

S3.5

Bi-Variate Data and Correlation

3.5 90645
Select and analyse continuous bi-variate data
www.ppresources.com
PC Sampler

Bi-variate Data – Scatter graph

Bi-variate data is best represented on a scatter graph

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 6

Bi-Variate Data

Examples of correlation coefficients.

Mike's carrot sprouts: $r = 0.4$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 11

Bi-variate Data - Description

Bi-variate data consists of pairs of values of two different variables associated with the same object. e.g.

- the heights and weights of members of a netball team.
- Each member of the team is represented by a pair of values (h, w)

$h =$ height $w =$ weight.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 2

Bi-variate Data – Correlation

Correlation is a measure of how closely the data follows a linear pattern.

To measure correlation we use the “product moment correlation coefficient”

This is usually called simply the “correlation coefficient”

The correlation coefficient (r) can take values in the range $-1 \leq r \leq 1$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 7

Examples of correlation coefficients.

Jenny's swede sprouts: $r = 0$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 12

Bi-variate Data - Description

Bi-variate data consists of pairs of values of two different variables associated with the same object. e.g.

The number of Km per day travelled by a patrol car, and the daily number of tickets issued.

- Each day is represented by a pair of values (K, t)

$K =$ Kilometers travelled
 $t =$ Number of tickets issued.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 3

Bi-variate Data – Correlation

$r = 0$ means the two variables are completely independent of each other.

$r = 1$ means the variables are directly related, and lie on a defined line with positive slope.

$r = -1$ means the variables are inversely related, and lie on a defined line with negative slope.

For example, consider these results of a Biology experiment.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 8

Examples of correlation coefficients.

Paul's mushroom spores: $r = -0.5$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 13

The Two Variables:

One is called the **explanatory variable** (Plotted on the horizontal \leftrightarrow axis)

The other is called the **dependent variable** (Plotted on the vertical \updownarrow axis)

- The explanatory variable is often (but not always) the one that can be controlled or selected.
- The explanatory variable is often (but not always) the one that may affect the other variable.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 4

Examples of correlation coefficients.

Amy's bean sprouts: $r = 0.9$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 9

Examples of correlation coefficients.

Kate's toadstool spores: $r = -0.9$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 14

Bi-variate Data - Description

Bi-variate data consists of pairs of values of two different variables associated with the same object. e.g.

- Number of customers each day at a dairy and the daily takings.

Each day is represented by a pair of values (c, t)

$c =$ number of customers
 $t =$ daily takings.

Explanatory (points to c)

Dependent (points to t)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 5

Examples of correlation coefficients.

Tom's bean sprouts: $r = 0.7$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 10

Examples of correlation coefficients.

Clyde made up his results: $r = 1$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 15

S3.5

Bi-Variate Data

Examples of correlation coefficients.

Sam made up his results for mushrooms:

Height of sprout $r = -1$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 16

Scatter graph of the data:

(1,4), (5, 4), (7, 10), (11, 10)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 21

Alternative Formula:

The coefficient of correlation (r) is given by the following alternative formula:

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

Note that this form can be used without first calculating the means. It is used in limited memory devices such as scientific calculators.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 26

Significance of (r)

In general,

If $|r| > 0.9$ we say that "there is a very strong linear relationship between the two variables."

If $|r| < 0.25$ we say that "there is very little or no linear association between the two variables."

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 17

Calculating the coefficient of correlation (r)

Example:

Calculate the coefficient of correlation for the following pairs of data:

(1,4), (5, 4), (7, 10), (11, 10)

First, create a table with the values needed for the formula:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 22

Alternative Formula for (r).

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

Notice that the calculator only needs to keep seven running totals in memory $\sum x$, $\sum x^2$, $\sum y$, $\sum y^2$, $\sum(xy)$, $\sum(xy)^2$ and n to calculate the coefficient for any number of pairs.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 27

Significance of (r) – More detail...

