Applied Biostatistics

Correlation and regression

Martin Bland

Professor of Health Statistics

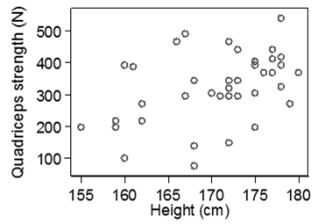
University of York

http://www-users.york.ac.uk/~mb55/msc/

Correlation

Example: Muscle strength and height in 42 alcoholics

A scatter diagram:

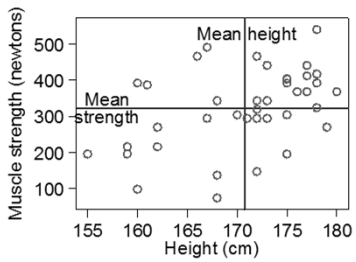


How close is the relationship?

Correlation: measures closeness to a linear relationship.

Correlation coefficient

Subtract means from observations and multiply.

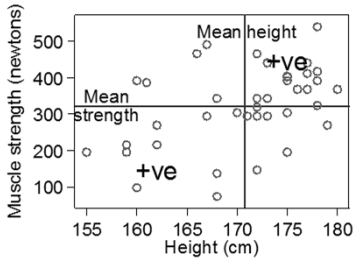


Sum of products about the means.

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

Correlation coefficient

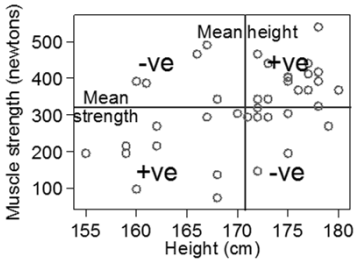
Subtract means from observations and multiply.



Products in top right and bottom left quadrants positive.

Correlation coefficient

Subtract means from observations and multiply.

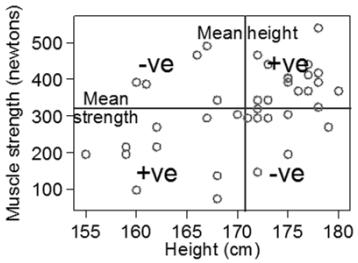


Products in top right and bottom left quadrants positive.

Products in top left and bottom right quadrants negative.

Correlation coefficient

Subtract means from observations and multiply.

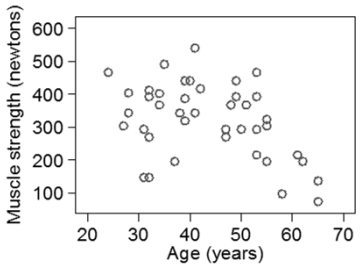


Sum of products positive.

Correlation positive.

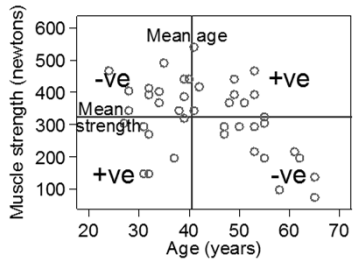
Correlation coefficient

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.

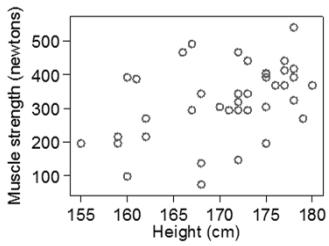
Correlation coefficient

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

Correlation coefficient, denoted by *r*.

Maximum value = 1.00.

Minimum value = –1.00.



r = 0.42.

Positive correlation of fairly low strength

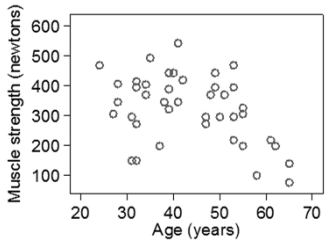
Correlation coefficient

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

Correlation coefficient, denoted by *r*.

Maximum value = 1.00.

Minimum value = –1.00.

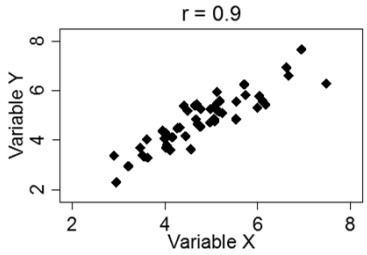


r = –0.42.

Negative correlation of fairly low strength.

