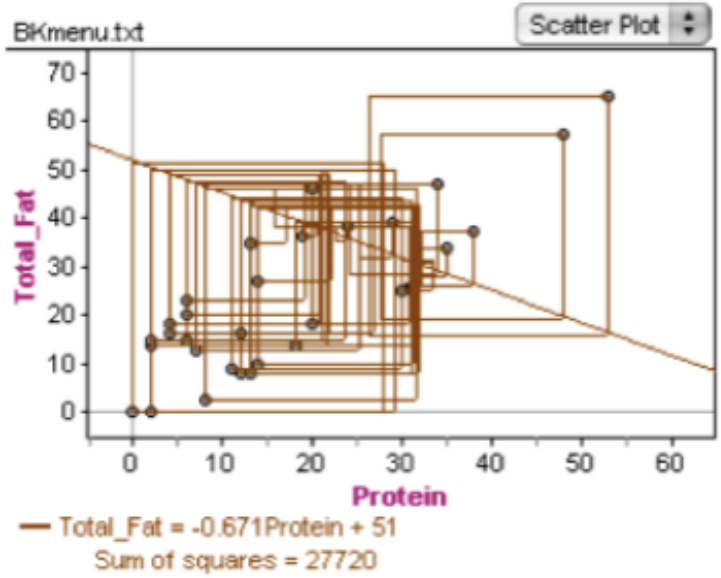


Fathom Example









Slope

- The slope can be calculated using the correlation coefficient, r, and the standard deviations of the explanatory variable (x) and the response variable (y).

$$\text{slope} = \text{r}\frac{s_y}{s_x}$$

- Interpretation of the slope: For each unit increase in x, we expect y to increase/decrease on average by the value of the slope.
- A positive slope implies that as x increases y increases.
- A negative slope implies that as x increases y decreases.
- When correlation (r) is positive slope (b) will be positive and when correlation (r) is negative slope (b) will be negative.

Example

Clicker!

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. If the correlation between protein and fat contents is 0.83, calculate the slope of the linear model for predicting fat content from protein content.

$$\text{slope} = r\frac{s_y}{s_x} = 0.83 \times \frac{16.4}{14} = 0.97$$

What's the y-variable?

Answer:

A. Fat content

B. Protein content

Example

Clicker!

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. If the correlation between protein and fat contents is 0.83, calculate the slope of the linear model for predicting fat content from protein content.

$$\text{slope} = r\frac{s_y}{s_x} = 0.83 \times \frac{16.4}{14} = 0.97$$

What's the x-variable?

Answer:

A. Fat content

B. Protein content

Example

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. If the correlation between protein and fat contents is 0.83, calculate the slope of the linear model for predicting fat content from protein content.

$$\text{slope} = \text{r}\frac{s_y}{s_x} = 0.83 \times \frac{16.4}{14} = 0.97$$

- We are interested in the linear model for predicting fat content from protein content.
- y-variable: fat content
- x-variable: protein content

Example

Clicker!

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. If the correlation between protein and fat contents is 0.83, calculate the slope of the linear model for predicting fat content from protein content.

$$\text{slope} = r\frac{s_y}{s_x} = 0.83 \times \frac{16.4}{14} = 0.97$$

Interpret the slope in context.

A. For one gram increase in protein content, we would expect the fat content to increase on average by 0.97 grams.

B. For one gram increase in fat content, we would expect the protein content to increase on average by 0.97 grams.

Example

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. If the correlation between protein and fat contents is 0.83, calculate the slope of the linear model for predicting fat content from protein content.

$$\text{slope} = \text{r}\frac{s_y}{s_x} = 0.83 \times \frac{16.4}{14} = 0.97$$

Interpret the slope in context.

For one gram increase in protein content, we would expect the fat content to increase on average by 0.97 grams.

Intercept

- Once we have the slope, we can calculate the intercept.

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

- We need to find the means of variables x and y.
- Interpretation of the intercept: When x equals 0, we expect y to equal the intercept.