$|r| > 0.9$ **very strong** linear relationship

$0.75 < |r| < 0.9$ **strong** linear relationship

$0.5 < |r| < 0.75$ **moderate** linear relationship

$0.25 < |r| < 0.5$ **weak** linear relationship

$|r| < 0.25$ **no significant** linear relationship

Exercises R-01, R-02, R-03 or 13.01, 13.02

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 18

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \quad (1,4), (5, 4), (7, 10), (11, 10)$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
1	4	-5	-3	25	9	15
5	4	-1	-3	1	9	3
7	10	1	3	1	9	3
11	10	5	3	25	9	15
24	28	Totals		52	36	36
6	7	Means				

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 23

Deceptive situations (1):

r is a measure of how one variable varies in a linear relation to the other.

An obvious pattern does not always indicate a high value of the coefficient of correlation (r)

A horizontal or vertical trend indicates that there is no relationship, and so r is (close to) zero.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 28

Calculating the coefficient of correlation (r).

The coefficient of correlation (r) is given by the following formula:

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}}$$

Obviously this would be a tedious process without a computer.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 19

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sqrt{\sum(x - \bar{x})^2 \sum(y - \bar{y})^2}} \quad (1,4), (5, 4), (7, 10), (11, 10)$$

x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(y - \bar{y})^2$	$(x - \bar{x})(y - \bar{y})$
24	28	Totals		52	36	36
6	7	Means				

Now substitute into the formula $r = \frac{36}{\sqrt{52 \times 36}}$

Exercise R-04 or 13.03 $r = 0.83$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 24

For example: Rata's bean sprouts: $r = 0$

Height of sprout

Minutes spent on measuring the plants

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 29

Calculating the coefficient of correlation (r)

Example:

Calculate the coefficient of correlation for the following pairs of data:

(1,4), (5, 4), (7, 10), (11, 10)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 20

Scatter graph of the data ($r = 0.83$):

(1,4), (5, 4), (7, 10), (11, 10)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 25

For example: Jenny's bean sprouts: $r = 0$

Height of sprout

Soil depth (cm)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 30

S3.5

For example: Kelly's shrubs: $r = 0.13$

r is a measure of a linear relationship only

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 31

Causality

For example, consider an analysis of some variables that vary with city size.

Bigger cities have more accidents, more police cars, more doctors, more teachers, more theatres, etc.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 36

Bi-Variate Data

Bi-Variate Data & Regression

3.5 90645

Select and analyse continuous bi-variate data

www.ppresources.com
PC Sampler

Correlation

Correlation can only be measured for numeric data.

For example you could not measure the correlation between being left-handed and being left-footed.

Nor could you measure the correlation between whether a vehicle is a car or motorcycle and the speed recorded.

Both variables must be numeric data

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 32

Correlation doesn't imply causality.

Does a large population cause accidents?

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 37

The regression line.

A regression line enables us to make approximate predictions about the value of the dependent variable, if we know the value of the explanatory or independent variable.

The regression line is sometimes referred to as "The line of best fit" when it is placed visually rather than calculated.

Reminder!

The terms "dependent variable" and "explanatory variable" do not imply that one causes the other.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 42

Correlation

The value of the Correlation Coefficient is independent of the order in which the values are calculated.

The link between any pair of values must be retained.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 33

Be careful about causality.

Larger population does "cause" more teachers

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 38

Calculating the Regression line: ("The Line of Best Fit")

A regression line shows the relationship between the variables.

Exercise R-07 or 14.01

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 43

Correlation

Correlation will have the same value regardless of the units used to measure the variables.

Correlation will have the same value regardless of which variable is X and which is Y.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 34

Causality?

Sometimes both are caused by a 3rd variable

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 39

Calculating the Regression line:

Consider the the vertical distance of each point from the line.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 44

Causality

Correlation is a measure of how closely the data follows a linear pattern.

A high level of correlation ($r = 1$ or $r = -1$) does not imply that one variable causes the variation in the other.

Although one is called the independent or explanatory variable and the other the dependent variable, the issue of what causes the relationship is not part of regression analysis.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 35

Correlation vs. Causality.

Note that:

Number of teachers is a predictor for city size.

