PG Dip in High Intensity Psychological Interventions

Correlation and regression

Martin Bland

Professor of Health Statistics

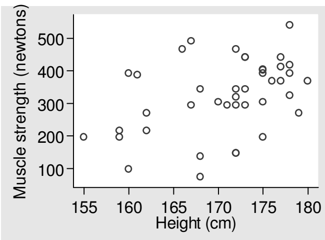
University of York

http://martinbland.co.uk/

Correlation

Example: Muscle strength and height in 42 alcoholics

A scatter diagram:

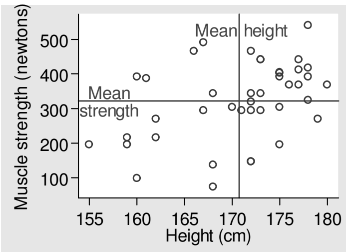


How close is the relationship?

Correlation: measures closeness to a linear relationship.

Correlation coefficient

Subtract means from observations and multiply.

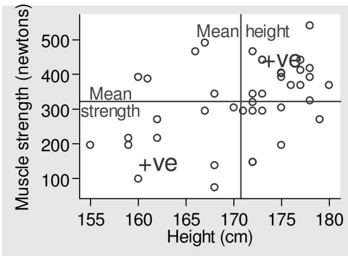


Sum of products about the means.

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

Correlation coefficient

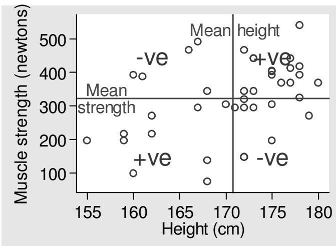
Subtract means from observations and multiply.



Products in top right and bottom left quadrants positive.

Correlation coefficient

Subtract means from observations and multiply.

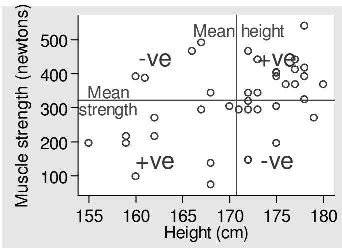


Products in top right and bottom left quadrants positive.

Products in top left and bottom right quadrants negative.

Correlation coefficient

Subtract means from observations and multiply.

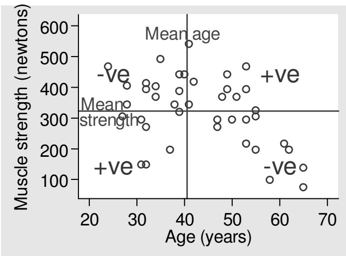


Sum of products positive.

Correlation positive.

Correlation coefficient

Example: Muscle strength and age in 42 alcoholics



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.

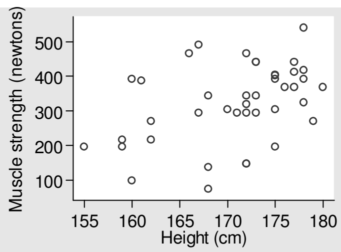
Correlation coefficient

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

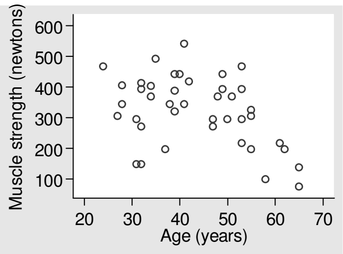
Correlation coefficient

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

Correlation coefficient, denoted by *r*.

Maximum value = 1.00.

Minimum value = –1.00.

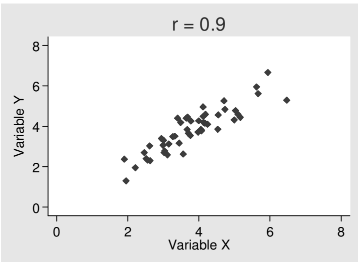


r = –0.42.

Negative correlation of fairly low strength.