Example

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. Previously we calculated the slope to be 0.97. Now, calculate the intercept.

$$\text{intercept} = \bar{y} - b\bar{x}$$

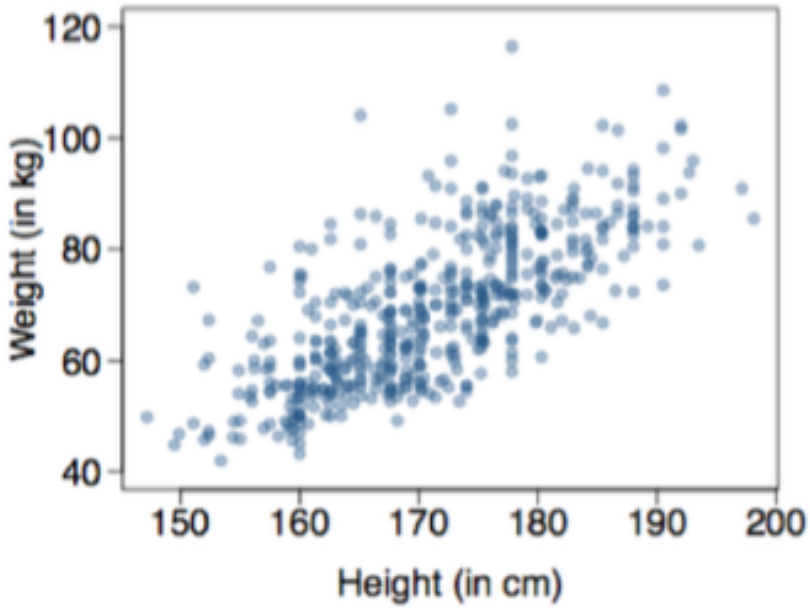
$$\text{intercept} = 23.5 - 0.97 \times 17.2 = 6.8$$

Interpret the intercept in context.

Burger King menu items with no protein are expected to have a fat content of 6.8 grams.

Intercept

- Sometimes the intercept by itself may not make sense. In these cases the intercept serves only to adjust the height of the line and is meaningless by itself.



$$\widehat{weight} = -105.0113 + 1.0176 * height$$

People who are 0 centimeters tall are expected to weigh -105.0113 kilograms. This is obviously not possible.

Burger King Example

- Let's look at the Burger King menu items example again. We found that the slope is 0.97 and intercept is 6.8. We were using the protein content to predict the fat content.

- What's the equation for the regression line?

Burger King Example

- Let's look at the Burger King menu items example again. We found that the slope is 0.97 and intercept is 6.8. We were using the protein content to predict the Fat content.

- What's the equation for the regression line?

$$\text{Predicted Fat} = 6.8 + 0.97 \times \text{Protein}$$

Regression Line: Order Matters

- Unlike the correlation coefficient, the order matters with regression.
- Referring back to the Burger King menu items example. What if we had switched the variables? We want to predict the amount of protein by looking at the amount of fat.
- To do this, we would switch protein with fat. Protein would be our y-variable and fat would be our x-variable.
- We would need to recalculate the slope and intercept.

Regression Line: Order Matters

Clicker!

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. The correlation between protein and fat contents is 0.83.

- We are interested in the linear model for predicting protein content from fat content.
- What's the y-variable?

Answer:

A. Fat content

B. Protein content

Regression Line: Order Matters

Clicker!

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. The correlation between protein and fat contents is 0.83.

- We are interested in the linear model for predicting protein content from fat content.
- What's the x-variable?

Answer:

A. Fat content

B. Protein content

Regression Line: Order Matters

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. The correlation between protein and fat contents is 0.83.

- We are interested in the linear model for predicting protein content from fat content.
- y-variable: protein content
- x-variable: fat content

Regression Line: Order Matters

Clicker!

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. The correlation between protein and fat contents is 0.83.

- Calculate the slope of the linear model for predicting protein content from fat content.

$$\text{slope} = r\frac{s_y}{s_x}$$

Answer:

A. 0.97

B. 0.71

C. 0.83

Regression Line: Order Matters

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. The correlation between protein and fat contents is 0.83.