City size is a predictor for the number of road accidents.

Therefore number of teachers is a predictor for the number of road accidents.

This does not imply that teachers cause road accidents.

Exercises R-05, R-06 or 13.04, 13.05

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 40

Residuals.

A residual is the difference between the value of the dependent variable (y) and the value predicted by the equation of the regression line (\hat{y})

i.e. residual = $(y - \hat{y})$

We represent residuals with vertical bars which extend above or below the regression line.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 45

S3.5

Bi-Variate Data

Calculating the Regression line:

A regression line is placed so that the sum of the squares of the residuals is a minimum.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 46

Calculating the regression line. Example (continued):

Calculate the equation of the regression line for the following pairs of data:
(1,4), (5, 4), (7, 10), (11, 10)

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \quad a = \frac{\sum y - b\sum x}{n}$$

First, create a table with the values needed for the formulae:

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 51

Caution: Extrapolating a trend

Do not assume that a trend or regression can be extrapolated beyond the range of the data.

For example,
Amy's bean sprouts will probably not reach 3 metres in height after 150 days.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 56

Equation of the regression line:

Assume that the equation of the regression line takes the form " $y = a + bx$ "

This is called linear regression.

a is called the regression coefficient.

b is the gradient of the regression line.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 47

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2}$$

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
1	4	-5	-3	25	15
5	4	-1	-3	1	3
7	10	1	3	1	3
11	10	5	3	25	15
24	28	Totals		52	36
6	7	Means			

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 52

Coefficient of determination (R^2)

The coefficient of determination (R^2) indicates the proportion of the variation in the dependent variable that can be explained by the regression function

Note: $0 \leq R^2 \leq 1$

If every pair of data points fits the function exactly, then $R^2 = 1$

For completely random data $R^2 \approx 0$.

A function is often considered a good fit if $R^2 > 0.9$ (Depends on situation)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 57

Equation of Linear Regression Line.

$$y = a + bx$$

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} \quad a = \frac{\sum y - b\sum x}{n}$$

$$b = \frac{(s_{xy})^2}{(s_x)^2} \quad a = \bar{y} - b\bar{x}$$

Again, these calculations are better carried out on a spreadsheet, a graphical calculator, or using statistics software.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 48

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
(1,4), (5,4), (7,10), (11,10)					
24	28	Totals		52	36
6	7	Means			

Now substitute:

$$b = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})^2} = \frac{36}{52} = 0.692$$

$$a = \frac{\sum y - b\sum x}{n} = \frac{28 - 0.692 \times 24}{4} = 2.85$$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 53

Calculating the Coefficient of determination (R^2)

The value of R^2 is calculated by comparing the squares of the residuals with the variance of y

$$R^2 = \frac{\sum(y-\hat{y})^2 - \sum(y-\bar{y})^2}{\sum(y-\bar{y})^2}$$

Where \hat{y} is the predicted value of y, and \bar{y} is the mean value of y

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 58

Calculating the regression line using the formula. Example:

Calculate the equation of the regression line for the following pairs of data:

(1,4), (5, 4), (7, 10), (11, 10)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 49

x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})^2$	$(x-\bar{x})(y-\bar{y})$
(1,4), (5,4), (7,10), (11,10)					
24	28	Totals		52	36
6	7	Means			

$$b = \frac{36}{52} = 0.692 \quad a = 2.85$$

Now substitute into $y = a + bx$

$$y = 2.85 + 0.692x \quad (3 \text{ s.f.})$$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 54

The Correlation Coefficient (r) and the Coefficient of Determination (R^2)

The Coefficient of Determination (R^2) determines how well a given regression line fits the data. Different regression lines will give different values of R^2 .

The Correlation Coefficient (r) is a measure of the strength of the linear relationship between the bi-variate data pairs.

