Two-Variable Data Analysis

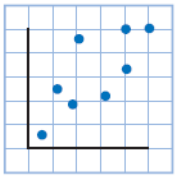
Extra Practice

MHR Page 436 #s 1 - 7

Solutions

1. Classify the nature of each linear correlation.

a)

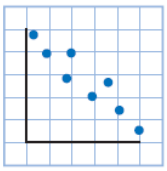


A

A moderate positive correlation
B strong positive correlation
C moderate negative correlation
D strong negative correlation

As x increases, y increases so there is a positive correlation. The data is not that close to forming a straight line, so the correlation is moderate.

b)

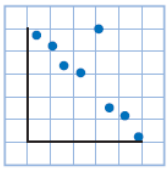


D

A moderate positive correlation
B strong positive correlation
C moderate negative correlation
D strong negative correlation

As x increases, y decreases so there is a negative correlation, The data is close to forming a straight line, so the correlation is strong.

2. Consider the scatter plot relating two variables.



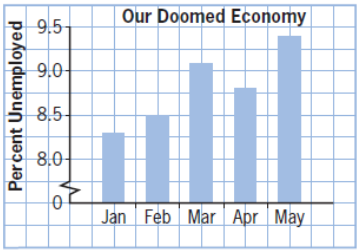
Which of the following best characterizes this correlation?

A strong positive correlation
B strong negative correlation
C strong negative correlation with a hidden variable
D strong negative correlation with an outlier

D

As x increases, y decreases so the correlation is negative. The data is close to forming a straight line, so the correlation is strong. However, there is one datum that does not fit, which can be regarded as an outlier.

3. The graph shows the unemployment rate over a five-month period.



Which of the following contribute to bias in this graph?

A the title
B the sample size
C the choice of vertical scale
D all of the above

D

The title "Our Doomed Economy" is biased, the sample size is very small (only five months) and the vertical scale has been altered (doesn't start at zero) so as to exaggerate the increase in the unemployment rate.

4. Classify each of the relationships. The independent variable is listed first.

a) Snow tire sales are positively correlated with hot chocolate sales.

b) Movie box office sales are negatively correlated with ticket price.

c) Driving test scores are negatively correlated with driver height.

d) Cheeseburger sales are negatively correlated with pita sales.

a) Both of these sales are associated with winter, so we can characterize this as a **common cause relationship**.

b) It is likely that box office sales will decrease as tickets become more expensive, so we can characterize this as a **reverse cause and effect relationship**.

c) We can say that height can play a role in driving safety but it is difficult to prove, so we can characterize this as a **presumed relationship**.

d) There would appear to be no obvious connection between the sales of cheeseburgers and pitas, so we can characterize this as an **accidental relationship**.

5. Attendance at football games is positively correlated with a team's position in the league standings.

a) Suggest a cause and effect relationship that could account for the results.

b) Pose and defend an argument for a reverse cause and effect relationship.

a) As a team is more successful, attendance at the games is likely to increase. People tend to attend more games when their team is being more successful. The better they do, the more fans will attend.

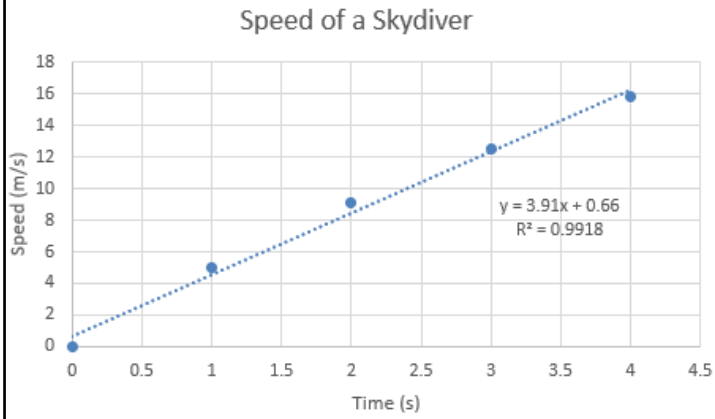
b) As a team gets more support from its fans, their results tend to improve. So as they get more support (an increase in attendance), they will be more successful, and therefore move up the league standings.

6. The table shows the speed of a skydiver as she falls from the instant she leaves a plane.

a) Construct a scatter plot of speed versus time. Describe the correlation.

b) Perform a linear regression. Interpret the equation of the line of best fit.

Time (s)	Speed (m/s)
0	0.0
1	5.0
2	9.1
3	12.5
4	15.8



a) There appears to be a strong, positive linear correlation.

b) The equation of the line of best fit is y = 3.91x + 0.66, where y is the speed of the skydiver in m/s and x is the time since jumping in seconds.

The r-value is 0.996 which means that the equation is a very good fit. The equation shows that the skydiver's speed started at 0.66 m/s and will increase a rate at 3.91 m/s.

