Advanced Mathematics Support Programme®

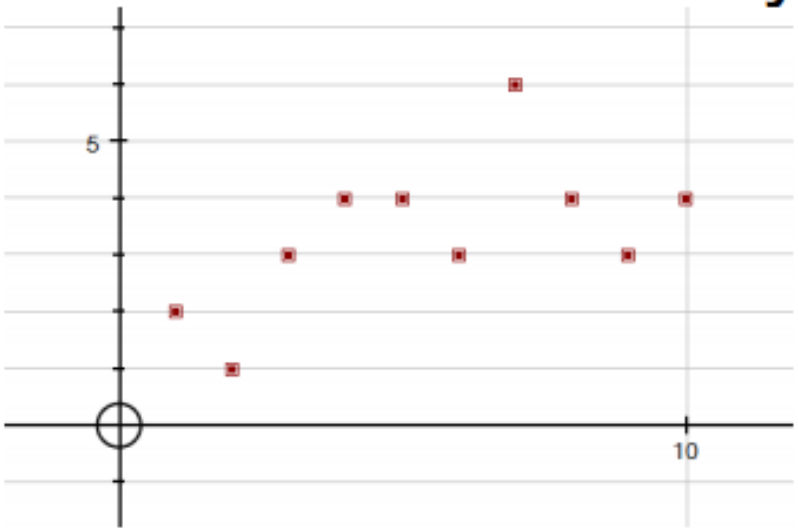
Overview of bivariate data in FM

- Scatter Diagrams
- PMCC
- Spearman's Rank CC
- Hypothesis Tests for PMCC
- Hypothesis Tests for SRCC
- Regression Lines
- Residuals.

(OCR year 1; MEI Statistics a; Edexcel FS2 year 1)

MSV 17: Ten point PMCC

Can you give ten points on a scatter diagram that give a PMCC of exactly 0.99?



www.making-statistics-vital.co.uk

Product Moment Correlation

The product moment correlation coefficient is given by $\frac{1}{n}\sum x'y'$ where x' and y' are the standardised values of x and y

That is, $x' = \frac{x-\bar{x}}{s_x}$ and $y' = \frac{y-\bar{y}}{s_y}$ *

It's the mean value of the products of the x and y deviations from their means, measured in standard units.

Pearson's product moment correlation coefficient, for a bivariate sample of size *n*:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

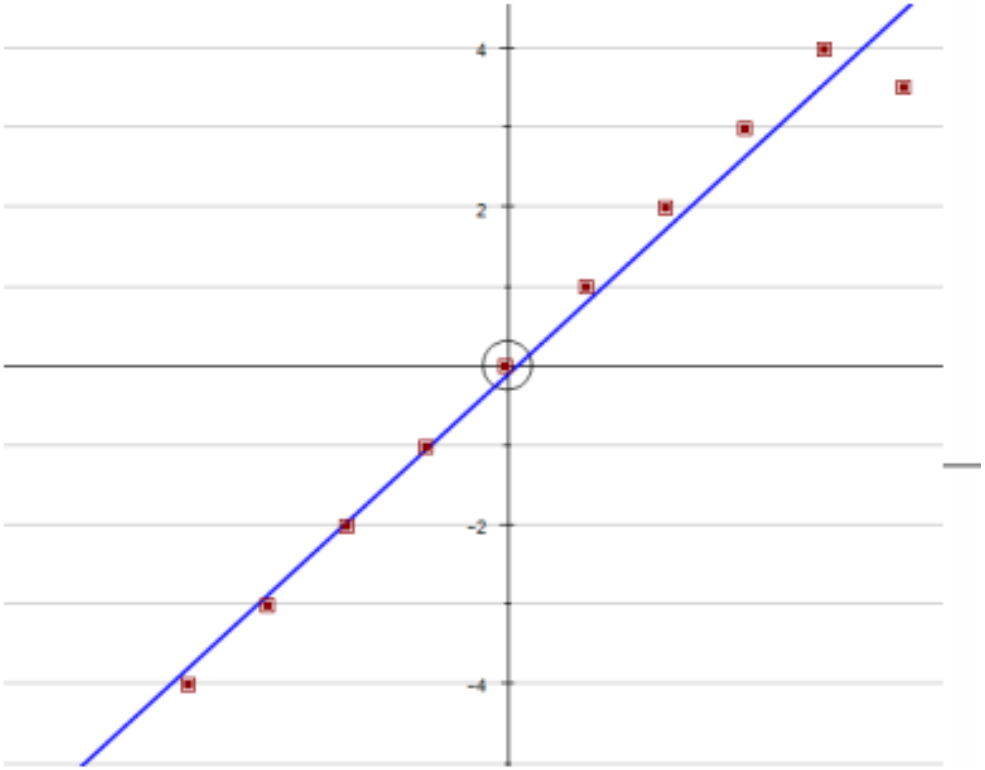
where $S_{xx} = \Sigma(x-\bar{x})^2 \equiv \Sigma x^2 - n\bar{x}^2 \equiv \Sigma x^2 - \frac{(\Sigma x)^2}{n}$

$S_{yy} = \Sigma(y-\bar{y})^2 \equiv \Sigma y^2 - n\bar{y}^2 \equiv \Sigma y^2 - \frac{(\Sigma y)^2}{n}$

$S_{xy} = \Sigma(x-\bar{x})(y-\bar{y}) \equiv \Sigma xy - n\bar{x}\bar{y} \equiv \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n}$

Hint:

Consider one point slightly out?



