Dynamic Analysis of Two-Variable Data

Lesson objectives

- I can identify outliers and account for their impact on a data trend
- I can recognize the presence of extraneous variables
- I can identify a hidden variable and account for its impact on a correlation

1.1

Lesson objectives | Teachers' notes | Lesson notes

MHR Page 413 #s 1 - 14

Warm up

Suppose two athletes are competing for a position on the track and field team in the long jump event. The chart shows their tryout distances, in metres.

	Jump 1	Jump 2	Jump 3	Jump 4	Jump 5	Mean
Hank	4.5	4.0	4.3	4.2	4.0	4.2
Vito	4.9	5.1	4.8	1.2	4.7	4.1

- Which athlete do you think should make the team?
- Could you make an argument for either athlete?
- Is there anything that seems unusual in the data?
- Would more information be helpful?

Vito should make the team

Yes, Hank was more consistent with his five jumps. Vito had three jumps better Hank's best jump.

Vito had one very short jump (1.2 m) which would affect his mean jump.

If there was more data we could be more confident with our selection. Five jumps is a very small sample size.

Definitions

Residual Plot

- Shows the value of each **residual** graphically as the vertical distance from a horizontal axis

Residual

- The difference between a data point's actual dependent value and the dependent value predicted by the line of best fit

Outlier

- A data point that does not fit an otherwise clear trend
- In a scatter plot, the outlier is either relatively far above or below the horizontal line

Hidden Variable

- A variable that affects or obscures the relationship between two other variables
- Can sometimes result in a false correlation or a fragmented trend

Example 1

Construct and Analyse a Residual Plot

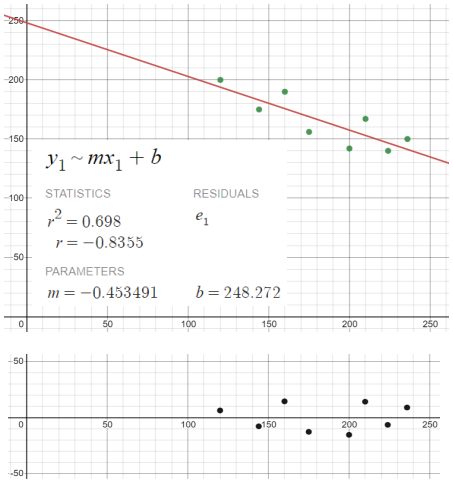
Will an increase in recycling result in a reduction of landfill? The table compares the mass of garbage and recycling for a town during a recycling campaign.

a) Create a scatter plot and perform a linear regression. Describe the trend. Is a linear model reasonable in this case?

b) Interpret the residuals for (120, 200) and (200, 142).

c) Construct a residual plot. Describe the pattern.

Amount Recycled (kg)	Amount of Garbage (kg)
120	200
144	175
160	190
175	156
200	142
210	167
224	140
236	150



a) The correlation coefficient, r, is about -0.84 which implies a strong, negative linear correlation.

b) The residuals tell the difference between the actual values and what the model predicts. Using the equation y = -0.453x + 248 a mass of 120kg of recycled material should give a mass of about 194kg of garbage. The residual for 120kg of recycled material is 200 - 194 = 6. This means that the actual mass is 6kg more than that predicted by the model. Similarly for a recycled mass of 200kg, the model gives a mass of about 158kg of garbage. The residual for this is 142 - 158 = -16. This time the actual mass is 16kg below the predicted value.

c) There seems to be no discernible pattern to the residual data which is normal for a good linear model. The points are above and below the horizontal axis.