- Calculate the slope of the linear model for predicting protein content from fat content.

$$\text{slope} = \text{r}\frac{s_y}{s_x} = 0.83 \times \frac{14}{16.4} = 0.71$$

Regression Line: Order Matters

Clicker!

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. The correlation between protein and fat contents is 0.83.

- Calculate the intercept of the linear model for predicting protein content from fat content.

$$\text{intercept} = \bar{y} - b\bar{x}$$

Answer:

A. 0.515

B. 11.288

Regression Line: Order Matters

The mean fat content of 30 Burger King menu items is 23.5g with a standard deviation of 16.4g, and the mean protein content of these items is 17.2g with a standard deviation of 14g. The correlation between protein and fat contents is 0.83.

- Calculate the intercept of the linear model for predicting protein content from fat content.

$$\text{intercept} = \bar{y} - b\bar{x} = 17.2 - 0.71 \times 23.5 = 0.515$$

Regression Line: Order Matters

Clicker!

- What is the equation of the regression line for predicting protein content from fat content?

Answer:

A. $\text{Predicted Fat} = 6.8 + 0.97 \times \text{Protein}$

B. $\text{Predicted Protein} = 0.515 + 0.71 \times \text{Fat}$

C. $\text{Predicted Protein} = 0.71 + 0.515 \text{ x Fat}$

D. $\text{Predicted Fat} = 0.515 + 0.71 \text{ x Protein}$

Regression Line: Order Matters

- What is the equation of the regression line for predicting protein content from fat content?

Answer:

A. $\text{Predicted Fat} = 6.8 + 0.97 \times \text{Protein}$

B. $\text{Predicted Protein} = 0.515 + 0.71 \times \text{Fat}$

C. $\text{Predicted Protein} = 0.71 + 0.515 \text{ x Fat}$

D. $\text{Predicted Fat} = 0.515 + 0.71 \text{ x Protein}$

Regression Line: Order Matters

- What is the equation of the regression line for predicting fat content from protein content?

$$\text{Predicted Fat} = 6.8 + 0.97 \times \text{Protein}$$

- What is the equation of the regression line for predicting protein content from fat content?

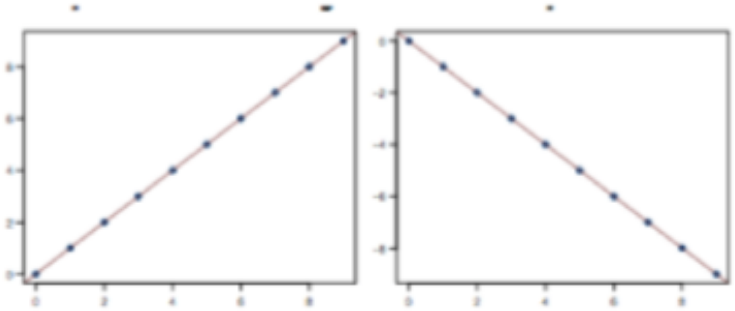
$$\text{Predicted Protein} = 0.515 + 0.71 \times \text{Fat}$$

Assessing Model Fit

- In order to assess if the linear model is a good fit, we look to see how much of the variation in the response variable is accounted for by the model, i.e. explained by the explanatory variable.
- Example: Predicting how well students score on an exam from the number of hours studied for the exam.
    - The response variable is the exam score. There will be variability in exam scores since each student scores differently.
    - One of the reasons for the variability is probably the variability in the number of hours studied for the exam (those who study more will tend to get higher grades).
    - Other reasons could be, previous exposure to material, time spent on homework, etc.
    - Number of hours studied will only explain “some” of the variability in exam scores, but probably not all.

A Measure for Assessing Model Fit

- If the correlation between x and y is r=1 or r=-1, then the model would predict y perfectly and the relationship between the variables would be perfectly linear.



- In the Burger King menu model r=0.83. This is not a perfect fit. We would be interested in knowing how much of the variation in the fat content is explained by the protein content.

