York Hospital: Introduction to Statistics for Research

Correlation and regression

Martin Bland

Emeritus Professor of Health Statistics

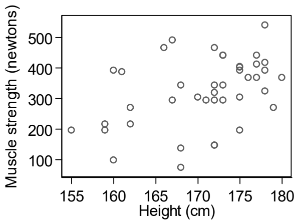
University of York

http://martinbland.co.uk/

Correlation

Example: Muscle strength and height in 42 alcoholics

A scatter diagram:

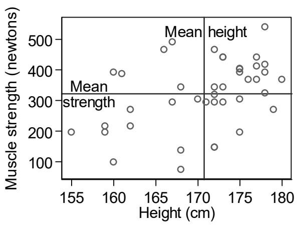


How close is the relationship?

Correlation: measures closeness to a linear relationship.

Correlation coefficient

Subtract means from observations and multiply.

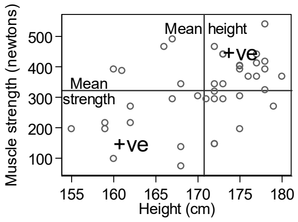


Sum of products about the means.

Like the sum of squares about the means used for measuring variability.

Correlation coefficient

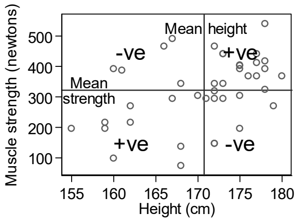
Subtract means from observations and multiply.



Products in top right and bottom left quadrants positive.

Correlation coefficient

Subtract means from observations and multiply.

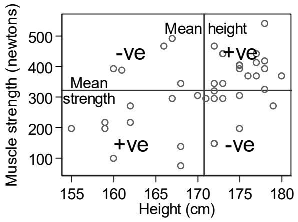


Products in top right and bottom left quadrants positive.

Products in top left and bottom right quadrants negative.

Correlation coefficient

Subtract means from observations and multiply.

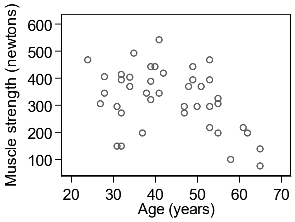


Sum of products positive.

Correlation positive.

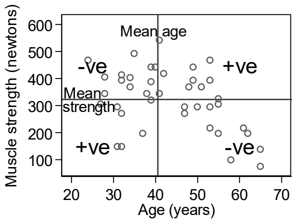
Correlation coefficient

Example: Muscle strength and age in 42 alcoholics



Correlation coefficient

Example: Muscle strength and age in 42 alcoholics



Sum of products negative.

Correlation negative.

Correlation coefficient

Divide sum of products by square roots of sums of squares.

Correlation coefficient, denoted by *r*.

Maximum value = 1.00.

Minimum value = –1.00.

Also known as:

- Pearson's correlation coefficient,
- product moment correlation coefficient.

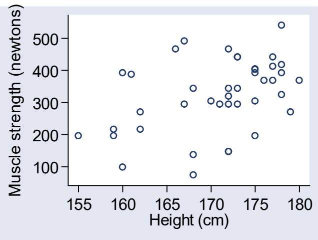
Correlation coefficient

Divide sum of products by square roots of sums of squares.

Correlation coefficient, denoted by *r*.

Maximum value = 1.00.

Minimum value = –1.00.



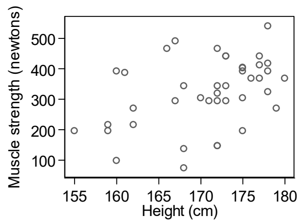
Correlation coefficient

Divide sum of products by square roots of sums of squares.

Correlation coefficient, denoted by *r*.

Maximum value = 1.00.

Minimum value = –1.00.



r = 0.42.

Positive correlation of fairly low strength

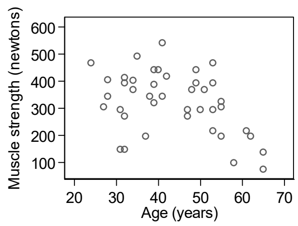
Correlation coefficient

Divide sum of products by square roots of sums of squares.

Correlation coefficient, denoted by *r*.

Maximum value = 1.00.

Minimum value = –1.00.

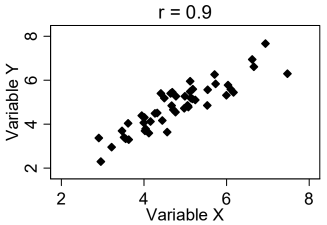


r = –0.42.

Negative correlation of fairly low strength.

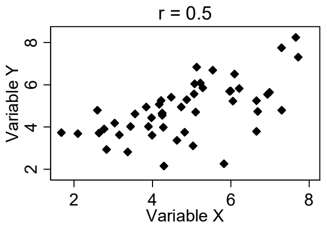
Correlation coefficient

Positive when large values of one variable are associated with large values of the other.