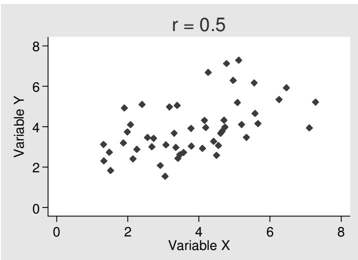
Correlation coefficient

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



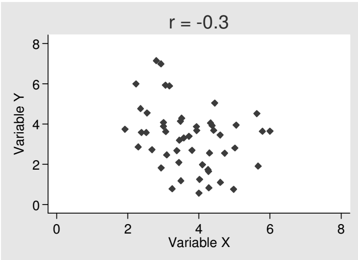
Correlation coefficient

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



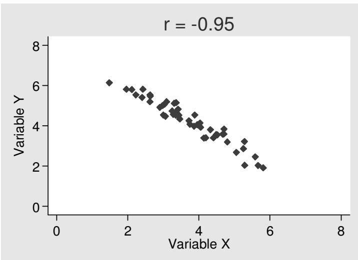
Correlation coefficient

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



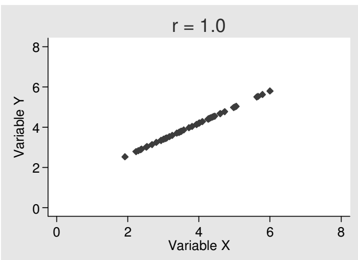
Correlation coefficient

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



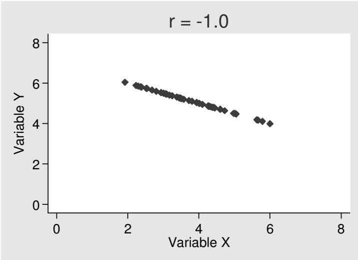
Correlation coefficient

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.



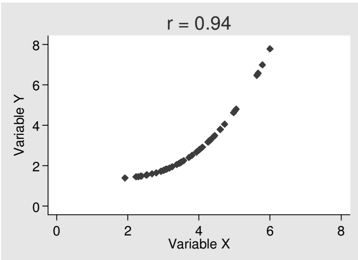
Correlation coefficient

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



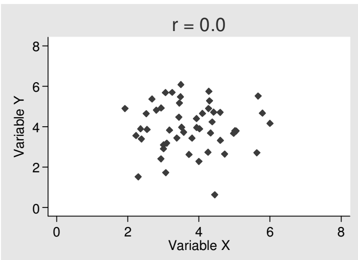
Correlation coefficient

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



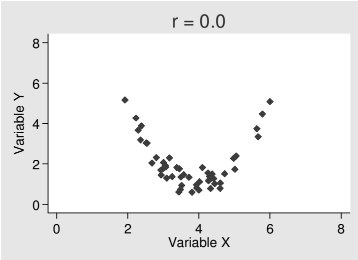
Correlation coefficient

$r = 0.00$ 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.



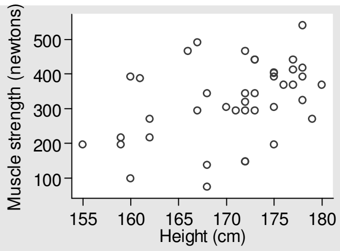
Correlation coefficient

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

Simple t test, tabulated.

Assume: 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.

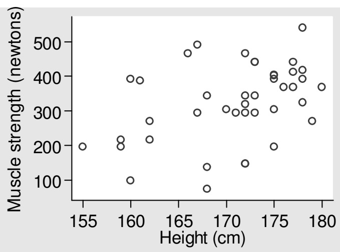
Correlation coefficient

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.



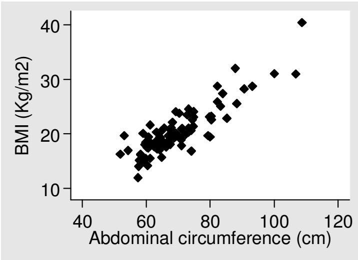
r = 0.42, approximate 95% confidence interval: 0.13 to 0.64

Tricky, approximate.

Computer programs rarely print this.

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?

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)?

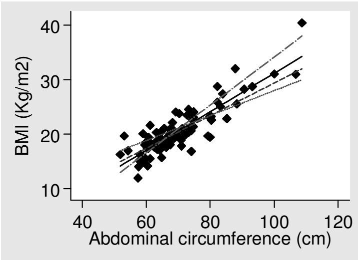
Linear relationship:

BMI = intercept + slope × AC

Equation of a straight line.

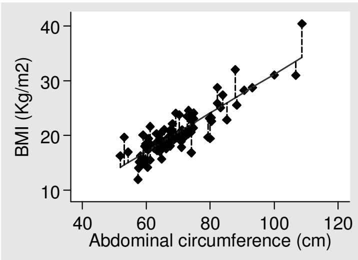
Simple Linear Regression

Which straight line should we choose?



Simple Linear Regression

Which straight line should we choose?

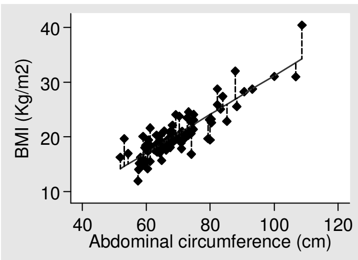


Choose the line which makes the distance from the points to the line ***in the y direction*** a minimum.

Differences between the observed strength and the predicted strength.

Simple Linear Regression

Which straight line should we choose?

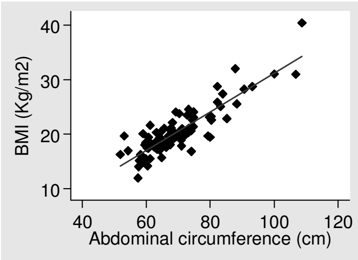


Minimise the sum of the squares of these differences.

Principle of least squares, least squares line or equation.