Coefficient of Determination

- The coefficient of determination or the correlation coefficient squared, r^2 (or r-squared), gives the percentage of the variance of y is explained by x.
- For the Burger King model: $r^2 = 0.83^2 = 0.69$
- When interpreting a regression model we need to explain in context of what r^2 means:
    - Burger King model: 69% of the variation in the fat content (response or y-variable) is explained by the protein content (explanatory or x-variable) of the burger.

Coefficient of Determination

- $r^2 = 0$ means that none of the variability in y is explained by x.
- $r^2 = 1$ means that all of the variability in y is explained by x.
- While the correlation coefficient is between -1 and 1, r^2 is between 0 and 1.

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

- We would like r^2 to be as close to 100% as possible.

Example

Clicker!

- The correlation between weight and gas mileage of cars is -0.96. Which of the following is the correct value and interpretation of r-squared of the linear model for predicting gas mileage from weight?

(A) 8% of the variation in gas mileage is explained by the weight of the cars.

(B) 92% of the variation in the weight is explained by the gas mileage of the cars.

(C) 92% of the variation in the gas mileage is explained by the weight of the cars.

(D) 96% of the variation in the gas milage is explained by the weighted of the cars.

Example

- The correlation between weight and gas mileage of cars is -0.96. Which of the following is the correct value and interpretation of r-squared of the linear model for predicting gas mileage from weight?

(A) 8% of the variation in gas mileage is explained by the weight of the cars.

(B) 92% of the variation in the weight is explained by the gas mileage of the cars.

(C) 92% of the variation in the gas mileage is explained by the weight of the cars.

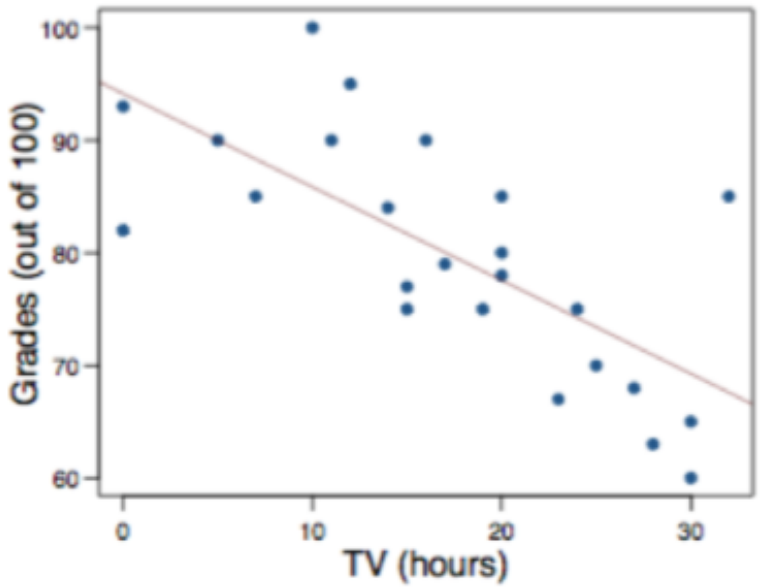
(D) 96% of the variation in the gas milage is explained by the weighted of the cars.

From r-squared to r

Clicker!

- The following scatterplot shows the relationship between number of hours students watch TV and their score on an exam. If r-squared is 50%, what is the correlation coefficient?