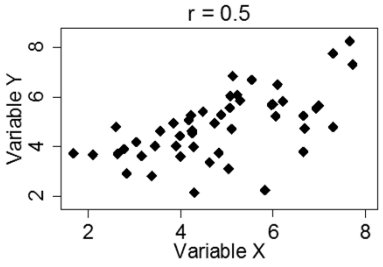
Correlation coefficient

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



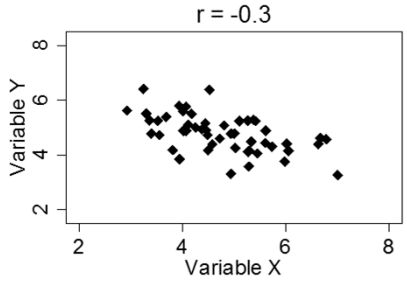
Correlation coefficient

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



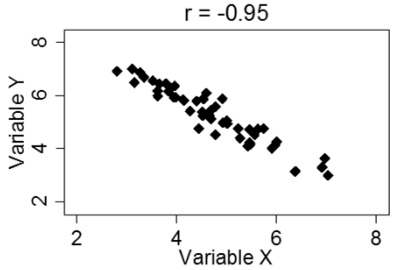
Correlation coefficient

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



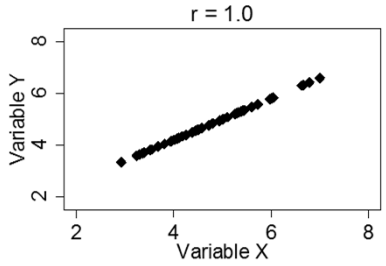
Correlation coefficient

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



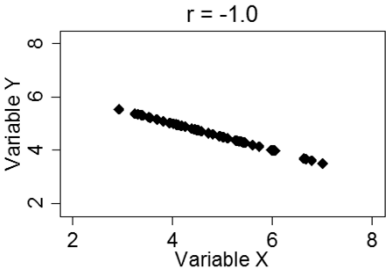
Correlation coefficient

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.



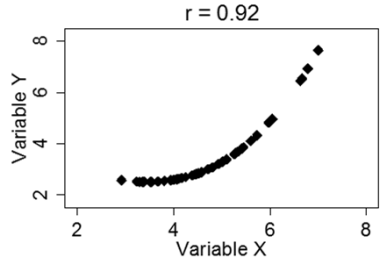
Correlation coefficient

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.



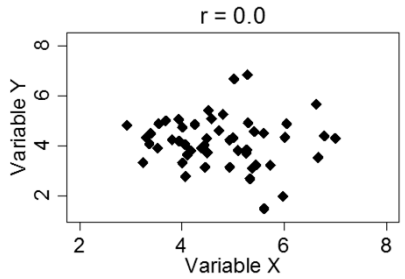
Correlation coefficient

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



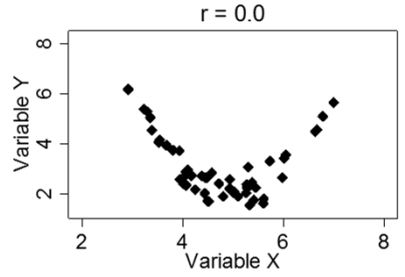
Correlation coefficient

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.



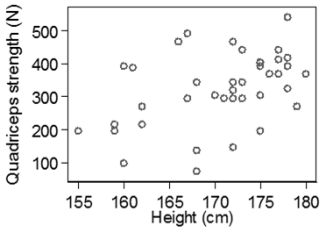
Correlation coefficient

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

Simple t test, tabulated.

Assume: independent observations, 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 programmes almost always print this.

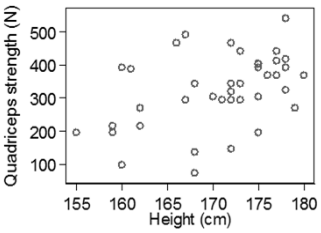
Correlation coefficient

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

Fisher's z transformation.

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

Large deviations from assumption ➔ CI very unreliable.



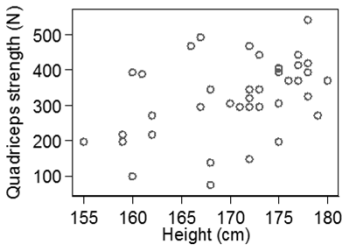
r = 0.42, approximate 95% confidence interval: 0.13 to 0.64

Tricky, approximate.

Computer programmes rarely print this.

Regression

Example: Muscle strength and height in 42 alcoholics



What is the relationship?