Let's choose the points (-4, -4), (-3, -3), (-2, -2), (-1, -1), (0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, k).

These give $\Sigma x = 5, \Sigma y = k, \Sigma x^2 = 85, \Sigma y^2 = 60+k^2, \Sigma xy = 60 + 5k.$

$$S_{xx} = 165/2, S_{yy} = 60 + 0.9k^2, S_{xy} = 60 + 4.5k.$$

So $\frac{99}{100} = \frac{60+4.5k}{\sqrt{(165/2)(60+0.9k^2)}}$, which gives $105.04485k^2 - 1080k + 2502.99=0$.

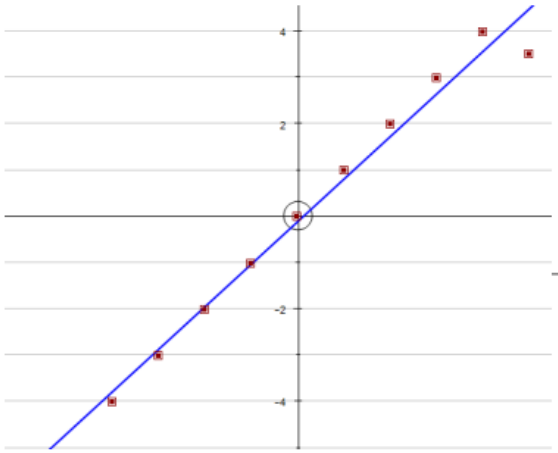
Let's choose the points (-4, -4), (-3, -3), (-2, -2), (-1, -1), (0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, k).

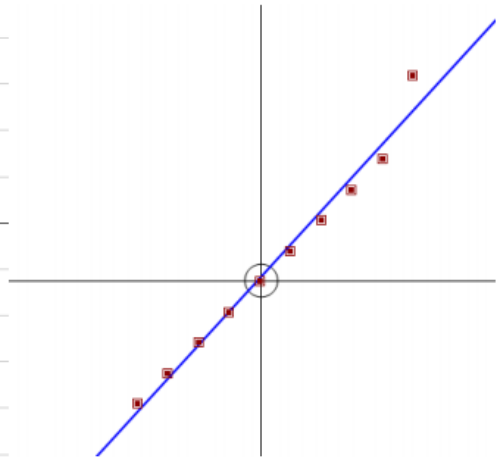
These give $\Sigma x = 5, \Sigma y = k, \Sigma x^2 = 85, \Sigma y^2 = 60+k^2, \Sigma xy = 60 + 5k.$

$$S_{xx} = 165/2, S_{yy} = 60 + 0.9k^2, S_{xy} = 60 + 4.5k.$$

So $\frac{99}{100} = \frac{60+4.5k}{\sqrt{(165/2)(60+0.9k^2)}}$, which gives $105.04485k^2 - 1080k + 2502.99=0$.

So k = 3.53 OR k = 6.75, giving the two scatter diagrams below.





Regression

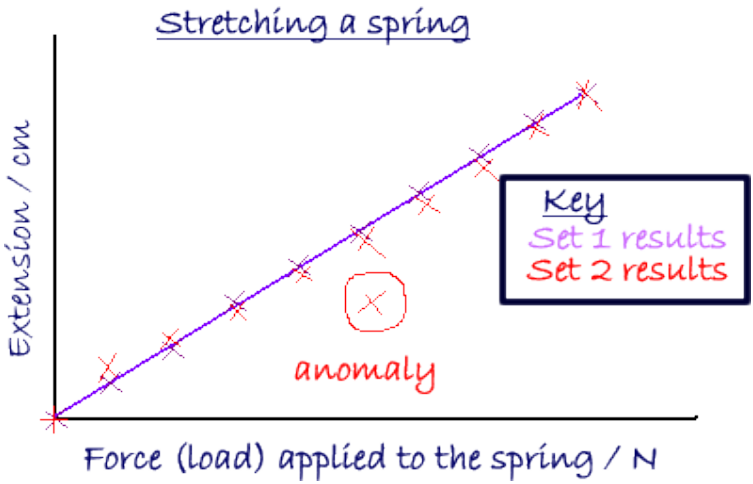
- The most fundamental idea in regression is that a relationship between two variables can be simplified by using a model – a mathematical function – which does two things:
  - It fits the data (reasonably well)
  - It makes sense in terms of what is being modelled

Regression

- The simplest regression model is **linear**. That is, the two variables satisfy a linear relationship.
- Sometimes one variable, the dependent variable, can be thought of as dependent on the other, the independent variable. Let's call this **Case A**.
- Sometimes neither variable is independent: they both depend, to some extent, on the other – or on some additional factors. Let's call this **Case B**.

