THE ERINDALE ACADEMY
1576 Dundas Street West, Mississauga ON L5C 1E5
www.teacademy.ca Tel: 905-232-1576
Email: info@teacademy.ca

Teacher: Ella

Unit 6: One and Two Variable data analysis

Lesson 6.1: Line of Best fit

Learning Goal: Determine a correlation coefficient to classify a linear relation

Two-Variable analysis

- Two-variable statistics provides methods for finding relationships _____ variables.
- Shows whether one variable (dependent variable) is affected by another variable (independent variable)

Correlations

- A linear correlation exists if changes in one variable tend to be _____ to changes in the other
- They have a _____ linear correlation if Y increases at a constant rate as X increases.
- They have a _____ linear correlation if Y decreases at a constant rate as X increases.
- A _____ is the straight line that passes as close as possible to all of the points on a scatter plot.
- The stronger the correlation, the more closely the data points cluster around the LOBF.

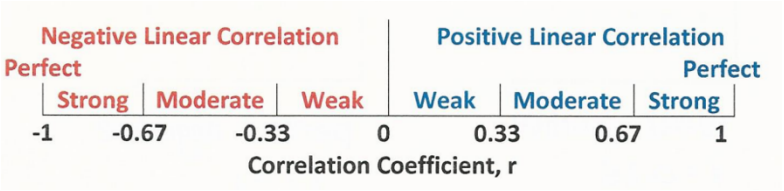
The correlation coefficient, r

- A scatter plot can give only a rough indication of the correlation between two variables
- The covariance depends on how the deviation of the two variables are related

Covariance

- Covariance will have a large positive value if both $(x - \overline{x})$ and $(y - \overline{y})$ tend to be large at the same time
- Covariance will have a negative value if one tends to be positive when the other is negative
  Correlation coefficient, r, will be obtained when the covariance divided by the product of the standard deviations for X and Y.

Correlation coefficient



- Correlation coefficient is a measure of how well a linear model fits a two-variable set of data
- Always has values in the range from – 1 to 1

THE ERINDALE ACADEMY
1576 Dundas Street West, Mississauga ON L5C 1E5
www.teacademy.ca Tel: 905-232-1576
Email: info@teacademy.ca

Teacher: Ella

Classifying linear correlations and Direction of relationship

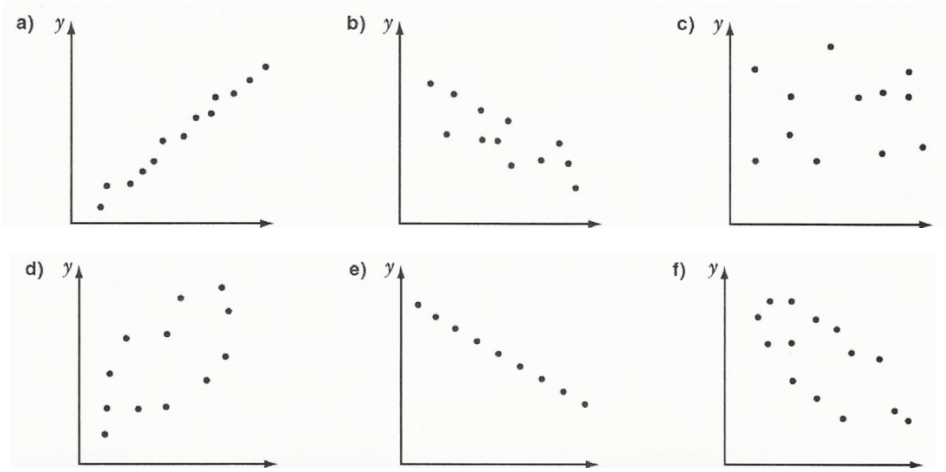
- linear correlations can have positive or negative relationships, or no relationship at all.
- The strength of the relationships are classified as weak, moderate, strong, or perfect.

For instance:

- – 1 indicates perfectly negative correlated
- – 0.8 indicates strongly negative correlated
- 0 indicates no correlation between dependent and independent variables
- 1 indicates perfectly positively correlated
- 0.8 indicates strongly positively correlated
- 0.4 or – 0.4 is moderately positive/negative correlated
- 0.1 or – 0.1 is weakly positive/negative correlated

Activity:

- classify the relationship between the variables X and Y for the data shown in the following diagrams using a combination of relationship direction and strength (i.e., strong negative or weak positive)
- identify an approximate correlation coefficient value for each graph.



Calculating the correlation coefficient

A survey of a group of randomly selected students compared the number of hours of television they watched per week with their grade average.

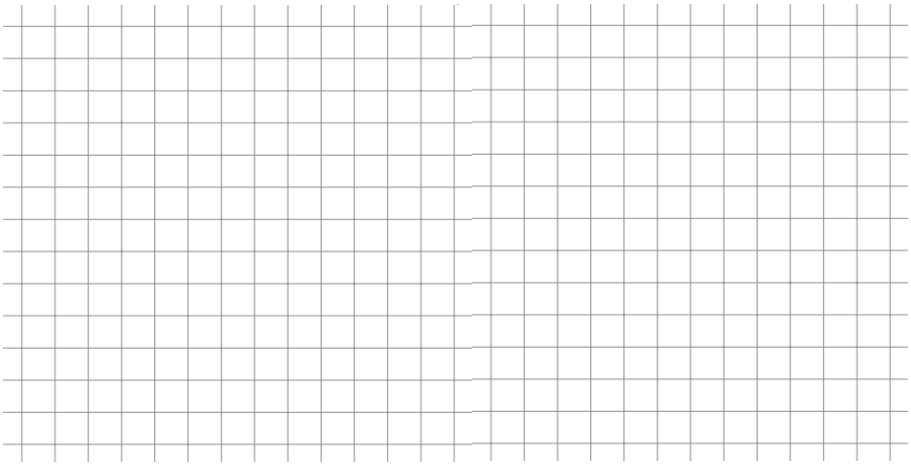
1. Create a scatter plot and classify the linear correlation between the two variables (positive or negative and strong, moderate, or weak)
2. Determine the correlation coefficient

