Review

Two-Variable Data Analysis

Page 439 #s 6 - 10

- When a change in one variable is accompanied by a proportional change in another variable, the variables share a linear correlation.
- The correlation coefficient, r, is a measure of the strength of linear correlation between two variables. The value of r, which can be between –1 and 1, gives an indication of how closely the data points relate to the line of best fit.
- The line of best fit can be used to model a linear correlation.
- Linear regression is the mathematical process that determines the line of best fit.

- A cause and effect relationship exists when one variable is directly responsible for a change in another variable.
- If two variables share a strong correlation, it does not imply that a cause and effect relationship exists.
- A common cause relationship exists when a common third variable is responsible for the correlation between two other variables.
- Several types of relationships can exist between two variables, including cause and effect, common cause, presumed, reverse cause and effect, and accidental.

- A residual is the difference between the actual dependent value of a datum and the value predicted by a line of best fit.
- A residual plot shows how close each data point is to a line of best fit.
- An outlier is a data point that does not fit well in an otherwise linear trend.
- An outlier can have a strong impact on a linear regression model if the number of data points is relatively small.
- A hidden variable can distort or obscure a linear correlation between two other variables.
- It is important to consider the impact of outliers and hidden variables when conducting a correlational study. It may help to remove or account for them when analysing the data.

- You can display data in multiple ways. Sometimes data are deliberately distorted to make an argument more convincing.
- The media often sensationalize data to generate public interest.
- Raw data can be awkward to work with. Software progams have tools to filter, organize, present, and analyse data.
- Contingency tables, side-by-side box plots, pivot tables, and pivot charts are methods for comparing quantitative data across different categories.
- Bubble plots and legend attributes are tools that can be used to recognize the possible impact of a hidden variable in a correlational study.

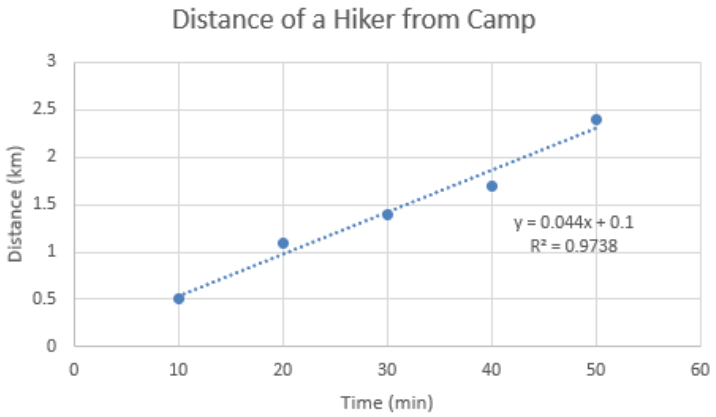
Solutions

6. The table gives the distance of a hiker from camp over time.

Time (min)	Distance (km)
10	0.5
20	1.1
30	1.4
40	1.7
50	2.4

a) Create a scatter plot of distance versus time. Describe the correlation.

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



a) There is a strong, positive linear correlation.

b) The line of best fit is telling us that the hiker started at 0.1 km from the camp and is walking at a rate of 0.044 km/min away from it. The correlation coefficient is $\sqrt{0.9738} \approx 0.99$ suggesting a strong correlation.

7. A study found that worker absenteeism is negatively correlated with income level.

a) Do you think this is a cause and effect relationship? Explain.

b) Suggest a possible common cause factor that could account for this relationship.

a) No. It is unlikely that a higher income level would cause less absenteeism. Even if you earn more, it doesn't mean that you work more. Both of these variables are likely to share a common trend with another (external) variable.

b) Possible common causes that could account for this relationship include level of job satisfaction, financial wellbeing, and other family issues.

8. Characterize each type of relationship.

a) Automotive sales is positively correlated with amount of rainfall.

b) Number of hours practised is negatively correlated with number of musical errors.

a) As there is no obvious reason to say that the more it rains, the more vehicles are sold, we can state the relationship is accidental.

b) Studies have shown that the more we practice the better we get, so it is reasonable to conclude that this is a cause and effect relationship.

9. The table shows student achievement for the grade 9 applied EQAO assessment over time.

Year	Percent Achieving at or Above Provincial Standard
2001–2002	21
2002–2003	21
2003–2004	25
2004–2005	27
2005–2006	35
2006–2007	35
2007–2008	34
2008–2009	38
2009–2010	40
2010–2011	42
2011–2012	44
2012–2013	44

a) Create a scatter plot for this time series. Call 2001–2002 year 0. Describe the correlation.