Regression, Case A – Example

- Clearly, the extension *y* is not observed exactly – it's actually quite difficult to measure the length of a spring to great accuracy.
- But the applied load *x* can be measured exactly – or so nearly exactly as makes no difference. The load is applied using standard calibrated weights.
- So the reason that the data points don't lie exactly on a line is because of the error.
- And the error is in the *y* direction.

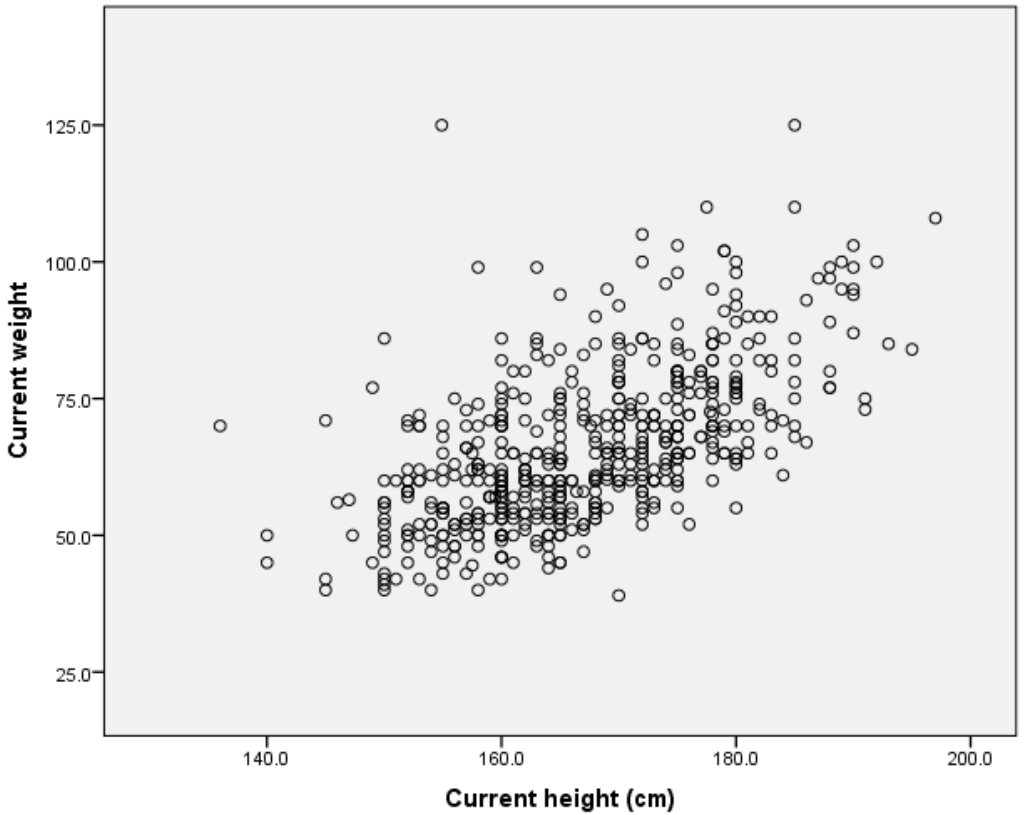


Regression, Case B

- We suppose that ***X* and *Y* are both random variables.**
- The observed data points lie in some sort of data cloud.
- And we suppose that neither can be thought of as directly dependent on the other.
- Both *X* and *Y* represent some real-world variation.

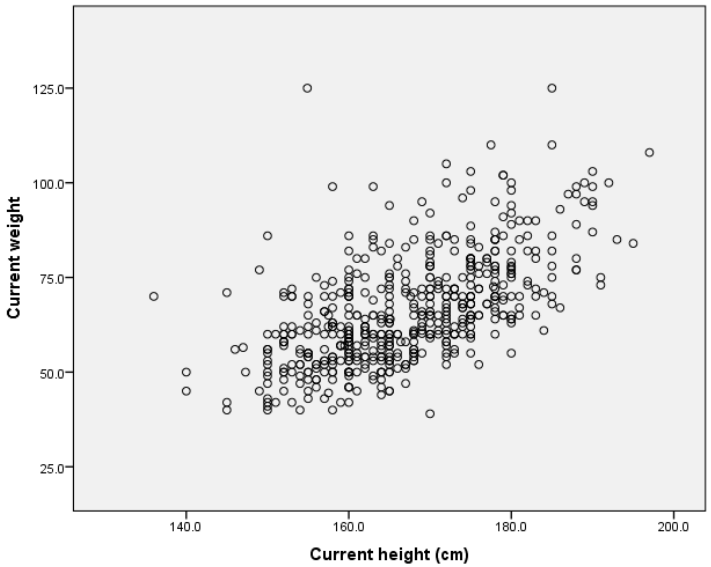
Regression, Case B – Example

- Heights and weights of a random sample of individuals form a data cloud.