The Correlation Coefficient (r) is independent of any regression line, and depends only on the data.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 59

Scatter graph of the data:

(1,4), (5, 4), (7, 10), (11, 10)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 50

Scatter graph of the data ($r = 0.83$):

(1,4), (5, 4), (7, 10), (11, 10)

$$y = 2.85 + 0.692x \quad (3 \text{ s.f.})$$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 55

Interpreting the coefficient of determination

When trying to determine the ability of one variable (x) to predict a second variable (y), R^2 gives a measure of how much the regression function explains the variation in y.

For example if $R^2 = 0.87$, we would suggest that 87% of the variation in y can be accounted for (predicted) by the regression function.

N.B. This is not saying that 87% of the variation in y is caused by variation in x.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 60

S3.5

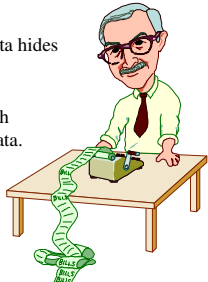
Bi-Variate Data

Coefficient of determination (R^2) and the correlation coefficient (r) – Note:
 The **coefficient of determination** (R^2) depends on the trend line chosen.
 The trend line can be curved or straight.
 The coefficient of determination (R^2) will be different for each trend line.
 The **correlation coefficient** is a measure of how well the data can fit a **straight line**.
 The correlation coefficient depends only on the data and does not relate to a trend line.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 61

Outlier Effects
 Sometimes outlier data hides the true pattern.

This is less likely with large amounts of data.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 66

Outlier Effects
 Of course we must not remove an outlier just to get a particular result.

However, we should investigate why an outlier does not fit the trend of the other data.

Incorrect measurement?
 Measured under different conditions?
 Clerical error?
 Random variation? Exercise R-10 or 14.03

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 71

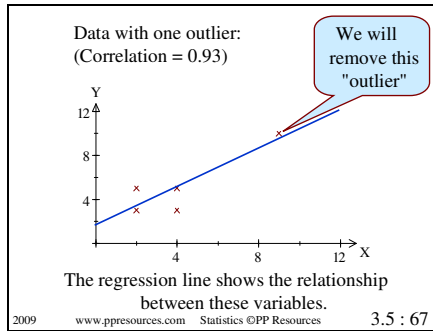
Residuals. $(y - \hat{y})$

A residual is the difference between the value of the dependent variable and the value predicted by the equation of the regression line.

Note: If R^2 is close to 1 then the residuals will be small.

After calculating the regression function, predicted values and the residuals should be calculated.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 62



Non-linear regression
 Sometimes data follows a non-linear pattern.

For example, the data may follow an exponential curve, or a power curve.

Sometimes the data may follow a different pattern under different values of the independent variable (piecewise regression).

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 72

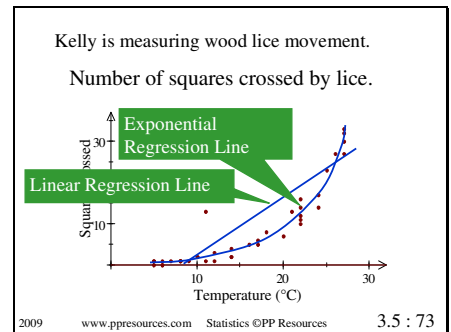
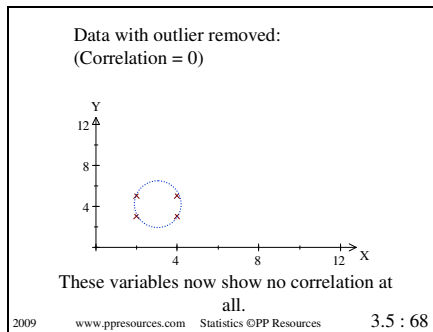
For Example: Earlier we calculated the regression equation for this data.