Regression: predict strength from observed height.

Simple Linear Regression

Example: Muscle strength and height in 42 alcoholics

What is the relationship?

Regression: predict strength from observed height.

What is the mean strength for people with any given observed height?

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

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

Linear relationship:

strength = intercept + slope × height

Equation of a straight line.

Simple Linear Regression

Strength will not be predicted exactly from height.

There will be other factors which we don't know about.

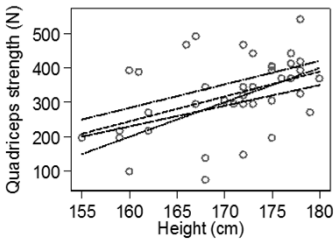
We call the other variation in the outcome variable **error**, or wandering.

Our regression model of the data is

strength = intercept + slope × height + error

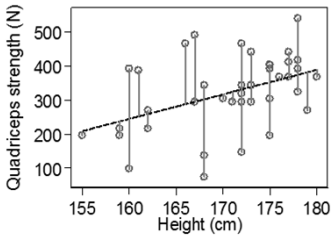
Simple Linear Regression

Which straight line should we choose?



Simple Linear Regression

Which straight line should we choose?

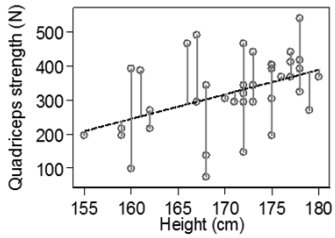


Choose the line which makes the distance from the points to the line ***in the y direction*** a minimum.

Differences between the observed strength and the predicted strength.

Simple Linear Regression

Which straight line should we choose?

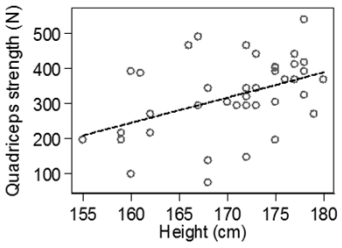


Minimise the sum of the squares of these differences.

Principle of least squares, least squares line or equation.

Simple Linear Regression

Strength = –908 + 7.20 × height



Simple Linear Regression

Strength = –908 + 7.20 × height

The intercept and slope are call **coefficients**. The slope of the line is sometimes called **the regression coefficient**.

These coefficients have units, unlike the correlation coefficient. They can take any value.

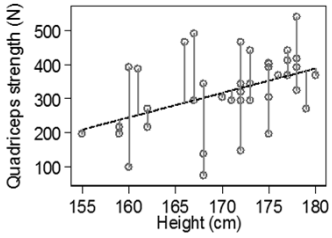
Strength in newtons =
–908 newtons + 7.20 newtons per cm × height in cm

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.

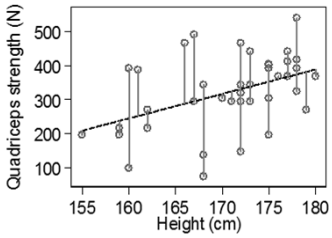


Simple Linear Regression

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

Slope = 7.20, 95% CI = 2.15 to 12.25 newtons/cm, P=0.006 against zero.

Intercept = –908, 95% CI = –45 to –1771 newtons.

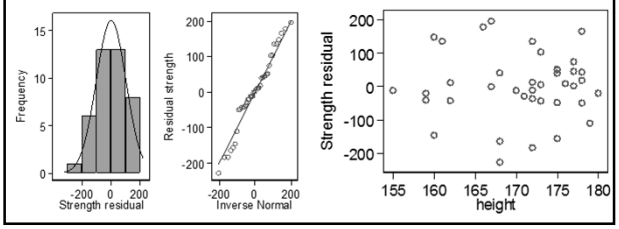


Simple Linear Regression

Assumptions: independent observations, 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:



Correlation or regression?

Correlation and regression provide two different ways to look at the relationship between two quantitative variables.

Correlation measures how closely they are related and makes no distinction between outcome and predictor.

Regression measures what the relationship is and has direction.

The regression of height on strength is not the same as the regression of strength on height, We must choose.

The tests of significance are identical for both regressions and for correlation.

Correlation and regression

Correlation and regression are closely related.

If we calculate the sum of squares about the mean for the outcome variable and the sum of squares of the deviations from the regression line, then

$$\frac{\text{SS of deviations}}{\text{SS about mean}} = 1 - r^2$$

We call r^2 the proportion of variability explained by the regression. This is often written as R^2.