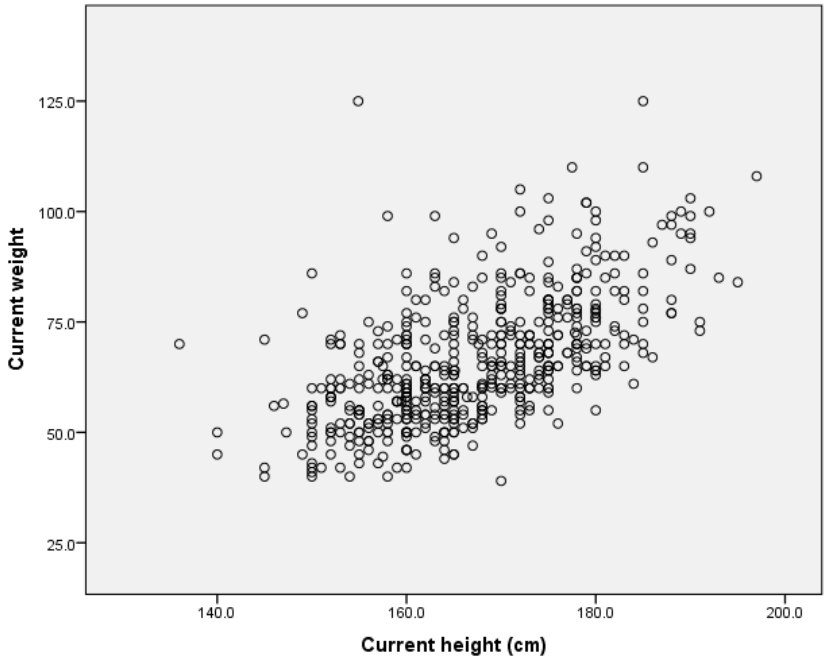
Regression, Case B – Example

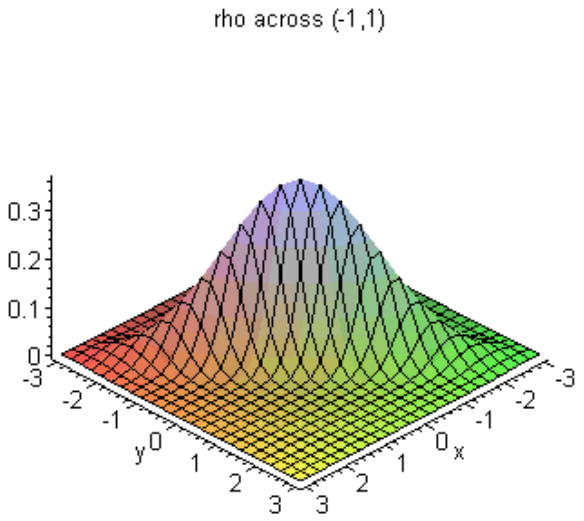
- Clearly, the weight *y* is not a function of the height *x*. For any given value of *x* there is a whole distribution of weights *y*.
- Likewise, there is no 'true' height for as given weight. For any given value of *y* there is a whole distribution of heights *x*.
- So the reason the data points don't lie on a line is not because of error. It's because of natural variation in the population.



Bivariate normal

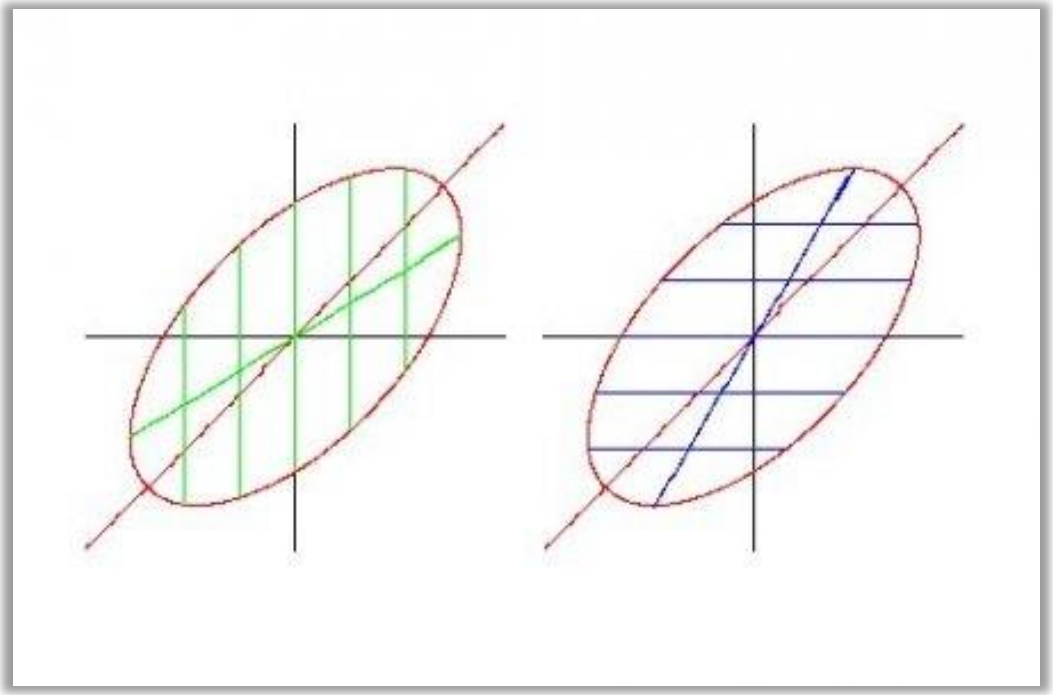
- For each x, the y's are normal with equal variances.
- For each y, the x's are normal with equal variances.