7. The number of users of a social networking website is shown as a time series.

a) Construct a scatter plot of this time series. Describe the trend.

b) Perform a linear regression. Interpret the equation of the line of best fit.

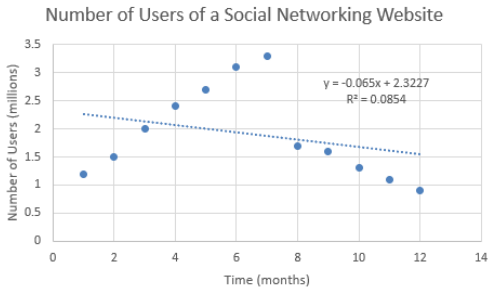
c) Construct a residual plot. Do you think this is a good model for this correlation? Explain why or why not.

d) Is there evidence of a hidden variable? Explain.

e) Initially this website was free. When did the website start charging a fee? How do you know? Explain the effect this had on the linear trend.

g) Construct a new linear model that will do a better job of predicting the future popularity of this website. Discuss your reasoning, including any assumptions you make.

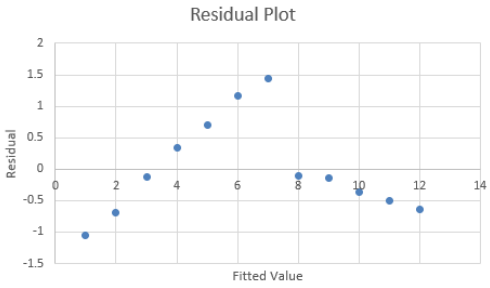
Time (months)	Users (millions)	Time (months)	Users (millions)
1	1.2	7	3.3
2	1.5	8	1.7
3	2.0	9	1.6
4	2.4	10	1.3
5	2.7	11	1.1
6	3.1	12	0.9



a) There seems to be two trends. At first there is an increase in user numbers and then there is a decrease in user numbers.

b) The r-value is 0.292 which suggests a weak, negative linear correlation.

The equation of best fit is y = -0.065x + 2.3227 where y is the number of users in millions, and x is the time in months. The equation shows that the numbers of users started at around 2,320,000 and is decreasing at a rate 65,000 users per month.



c) The residual plot suggests that the model is not a good fit. Whenever the residuals look a bit "parabolic" (u-shaped or n-shaped) this suggests that a better model is possible. We actually have two linear patterns here.

7. The number of users of a social networking website is shown as a time series.

a) Construct a scatter plot of this time series. Describe the trend.

b) Perform a linear regression. Interpret the equation of the line of best fit.

c) Construct a residual plot. Do you think this is a good model for this correlation? Explain why or why not.

d) Is there evidence of a hidden variable? Explain.

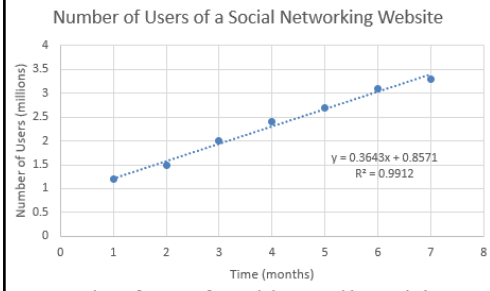
e) Initially this website was free. When did the website start charging a fee? How do you know? Explain the effect this had on the linear trend.

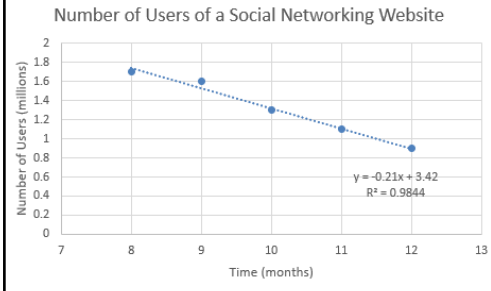
g) Construct a new linear model that will do a better job of predicting the future popularity of this website. Discuss your reasoning, including any assumptions you make.

Time (months)	Users (millions)	Time (months)	Users (millions)
1	1.2	7	3.3
2	1.5	8	1.7
3	2.0	9	1.6
4	2.4	10	1.3
5	2.7	11	1.1
6	3.1	12	0.9

d) There could well be a hidden variable as in month 8 there is a sharp drop in user numbers.

e) The website started charging a fee in month 8. This is what caused the drop in user numbers and turned the trend from positive to negative.





f) Looking at the two different trends separately, we can see that they both have strong models (r-values of 0.996 and 0.992 respectively). The first model is for when there was no fee and it showed an increase of around 364,000 users per month. When they changed to charging a fee there was a sharp drop in numbers, so that they started with around 1.7 million users and were then losing around 210,000 users per month. Using the model of y = -0.21x + 3.42 the website will have no users by month 17.