(1,4), (5, 4), (7, 10), (11, 10)

$$y = 2.85 + 0.692x \text{ (to 3 s.f.)}$$

Now we need to calculate the predicted values and the residuals.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 63

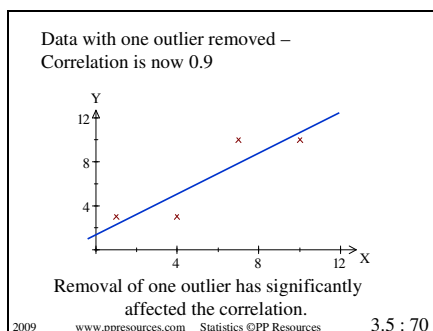
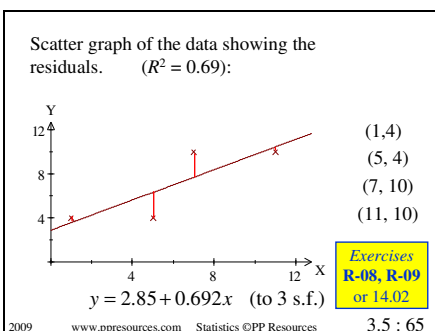
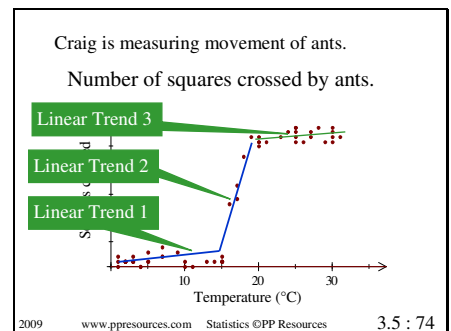
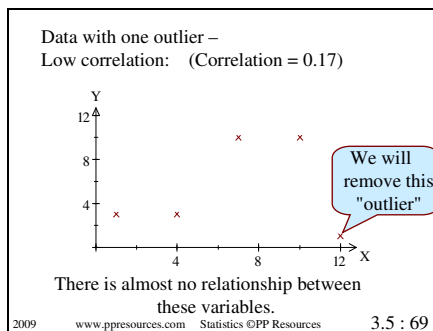


$y = 2.85 + 0.692x \text{ (to 3 s.f.)}$

x	y	Predicted	Residuals
1	4	3.54	0.46
5	4	6.31	-2.31
7	10	7.69	2.31
11	10	10.46	-0.46
Total residuals			0

Note that for a correctly placed regression line the sum of the residuals should be small.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 64



The Process:

Graph the data on a scatter plot
 Visually examine the points to:

- look for a pattern (curved or straight?)
- Ask why you might expect a particular pattern.
- Look for outliers

Use technology to find a trend line

- Use the R^2 value as a **guide only** to choosing the most appropriate trend line.

Write up your report.

Why? [Click here](#)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 75

S3.5

Bi-Variate Data


Humans are:

- Slow
- Inaccurate
- Capable of intelligent thought.

Computers are:

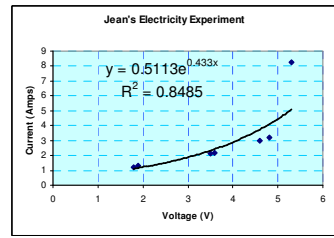
- Fast
- Accurate
- Brainless, mindless, unintelligent.

Your challenge, when using technology, is to combine the best features of both.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 76

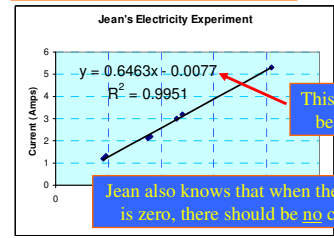
Jean's Electricity Project



Jean tries an exponential trend line. $R^2 = 0.85$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 81

More intelligent stuff needed yet!!



This should be zero.

Jean also knows that when the voltage is zero, there should be no current.

Jean adds a linear trend line. $R^2 = 0.995$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 86

Caution when using technology to choose the most appropriate trend line:

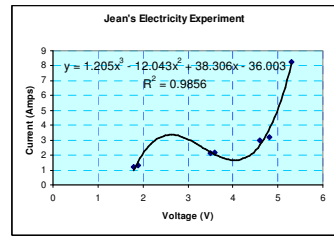
Use your intelligence to choose the best trend line.