Regression – Case B

- When the population is bivariate Normal, the data cloud has a broadly elliptical shape.
- Under those conditions the mean of Y is a linear function of the value of X.
- $E(Y \mid X = x) = a + bx$
- We can write this as $y = a + bx$ where $y = E(Y \mid X = x)$.



Regression – Combining the cases

- So regression can mean
  - finding the 'true' underlying linear relation between the dependent variable *y* and the independent variable *x* (case A), *or*
  - finding how the expectation of one variable depends on the value of the other (case B).
- The equation $y = a + bx$ means different things in the two cases, but
- Provided in case B the underlying distribution is bivariate Normal, the formulae for *a* and *b* are the same as in case A.

Regression

- Note that specifications vary.
  - Some deal only with case A
  - Some deal only with case B
  - Some deal with both
- Those that deal with both often fail to make clear distinction between them.
- In particular, it is quite common in case B to be asked to estimate or predict a value of one variable given the value of another …

Regression

- … for example:

Find the regression line for weight on height (from a given set of data for adult males). Hence estimate the weight of an adult male of height 1.8m.

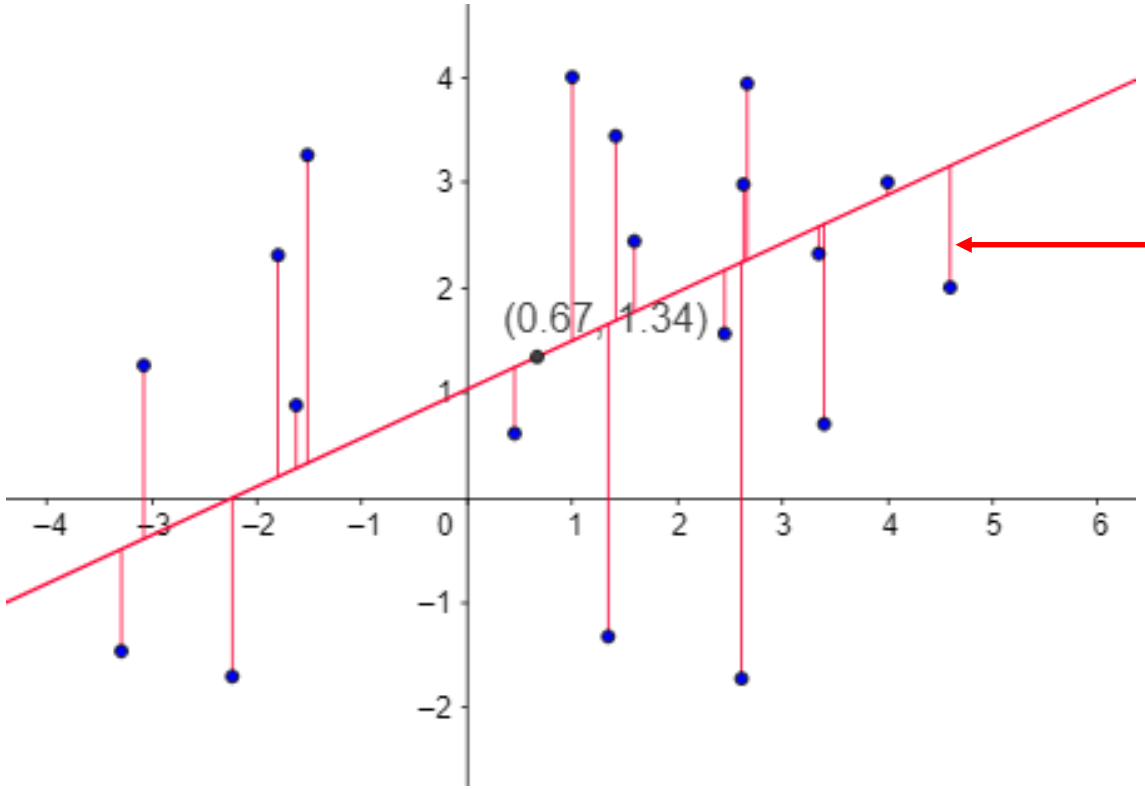
Can you see what is wrong with that?