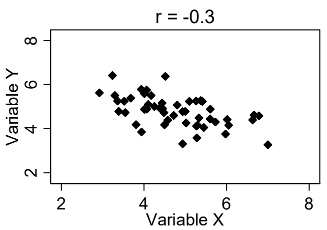
Correlation coefficient

Positive when large values of one variable are associated with large values of the other.



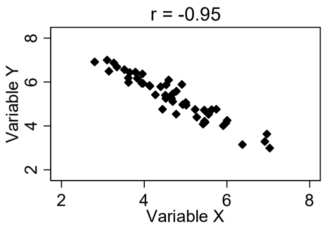
Correlation coefficient

Negative when large values of one variable are associated with small values of the other.



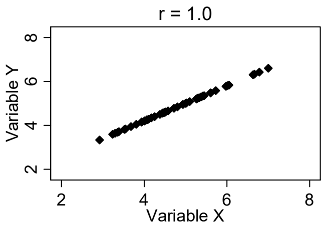
Correlation coefficient

Negative when large values of one variable are associated with small values of the other.



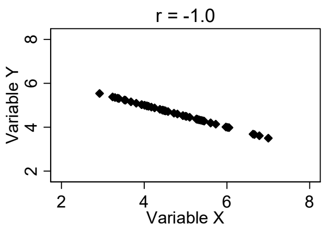
Correlation coefficient

r = +1.00 when large values of one variable are associated with large values of the other and the points lie on a straight line.



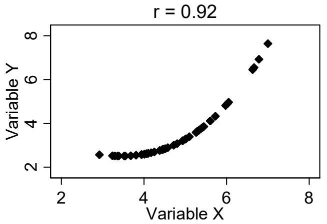
Correlation coefficient

r = –1.00 when large values of one variable are associated with small values of the other and the points lie on a straight line.



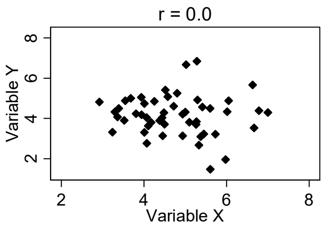
Correlation coefficient

r will not equal –1.00 or +1.00 when there is a perfect relationship unless the points lie on a straight line.



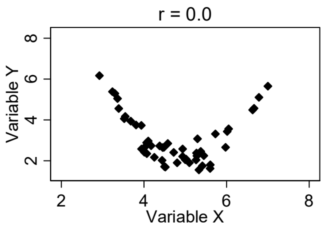
Correlation coefficient

r = 0.00 when there is no linear relationship.



Correlation coefficient

It is possible for *r* to be equal to 0.00 when there is a relationship which is not linear.



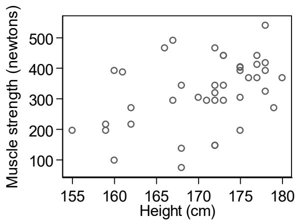
Correlation coefficient

We can test the null hypothesis that the correlation coefficient in the population is zero.

Simple t test, tabulated.

Assume: one of the variables is from a Normal distribution.

Large deviations from assumption ➔ P very unreliable.



r = 0.42, P = 0.006.

Easy to do, simple tables.

Computer programs almost always print this.

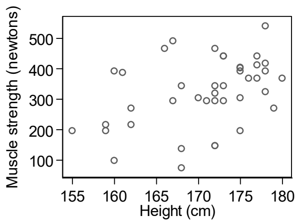
Correlation coefficient

We can find a confidence interval for the correlation coefficient in the population.

Fisher's z transformation.

Assume: both of the variables are from a Normal distribution.

Large deviations from assumption ➔ CI very unreliable.



r = 0.42, approximate 95% confidence interval: 0.13 to 0.64

Tricky, approximate.

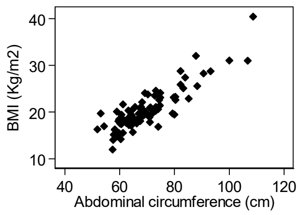
Computer programs rarely print this.

Regression analyses

- Simple linear regression
- Multiple linear regression
- Curvilinear regression
- Dichotomous predictor variables
- Regression in clinical trials
- Dichotomous outcome variables and logistic regression
- Interactions
- Factors with more than two levels
- Sample size

Simple Linear Regression

Example: Body Mass Index (BMI) and abdominal circumference in 86 women



(Data of Malcom Savage)

What is the relationship?

Regression: predict BMI from observed abdominal circumference.

Simple Linear Regression

Example: Body Mass Index (BMI) and abdominal circumference in 86 women.

What is the relationship?

Regression: predict BMI from observed abdominal circumference.

What is the mean BMI for women with any given observed abdominal circumference?

Simple Linear Regression

Example: Body Mass Index (BMI) and abdominal circumference in 86 women.

What is the relationship?

Regression: predict BMI from observed abdominal circumference.

What is the mean BMI for women with any given observed abdominal circumference?

BMI is the **outcome**, **dependent**, **y**, or **left hand side** variable.