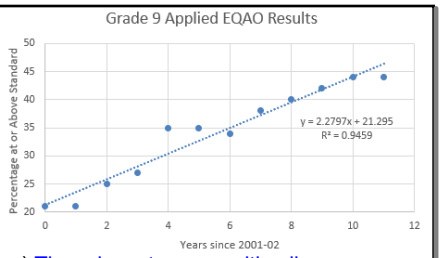
b) Perform a linear regression. Interpret the equation of the line of best fit.

c) Construct a residual plot. Does there appear to be any evidence of a hidden variable in the data? Explain.

d) In 2005, the math curriculum was revised. Could this fact be considered a hidden variable? Why or why not?

e) Repeat the analysis performed in parts a) to c) for 2005–06 to 2012–2013.

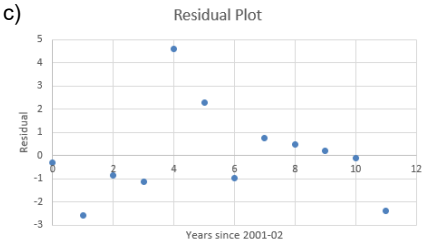
f) Is there evidence that this linear model is better than the original one? Explain.



a) There is a strong, positive linear correlation. However, there does appear to be two separate trends.

b) The line of best fit is implying that for every year that has passed since 2001-02, the number of students who are at or above standard is increasing by 2.28%. The percentage at or above standard in 2001-02 was 21.3%. The correlation coefficient is $\sqrt{0.9459} \approx 0.97$ suggesting a strong correlation.

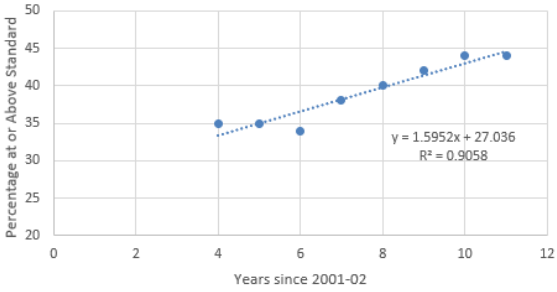
c)

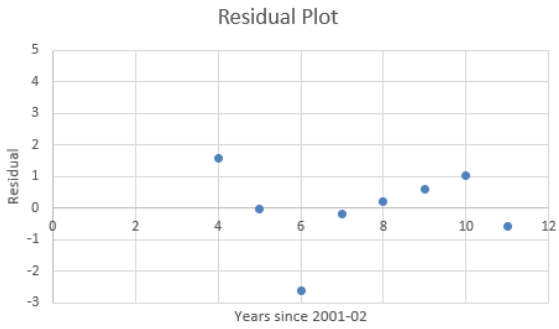


There does appear to be evidence of a hidden variable as the residuals form a "mountain" shape, with the peak in 2004-05 (x = 4).

d) Yes, the change in curriculum could be considered a hidden variable. Perhaps in became more in line with the EQAO exam than it had previously, resulting in higher marks.

e)





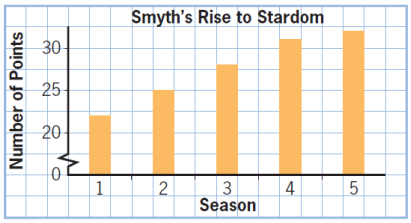
There is a still a strong, positive linear correlation.

The line of best fit is implying that for every year that has passed since 2004-05, the number of students who are at or above standard is increasing by 1.60%. The percentage at or above standard in 2004-05 was 27.0%. The correlation coefficient is $\sqrt{0.9059} \approx 0.95$ which still suggests a strong correlation.

There does not appear to be any evidence of a hidden variable as the residuals are above and below the residual line and don't form a linear or parabolic shape.

f) This model appears to be a better fit, with one potential outlier (x = 6). We need to exercise caution though as we are now using a smaller sample (less data) to draw these conclusions.

10. The graph shows the point totals for a hockey player's first five seasons.



a) Identify any sources of bias in the graph.

b) Smyth's contract is up for renewal. Do you think the graph was made by the team manager or by Smyth's agent? Explain your thinking.

c) How could you remove all bias in the graph?

a) The there is bias in the title of the graph, the small sample size (five seasons), and also with the vertical axis (doesn't start at zero).

b) It was likely made by Smyth's agent. The sensationalist title as well as the manipulation of the vertical axis to accentuate rapid progress are all designed to make Smyth seem better than he/she probably is.

c) Bias can be removed by using a neutral title (Smyth's First Five Seasons) and maing the vertical axis start at zero.