Case A or B?

x	y
Life expectancy of a country	Birth rate of a country
Gestational age at birth (weeks)	Birth Mass (g)
Load applied to a spring (N)	Extension in spring (mm)
English test scores for a year group (%)	Maths test scores for year group (%)
GDP of a country	Population of a country
Amount of rainfall (mm)	Crop yield (kg)

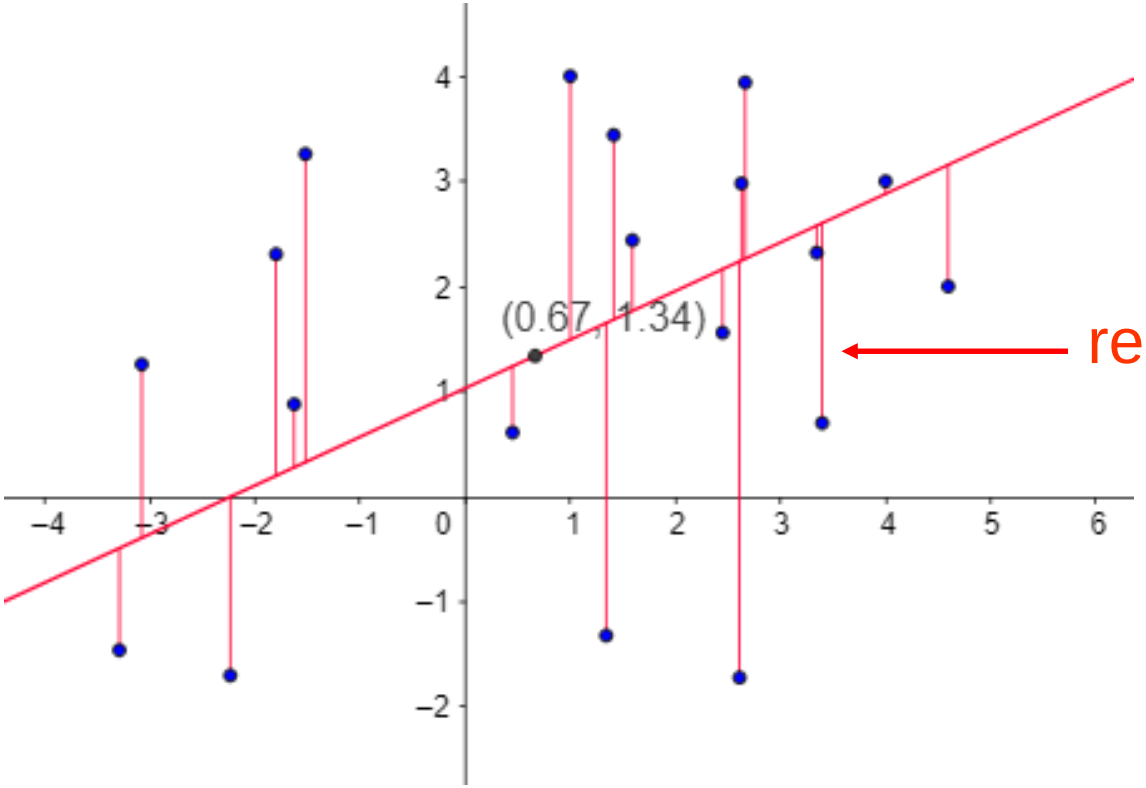
The Regression Line

Aims to minimize the distances between points and the line of best fit.



The Regression Line

In particular the sum of the squares of the residuals is minimized.



The Regression Line

In particular the sum of the squares of the residuals is minimized.

https://www.geogebra.org/m/kWA7IcSE

Y on X regression line y = a +bx

Each residual is of the form:

$$y_i - (a + bx_i)$$

Y on X regression line y = a +bx

Each residual is of the form:

$$y_i - (a + bx_i)$$

So the sum of the squares of the residuals is:

$$S = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2$$

Y on X regression line y = a +bx

Each residual is of the form:

$$y_i - (a + bx_i)$$

So the sum of the squares of the residuals is:

$$S = \sum_{i=1}^{n} \left(y_i - (a + bx_i)\right)^2$$

By partial differentiation with respect to both a and b you can show…..

Y on X regression line y = a +bx

$\bar{y} = a + b\bar{x}$

So line passes through $(\bar{x}, \bar{y})$.

Y on X regression line y = a +bx

$$\bar{y} = a + b\bar{x}$$

$$b\sum x^2 = \sum xy - a\sum x$$

Y on X regression line y = a +bx

$$\bar{y} = a + b\bar{x}$$

$$b\sum x^2 = \sum xy - a\sum x$$

$$b = \frac{\sum xy - \frac{1}{n}\sum x \sum y}{\sum x^2 - \frac{1}{n}(\sum x)^2} = \frac{S_{xy}}{S_{xx}}$$

Regression line of X on Y

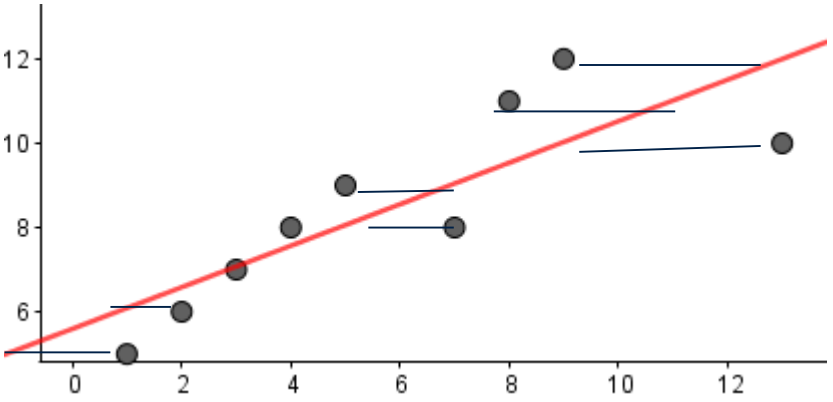
A similar process can be employed to find the regression line of X on Y.