Your Turn

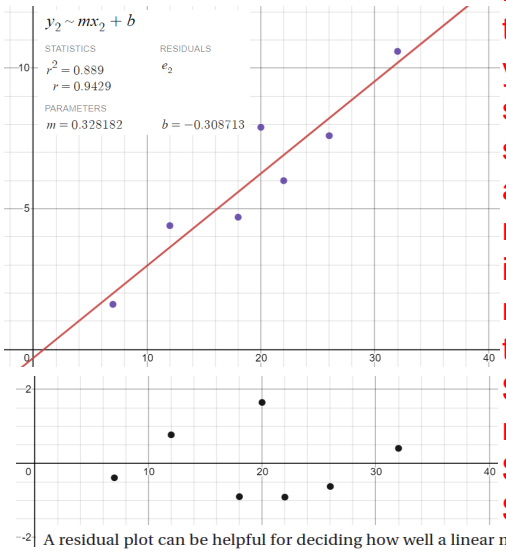
The table compares a jewellery store's weekly diamond ring sales to the minutes of radio advertising purchased for the same week.

a) Create a scatter plot and perform a linear regression. Describe the trend. Is a linear model reasonable in this case?

b) Interpret the residuals for (20, 7.9) and (18, 4.7).

c) Construct a residual plot. Describe the pattern in it.

Advertising (min)	Diamond Ring Sales ($1000s)
7	1.6
12	4.4
20	7.9
32	10.6
22	6.0
18	4.7
26	7.6



A residual plot can be helpful for deciding how well a linear model fits a set of data. If there is a pattern or irregularity in a residual plot, the linear model may have to be re-evaluated.

a) The correlation coefficient, r, is about 0.94 which implies a strong, positive linear correlation.

b) The residuals tell the difference between the actual values and what the model predicts. Using the equation y = 0.328x - 0.31 a 20 minute ad should bring in $6250 of diamond ring sales. The residual for 20 minutes of ad time is 7900 - 6250 = 1650. This means that the actual amount of sales is $1650 more than predicted by the model. Similarly for an 18 minute ad, the model pedicts sales of about $5600 of diamond ring sales. The residual for this is 4700 - 5600 = -$900. This time the actual sales are $900 less than the predicted value.

c) Again, there seems to be no discernible pattern to the residual data which is normal for a good linear model. There should be roughly the same number of points above and below the horizontal axis.

Example 2

Account for Outliers

The table shows repair costs over the course of a year for several cars of the same model.

a) Create a scatter plot of repair costs versus age.

b) Perform a linear regression and discuss the goodness of fit.

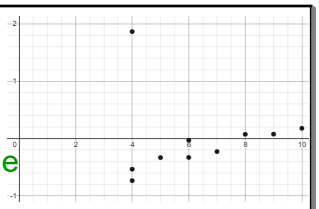
c) Construct a residual plot and identify any outliers.

d) Repeat the regression with the outlier removed.

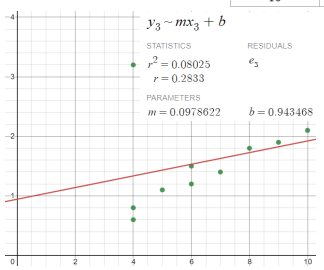
e) Compare the two linear models.

Age of Car (years)	Repair Costs ($1000s)
4	0.6
4	0.8
4	3.2
5	1.1
6	1.2
6	1.5
7	1.4
8	1.8
9	1.9
10	2.1

c) The point (4,3.2) appears to be an outlier. It is much further away from the horizontal axis than the other points.

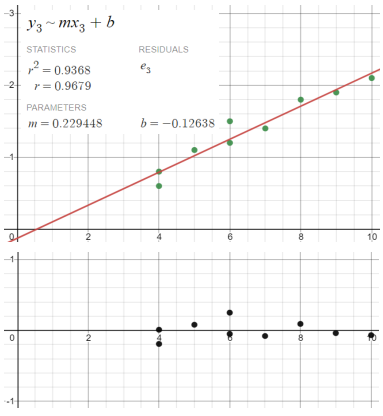


a)



b) The correlation coefficient, r, is about 0.28 which implies a weak, positive linear correlation. Also the y-intercept would imply an almost $1000 repair bill for a brand new car. The point (4,3.2) appears to be an outlier.

d) After removing the outlier, the correlation coefficient, r, is about 0.97 which implies a very strong, positive linear correlation. Also the y-intercept would imply a slightly negative repair bill for a brand new car which is more sensible. The residuals are all close to the horizontal axis, so we can say that the point (4, 3.2) is an outlier.