Abdominal circumference is the **predictor**, **explanatory**, **independent**, **x**, or **right hand side** variable.

Simple Linear Regression

Example: Body Mass Index (BMI) and abdominal circumference in 86 women.

What is the relationship?

Regression: predict BMI from observed abdominal circumference.

What is the mean BMI for women with any given observed abdominal circumference (AC)?

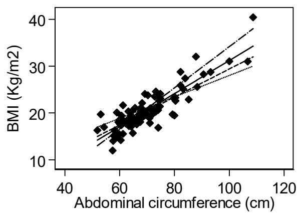
Linear relationship:

BMI = intercept + slope × AC

Equation of a straight line.

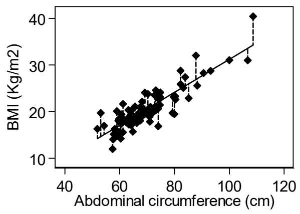
Simple Linear Regression

Which straight line should we choose?



Simple Linear Regression

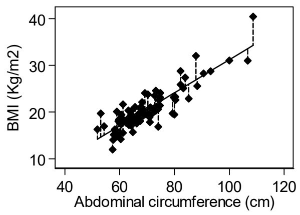
Which straight line should we choose?



Choose the line which makes the distance from the points to the line ***in the y direction*** a minimum.
Differences between the observed strength and the predicted strength.

Simple Linear Regression

Which straight line should we choose?

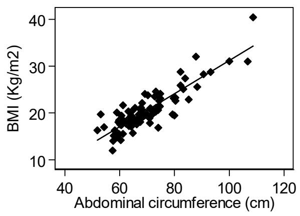


Minimise the sum of the squares of these differences.

Principle of least squares, least squares line or equation.

Simple Linear Regression

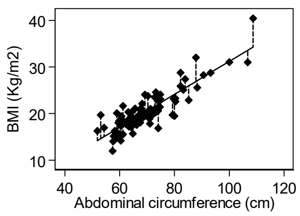
BMI = –4.15 + 0.35 × AC



We can find confidence intervals and P values for the coefficients subject to assumptions.

Simple Linear Regression

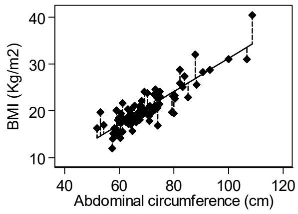
We can find confidence intervals and P values for the coefficients subject to assumptions.



Deviations from line should have a Normal distribution with uniform variance.

Simple Linear Regression

Can find confidence intervals and P values for the coefficients subject to assumptions.



Slope = 0.35 $Kg/m^2/cm$, 95% CI = 0.31 to 0.40 $Kg/m^2/cm$, P<0.001 against zero.

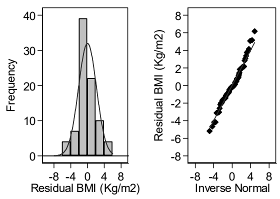
Intercept = –4.15 Kg/m^2, 95% CI = –7.11 to –1.18 Kg/m^2.

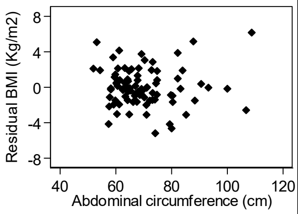
Simple Linear Regression

Assumptions: deviations from line should have a Normal distribution with uniform variance.

Calculate the deviations or residuals, observed minus predicted.

Check Normal distribution: Check uniform variance:





Dichotomous predictor variable

24 hour energy expenditure (MJ) in two groups of women

Lean			Obese	
6.13	7.53	8.09	8.79	9.69
7.05	7.58	8.11	9.19	9.97
7.48	7.90	8.40	9.21	11.51
7.48	8.08	10.15	9.68	11.85
		10.88		12.79

Can carry out linear regression.

Define variable: obese = 0 if woman lean,

obese = 1 if woman obese.

Regression equation:

energy = 5.83 + 2.23 × obese

slope: 95% CI = 1.05 to 3.42 MJ, P=0.0008.

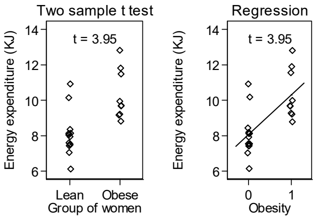
Regression and the two sample t method

Regression:

energy = 5.83 + 2.23 × obese

slope: 95% CI = 1.05 to 3.42 MJ, P=0.0008.

The two methods are identical.



Difference (obese – lean) = 10.298 – 8.066 = 2.232.

Two sample t method:

95% CI = 1.05 to 3.42 MJ,

P=0.0008.

Regression and the two sample t method

Assumptions of two sample t method

1. Energy expenditure follows a Normal distribution in each population.
2. Variances are the same in each population.

Assumptions of regression

1. Differences between observed and predicted energy expenditure follow a Normal distribution.
2. Variances of differences are the same in whatever the value of the predictor.

These are the same.