This time we look to minimize the sums of squares of the horizontal distances.

Regression line of X on Y

A similar process can be employed to find the regression line of X on Y.

This time we look to minimize the sums of squares of the horizontal distances.



Once again $(\bar{x}, \bar{y})$ lies on the line and $b = \frac{S_{xy}}{S_{yy}}$

RSS (Residual Sum of Squares)

By substituting the values of a and b for the Y on X regression line back into the expression for the sum of the squares of the residuals, we will get:

$$RSS = S_{yy} - \frac{(S_{xy})^2}{S_{xx}}$$

Rank Correlation

- The (Spearman*) rank correlation coefficient is just the pmcc applied to the ranks of the data
- Formula: $1 - \frac{6\sum D^2}{n(n^2-1)}$ where *D* is difference in ranks
- Unlike the pmcc, the rank correlation does not require any particular underlying distribution. It is 'distribution free'. It is 'non-parametric'.
- It measures the amount of 'association'.

*other correlation coefficients are available

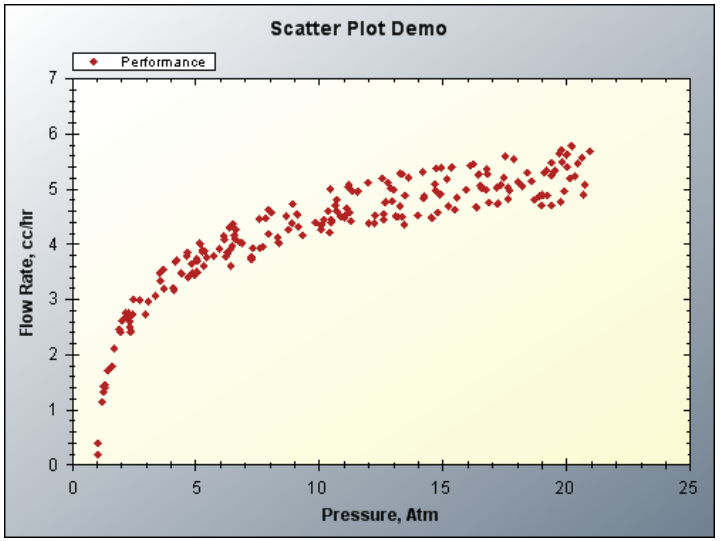
When would you use r_s?

When would you use r_s?

- If you are only given the ranks

When would you use r_s?

- If you are only given the ranks
- Where you suspect there is not a linear relationship but where one variable generally increases or decreases as other increases; in that case we are looking for association rather than correlation.



Gestational Age V Birth Mass

Infant ID #	Gestational Age (weeks)	Birth Mass (grams)
1	34.7	1895
2	36	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38	2680
17	38.7	2005

The data has been input for you at:

https://www.desmos.com/calculator/pznbviilj2

(x_1, y_1) are the original data and (x_2, y_2) are the ranks.

https://www.geogebra.org/classic/usphfxwu

Gestational Age V Birth Mass

Infant ID #	Gestational Age (weeks)	Birth Mass (grams)
1	34.7	1895
2	36	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38	2680
17	38.7	2005

(a) Give the values of r (pmcc) and r_s (spearman's cc)

(b) Give the equations of the two regression lines

(c) Describe the correlation and/or association, comparing the values of r and r_s

(d) Estimate the mean mass of an infant of age 39 weeks

(e) Estimate the mean age of an infant of mass 3kg.

Desmos

To get a regression line and r type:

$$y_1 \sim m x_1 + c$$

GeoGebra

Produces summary statistics

Swaps the Axes over

Gestational Age V Birth Mass

Infant ID #	Gestational Age (weeks)	Birth Mass (grams)
1	34.7	1895
2	36	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38	2680
17	38.7	2005

(a) Give the values of r (pmcc) and r_s (spearman's cc)

(b) Give the equations of the two regression lines

(c) Describe the correlation and/or association, comparing the values of r and r_s

(d) Estimate the mean mass of an infant of age 39 weeks

(e) Estimate the mean age of an infant of mass 3kg.

r V r_s

Can rank correlation be low but pmcc high?

Can rank correlation be high but pmcc low?

r V r_s

Can rank correlation be low but pmcc high?

Can rank correlation be high but pmcc low?