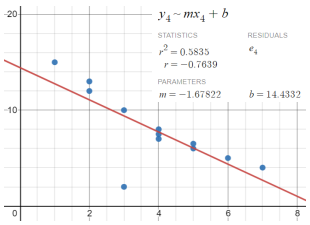
Your Turn

The table shows the sale price for several used motorcycles of the same model and their age.

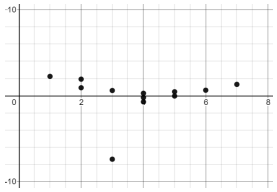
Age (years)	Price ($1000s)
1	15
2	13
2	12
3	2
3	10
4	8
4	7.5
4	7
5	6.5
5	6
6	5
7	4

a) Create a scatter plot of sale price versus age.
b) Perform a linear regression and discuss the goodness of fit.
c) Construct a residual plot and identify any outliers. Suggest reasons why any outliers may exist.
d) Repeat the regression with the outlier removed.
e) Compare the two linear models.

a)

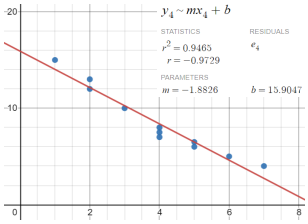


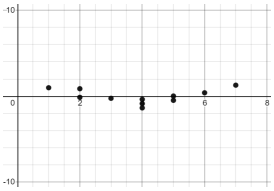
b) The correlation coefficient, r, is about -0.76 which implies a moderate, negative linear correlation. The point (3,2) appears to be an outlier.



c) The point (3,2) appears to be an outlier. It is much further away from the horizontal axis than the other points.

d)





e) After removing the outlier, the correlation coefficient, r, is about -0.97 which implies a very strong, negative linear correlation. The residuals are all close to the horizontal axis, so we can say that the point (3,2) is an outlier. The point (3,2) may indicate that the motorcycle was damaged, or had been neglected, or maybe it was just a lemon.

Example 3

Account for a Hidden Variable

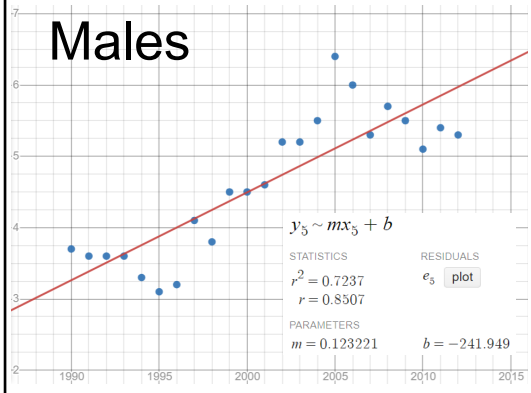
Is absenteeism among professional workers on the rise in Canada? Are there similar trends for males and females? The contingency table shows the average number of absences for all Canadian males and females who have at least one university degree. Perform dynamic statistical analysis to determine any absenteeism trends.

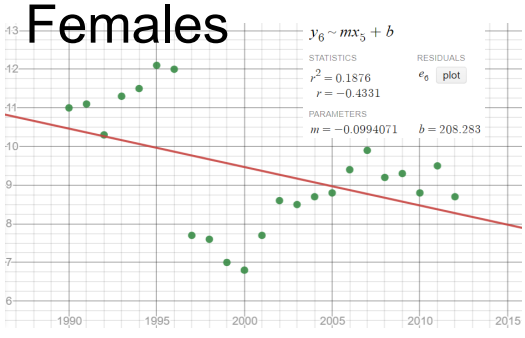
Absences					
Year	Males	Females	Year	Males	Females
1990	3.7	11	2002	5.2	8.6
1991	3.6	11.1	2003	5.2	8.5
1992	3.6	10.3	2004	5.5	8.7
1993	3.6	11.3	2005	6.4	8.8
1994	3.3	11.5	2006	6.0	9.4
1995	3.1	12.1	2007	5.3	9.9
1996	3.2	12.0	2008	5.7	9.2
1997	4.1	7.7	2009	5.5	9.3
1998	3.8	7.6	2010	5.1	8.8
1999	4.5	7.0	2011	5.4	9.5
2000	4.5	6.8	2012	5.3	8.7
2001	4.6	7.7			

Source: CANSIM Table 279-0036, Absence rates of full-time employees, by sex and education, Canada, Statistics Canada, September 18, 2013

There appears to be an upward trend for male absenteeism. The correlation coefficient is about 0.85. Using the model, it implies that male absenteeism is increasing at a rate of about 0.12 days per year.