Continues on next page…

THE ERINDALE ACADEMY
1576 Dundas Street West, Mississauga ON L5C 1E5
www.teacademy.ca Tel: 905-232-1576
Email: info@teacademy.ca

Teacher: Ella

Hours per Week	Grade Avg (%)
12	70
10	85
5	82
3	88
15	65
16	75
8	68



Hours/week, X	Grade %, Y	X^2	Y^2	XY

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{\left[n\sum x^2 - (\sum x)^2\right]\left[n\sum y^2 - (\sum y)^2\right]}}$$

Can you make any conclusions about the effect that watching television has on academic achievement?

Teacher: Ella

Activity:

Using Technology to Find the LOBF and Correlation Coefficient

The table shows distance from home for a cyclist over time.

Time (min)	Distance (km)
10	9.8
20	8.1
30	5.8
40	4.2
50	2.3

a) Create a scatter plot relating distance, *d*, and time, *t*

- ☐ Enter the data into Excel as shown
- ☐ Highlight the data in both columns (including the column labels)
- ☐ Insert a scatter plot
- ☐ Add a chart title and axes labels

b) Determine the strength of linear correlation with technology.

- ☐ In an empty cell type **=CORREL(**
- ☐ Highlight data cells for the independent variable (not including the column label)
- ☐ Type **,** (comma)
- ☐ Highlight data cells for dependent variable (not including the column label)
- ☐ Hit Enter
- ☐ Classify the strength and direction of the correlation

c) Determine the equation of the line of best fit and explain what it means

- ☐ Right-click on any data point on the graph and select Add Trendline
- ☐ Choose a trend line to insert (the default is linear)
- ☐ Select the bottom two checkboxes to display the equation and the R-squared
  - ○ R-square is the coefficient of determination which is the square of the correlation coefficient
- ☐ Click Close
- ☐ Describe the relationship

Teacher: Ella

Learning Goal: Analyze & interpret statistical information to assess reliability and validity of conclusions

Linear Regression

- An analytic technique for determine the relationship between a dependent and independent variable
- By finding and using the equation for the LOBF, you can answer questions about trends in the data
    - _____ is when you estimate _____
    - _____is when you predict _____
- Creating a LOBF using a least-squares fit gives more accurate results, especially for weak correlations

Least-squares fit

- Determines the residuals (vertical deviation from the LOBF)
    - Residuals are positive for points above the line and negative for points below the line
    - The sum of the residuals is zero (the positive and negative residuals cancel out)
    - The sum of the squares of the residuals has the least possible value

You can find the equation for any LOBF in a scatter plot using this method:

THE ERINDALE ACADEMY
1576 Dundas Street West, Mississauga ON L5C 1E5
www.teacademy.ca Tel: 905-232-1576
Email: info@teacademy.ca

Teacher: Ella

Example 1: This table shows data for the full-time employees of a small company. Analytically, find the equation of the LOBF and classify the correlation.

Age	Annual Income ($000)
33	33
25	31
19	18
44	52
50	56
54	60
38	44
29	35

Age	Income ($000)	x^2	y^2	xy
33	33			
25	31			
19	18			
$\sum x =$	$\sum y =$	$\sum x^2 =$	$\sum y^2 =$	$\sum xy =$

Linear Regression:

$$a = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{n\sum x^2 - \left(\sum x\right)^2} \quad \textbf{and} \quad b = \bar{y} - a\bar{x} \quad \text{, where } \bar{x} = \frac{\sum x}{n},\ \bar{y} = \frac{\sum y}{n}$$

Correlation coefficient:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

Predict the income for a new employee who is 21 and an employee at age 65. Are these examples of interpolation or extrapolation?

Teacher: Ella

Your turn: Spend 30mins to work on this Task, then hand it in to your teacher to photocopy, so the copy will be kept in your student's folder.

Researchers monitoring the number of wolves and rabbits in a wildlife reserve think that the wolf population depends on the rabbit population since wolves pry on rabbits. Over a period of eight years researchers have collected the following data.

Year	1994	1995	1996	1997	1998	1999	2000	2001
Rabbit Population	61	72	78	76	65	54	39	43
Wolf Population	26	33	42	49	37	30	24	19

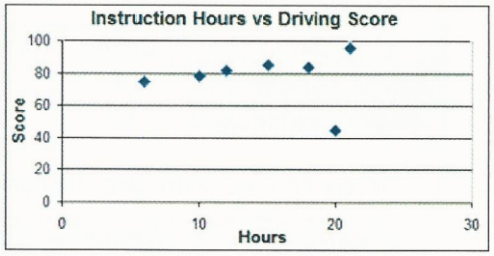
a) Determine the LOBF and the correlation coefficient for these data.

b) Use Excel to graph the data and the LOBF to check if your values from part a) are correct. And suggest if these data support the researchers' theory?

Teacher: Ella

Homework: To evaluate the performance of one of its instructors, a driving school tabulates the number of hours of instruction and the driving-test scores for the instructor's students.

Hours	Score
10	78
15	85
21	96
6	75
18	84
20	45
12	82



a) analyze these data to determine whether they suggest that the instructor is an effective teacher. (Hint: focus on correlation coefficient, close to zero, 1, or – 1)

b) Use outlier check (1.5 times of box width) to comment on any data that seem unusual.

c) Determine the effect of any outliers on your analysis. (Hint: double check correlation coefficient if you remove the outlier)

Teacher: Ella