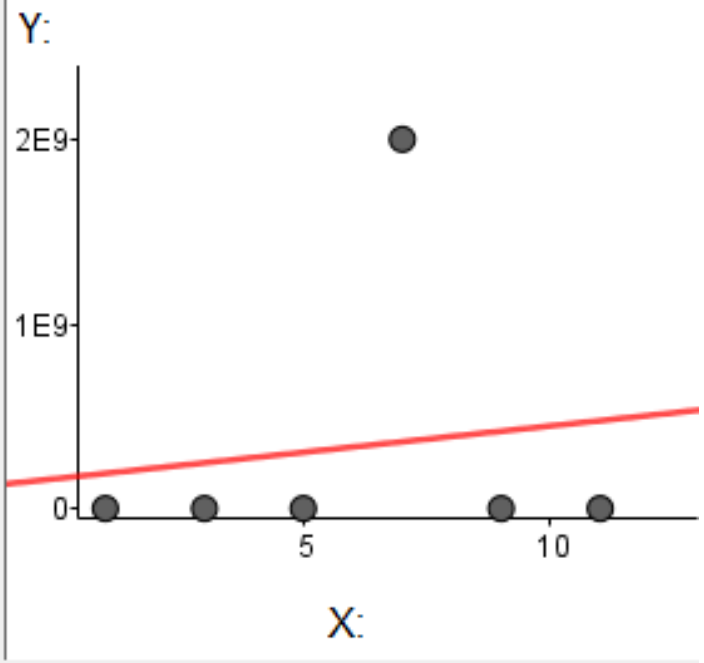
Provide some datasets to illustrate these.

A	B	C
1	0	
3	1	
5	7	
7	2000000000	
9	13	
11	15	

Statistics

MeanX	6
MeanY	333333339.3...
Sx	3.7417
Sy	816496577.9...
r	0.1309
ρ	0.8286
Sxx	70
Syy	33333333093...
Sxy	2000000104
R^2	0.0171

Scatterplot



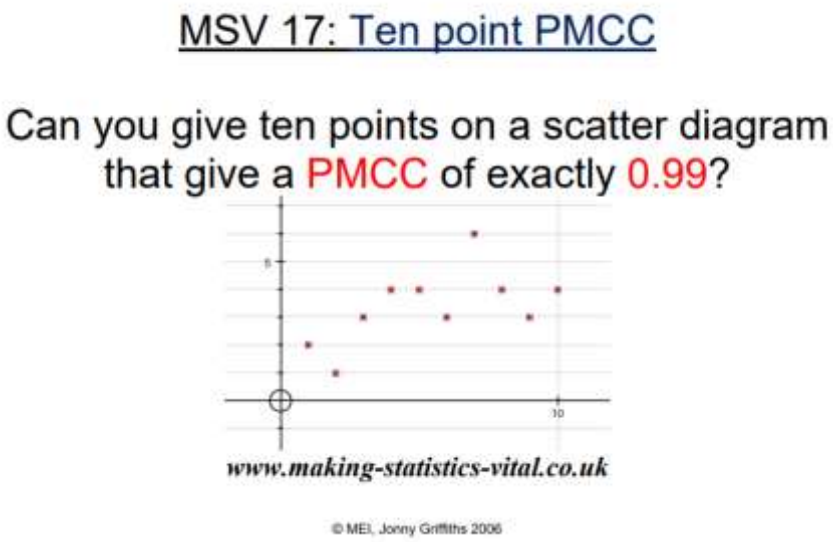
Regression Model

None

$y = 28571430.06\ x + 161904758.99$

Advanced Mathematics Support Programme®

Analysing Bivariate Data



Case A (Independent, dependent) or Case B (Independent, Independent)?

x	y
Life expectancy of a country	Birth rate of a country
Gestational age at birth (weeks)	Birth Mass (g)
Load applied to a spring (N)	Extension in spring (mm)
English test scores for a year group (%)	Maths test scores for year group (%)
GDP of a country	Population of a country
Amount of rainfall (mm)	Crop yield (kg)

Minimising the sum of the squares of the residuals

Type https://www.geogebra.org/m/kWA7lcSE into your browser window and then follow the instructions.

Gestational Age V Birth Mass

Infant ID #	Gestational Age (weeks)	Birth Mass (grams)
1	34.7	1895
2	36	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38	2680
17	38.7	2005

The data shows age (in weeks) at birth and birth mass (g) for 17 children born on the maternity ward of a local hospital.

Using technology, plot the data and answer the following

(a) Give the values of r (pmcc) and r_s (spearman's cc)
(b) Give the equations of the two regression lines
(c) Describe the correlation and/or association, comparing the values of r and r_s
(d) Estimate the mean mass of an infant of age 39 weeks
(e) Estimate the mean age of an infant of mass 3kg.

The data has been input for you at:

https://www.desmos.com/calculator/pznbviilj2

(x_1, y_1) are the original data and (x_2, y_2) are the ranks.

or

https://www.geogebra.org/classic/usphfxwu

Which is bigger r or r_s?

Is it possible to predict which of r or r_s will be numerically greater?

Can rank correlation be low but pmcc high?

Can rank correlation be high but pmcc low?