There appears to be a slight downward trend for female absenteeism. The correlation coefficient is about -0.43. Using the model, it implies that female absenteeism is decreasing at a rate of about 0.1 days per year. However, caution is needed as the correlation is barely moderate at best.

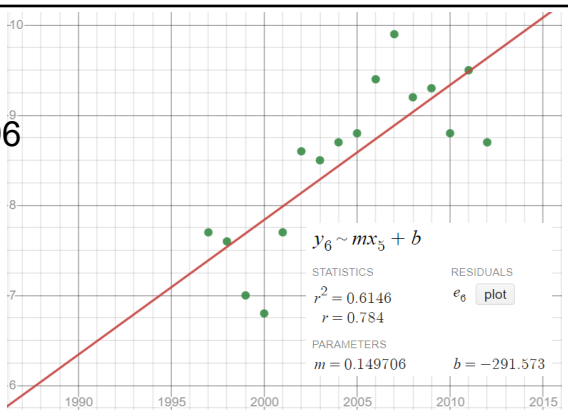




Footnotes:

1. Data from 1987 to 1996 include maternity leave. Also, men using paid paternity (in Quebec only) and parental leave are included in the calculation till 2006.

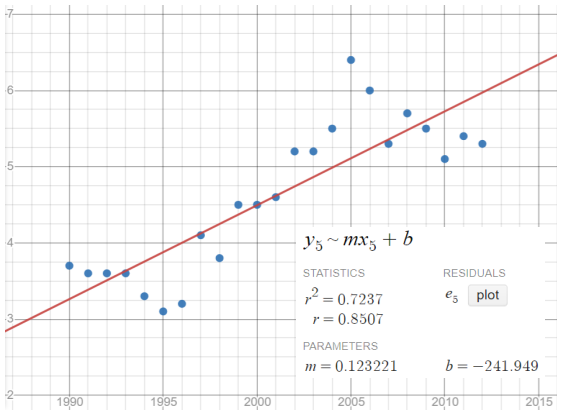
When we remove the data from 1990 to 1996 we get the following plot. The r value changes to 0.78 which confirms a strong positive correlation. The model is now showing an increase in absenteeism for females of 0.15 days per year.



Overall, absenteeism seems to be on the rise in Canada!

Your Turn

Do you think men using paid paternity leave in Quebec until 2006 is a significant hidden variable in this analysis? Why or why not?



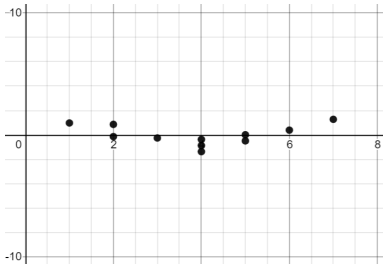
Looking at the data from 1996 to 2006 there only appears to be one outlier (2005,6.4). If paternity leave was having a big impact on absenteeism, then the r value would be much lower than 0.8507.

Key Concepts

- 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.

R1. Explain what a residual plot shows for a set of two-variable data. Draw a sketch to support your answer.

A residual plot shows the value of each residual graphically as the vertical distance from a horizontal axis. A random pattern indicates a good fit for a linear model.



R2. **a)** What is an outlier?

b) Describe two ways of identifying an outlier in a set of two-variable data.

a) An outlier is a data point that does not fit well in an otherwise linear trend.

b) In a scatter plot, the outlier is relatively far from the line of best fit. In a residual plot, the outlier is either relatively far above or below the horizontal line.