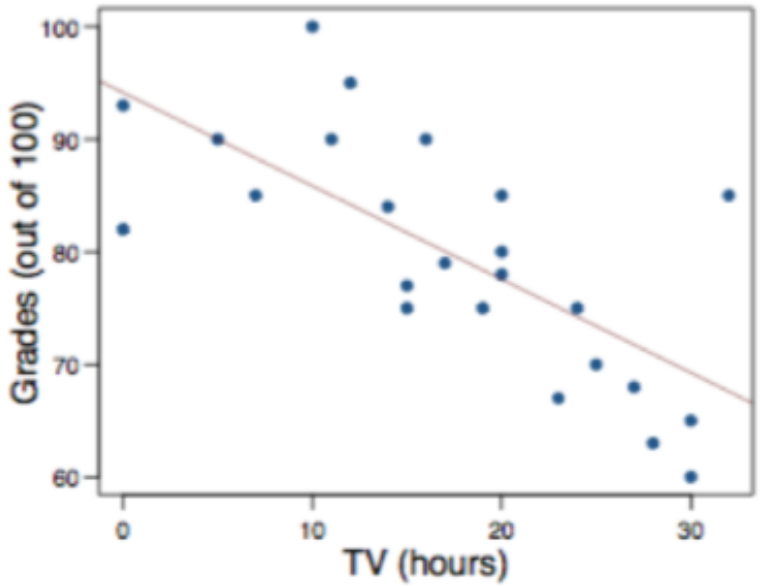
a) -0.71
b) 0.71
c) 0.25
d) 7.07



From r-squared to r

- The following scatterplot shows the relationship between number of hours students watch TV and their score on an exam. If r-squared is 50%, what is the correlation coefficient?

a) -0.71
b) 0.71
c) 0.25
d) 7.07



Prediction

- When asked to make a prediction using the linear model, simply plug in the value for x in the equation for the regression line and calculate predicted y.
- If the Burger King chicken sandwich has a protein content of 30 grams, what is the predicted fat content?
- Using the following regression line, find the predicted fat content for a Burger King chicken sandwich:

$$\text{Predicted Fat} = 6.8 + 0.97 \times \text{Protein}$$

Prediction

- When asked to make a prediction using the linear model, simply plug in the value for x in the equation for the regression line and calculate predicted y.
- If the Burger King chicken sandwich has a protein content of 30 grams, what is the predicted fat content?
- Using the following regression line, find the predicted fat content for a Burger King chicken sandwich:

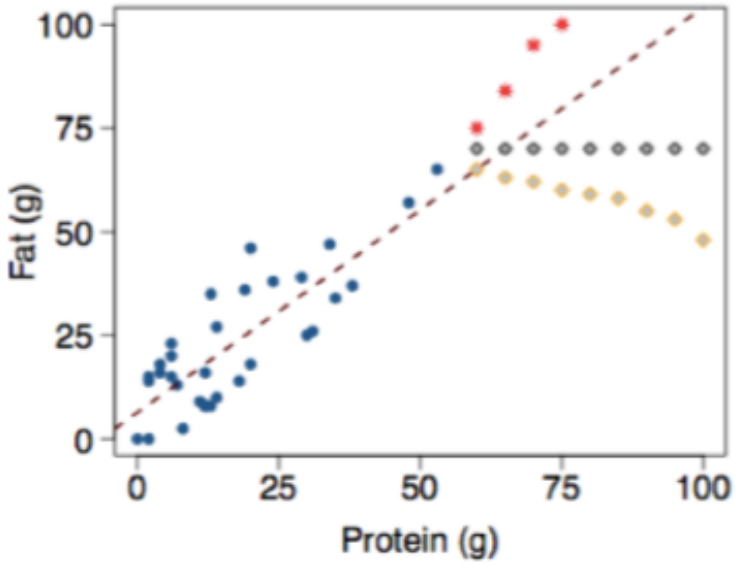
$$\text{Predicted Fat} = 6.8 + 0.97 \times \text{Protein}$$

$$\text{Predicted Fat} = 6.8 + 0.97 \times 30 = 35.9g$$

$$\hat{y} = 35.9g$$

Extrapolation

- Do not extrapolate beyond the data, i.e., do not make a prediction for a x value outside the range of the data - the linear model may no longer hold outside that range.
- If the Burger King chicken sandwich has a protein content of 75 grams, what is the predicted fat content?
- We should not use the linear model to predict the amount of fat in a burger with 75 g of protein since the data we used to create the model is for burgers with approximately 0 to 50 g of protein.



Pitfalls to Avoid

- Don’t fit linear models to nonlinear associations.
- Correlation is not causation.
- Beware of outliers! Try the regression and correlation with and without influential points to see the differences
- Be careful of regressions of aggregate data (data of means rather than individuals).
- Don’t extrapolate.