Use research to see what shape you would expect the trend line to be:

- Growth or decay – expect exponential
- Scientific experiments – Look it up
- Sociology – R^2 is often so low that there is no reason not to use a linear trend.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 77

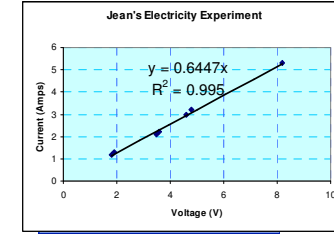
Jean's Electricity Project



Jean tries a polynomial (x^3) trend line. $R^2 = 0.99$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 82

Jean's formats the trend line to pass through zero.



Jean has a linear trend line at last. $R^2 = 0.995$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 87

Caution when using technology to choose the most appropriate trend line:

Use your intelligence to choose the best trend line. Sometimes a computer fits an inappropriate trend line with a high value of R^2 . You need to ask whether this trend is likely to apply for larger or smaller values, or whether such a trend is reasonable.

For example, if four points lie close to a straight line ($R^2 = 0.9$). The computer could fit a cubic function exactly ($R^2 = 1$) Even though this would not be a sensible option.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 78

Jean's Electricity Project

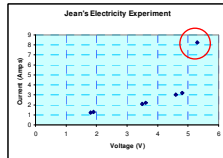
Jean knows enough science to know that this is not the usual connection between voltage and current.

So Jean re-examines the raw data.

One of the data points stands out as being very different from the others.

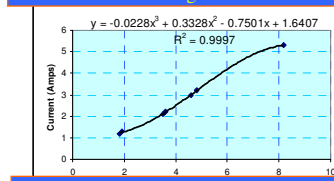
i.e an outlier.

So she checks her raw data.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 83

However a polynomial trend would show a lower current as voltage continues to rise.

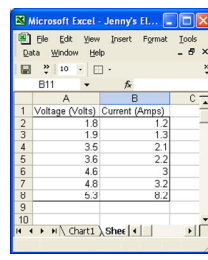


For this reason Jean chooses a linear trend line.

A polynomial trend line fits even better. $R^2 = 0.9997$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 88

Jean's Electricity Project



Jean is recording the electrical current through a resistor for a range of voltages.

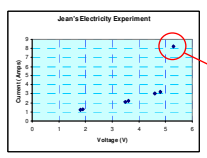
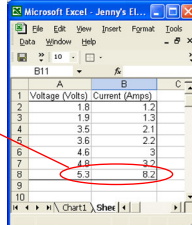
To find the relationship she will plot the data and find a trend line.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 79

Jean's Electricity Project

Jean finds that she has entered her last data point reversed.

The last point (5.3, 8.2) should be (8.2, 5.3)

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 84

To Summarise the Process:

Graph the data on a scatter plot

Visually examine the points to:

- look for a pattern (curved or straight?)
- Ask why you might expect a particular pattern.
- Look for outliers

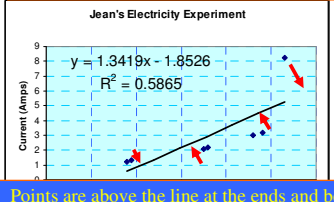
Use technology to find a trend line

- Use the R^2 value as a guide only to choosing the most appropriate trend line.

Write up your report. **Exercise R-11 or 14.04**

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 89

A visual check shows this is not a good fit.



Points are above the line at the ends and below it in the middle. This implies a curved trend line.

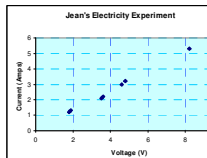
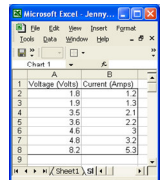
Jean adds a linear trend line. $R^2 = 0.59$

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 80

Jean's Electricity Project

So Jean could leave out this data point (outlier)


Jean chooses to enter the correct data and try again.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 85

S3.5

Bi-Variate Data




Bi-Variate Data – Producing a Report

3.5 90645
Select and analyse
continuous bi-variate data

www.ppresources.com
PC Sampler

For excellence, your report will include a critical evaluation:

- i.e. where appropriate, include:
 - » Limitations
 - » Possible improvements
 - » Justification of method
 - » Alternative approaches
 - » Assumptions made
 - » Potential sources of bias
 - » Relevance and usefulness of evidence
 - » How widely the findings can be applied




2009 www.ppresources.com Statistics ©PP Resources 3.5 : 95

Finally – a few practical considerations:

1. From the start, on every page place a footer which contains your name, the page number and the title of the report.

Not only does this help with sorting and identifying your assignment, it is very useful when many students share a network printer.


2. Graphs have minimum requirements such as a full title, labels, etc. Do not assume that the computer will do all of this for you.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 100

An investigation of bi-variate data should (minimum requirements):

- Be an interesting task. (fun!)
- Start with a plan, which includes:
 - » A purpose statement
 - » Identification of appropriate variables (May be more than one pair)
 - » A description of the data collection method or the source of the data.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 91

Presenting your report

When presenting or discussing results, keep in mind the limitations of the study.

Do not generalize the results beyond its scope. For example, a study that examines the effect of population on housing costs can only conclude that housing costs might be associated with population.

It cannot establish cause and effects.


2009 www.ppresources.com Statistics ©PP Resources 3.5 : 96

Even more practical considerations:

3. **ALWAYS** keep a backup of your computer files. (e.g. One on the school network, one at home and one on a memory stick.)

Although computers do not have brain, they seem to know when you are stressed, and this is when things will go wrong.

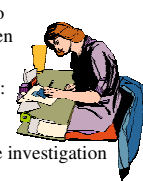
Keep a backup of every file you create.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 101

An investigation of bi-variate data should (bare minimum requirements):

- Involve the use of regression to explore the relationship between pairs of variables
- Finish with a report which will:
 - » Relate to the context
 - » Address the purpose of the investigation
 - » Describe the relationship between at least one pair of variables.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 92

The Findings:

The final step of a study is to communicate its findings to others.

The report should:

- Clearly describe the background of the experiment, and/or the source of the data
- State the model on which the analysis was based.
- State the assumptions made in the analysis.
- Identify the variables used in the analysis.
- Justify the exclusion of any data from the analysis (e.g., outliers).
- Provide reasons for doing subgroup analysis.


2009 www.ppresources.com Statistics ©PP Resources 3.5 : 97

One last practical consideration:

4. The most likely time for computer problems is the day before it is due!

That is when the printer at home will run out of ink and the school network is down.

The remedy is –
Start early and aim to finish early.



2009 www.ppresources.com Statistics ©PP Resources 3.5 : 102

An investigation of bi-variate data should (merit requirements):

- Be an in-depth analysis of bi-variate data.
- “In depth” means you should:
 - » Establish a model using regression
 - » Interpret correlation coefficients
 - » Interpret coefficient of determination
 - » Interpolate and extrapolate
 - » Discuss the appropriateness of the model
 - » Discuss the effect of outliers
 - » Consider piecewise or non-linear models
 - » Discuss causality and correlation

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 93

The Report

The report should provide sufficient quantitative information to allow readers to evaluate the appropriateness of the analysis and to draw their own interpretations.

The report should include summary statistics (e.g., mean, standard error, and sample size) to show the data structure

The report should include sufficient and appropriate graphs.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 98

An investigation of bi-variate data should (merit requirement):

- Be consistent with the analysis
- » “Consistent” means that what you describe in English should already be in your mathematical analysis, the data, or the graphs.
- » Don’t “waffle” (don’t write to fill the page)
- » Guideline: If you are unable to make your point in two sentences, you possibly do not know what you are talking about!

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 94

Regression Analysis

Explain why a particular equation was chosen for the final regression (R^2 closest to 1 ?).

Interpret the form of the regression equation.

Give the range of the x -variable for which predictions are valid.

Extrapolate regression results beyond the range exhibited in the original data when justified.

Plot raw data and fitted equation on the same graph to demonstrate the fit of the regression model.

2009 www.ppresources.com Statistics ©PP Resources 3.5 : 99