Simple Linear Regression

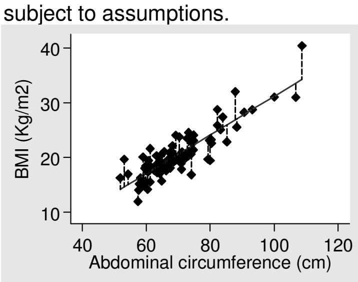
BMI = –4.15 + 0.35 × AC



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

Simple Linear Regression

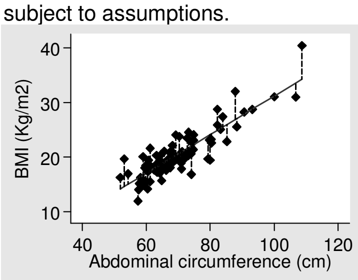
We can find confidence intervals and P values for the coefficients subject to assumptions.



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.



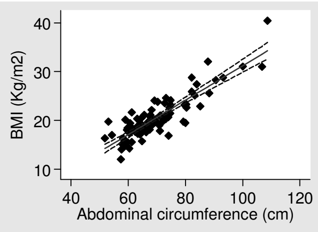
Slope = 0.35 $Kg/m^2/cm$, 95% CI = 0.31 to 0.40 $Kg/m^2/cm$, P<0.001 against zero.

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

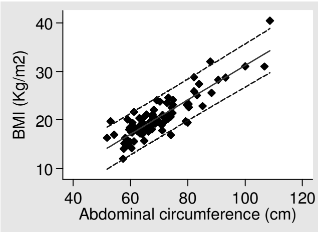
Simple Linear Regression

We can also find confidence intervals for regression estimates and predicted value for a new subject.

95% confidence intervals for regression estimates for BMI and abdominal circumference



Prediction intervals or 95% confidence intervals for prediction of BMI from abdominal circumference



Simple Linear Regression

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

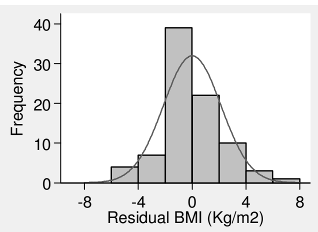
Calculate the deviations or residuals, observed minus predicted.

Simple Linear Regression

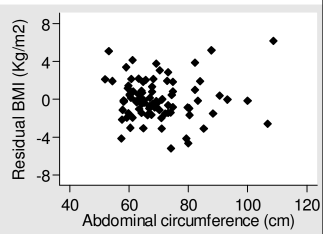
Assumptions: deviations from line should have a Normal distribution with uniform variance.

Calculate the deviations or residuals, observed minus predicted.

Check Normal distribution:

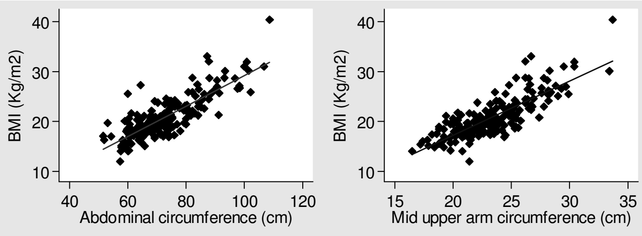


Check uniform variance:



Multiple Linear Regression

More than one predictor:



BMI = –1.35 + 0.31 × AC BMI = –4.59 + 1.09 × MUAC

BMI = –5.94 + 0.18 × AC + 0.59 × MUAC

Multiple Linear Regression

More than one predictor:

BMI = –1.35 + 0.31 × AC BMI = –4.59 + 1.09 × MUAC

BMI = –5.94 + 0.18 × AC + 0.59 × MUAC

We find the coefficients which make the sum of the squared differences between the observed BMI and that predicted by the regression a minimum.

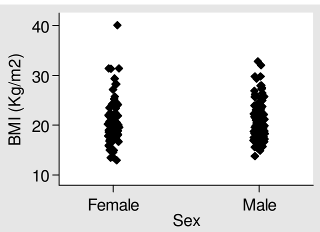
This is called **ordinary least squares** regression or **OLS** regression.

Multiple Linear Regression

Dichotomous predictor: sex.

Variable male = 0 for a female, = 1 for a male.

	BMI =	20.51 +	0.40 × male
95% CI		19.64 to 21.38	–0.75 to 1.55
			P = 0.5



Sex is not a significant predictor alone.

Multiple Linear Regression

Dichotomous predictor: sex.

Variable male = 0 for a female, = 1 for a male.

	BMI =	20.51 +	0.40 × male
95% CI		19.64 to 21.38	–0.75 to 1.55
			P = 0.5

BMI =	–6.44 +	0.18 × AC +	0.64 × MUAC	– 1.39 × male
	–8.49 to –4.39	0.14 to 0.22	0.50 to 0.78	–1.94 to –0.84
		P<0.001	P<0.001	P<0.001

Male has become a significant predictor because abdominal circumference and arm circumference have removed a lot of variability.

Mean BMI is lower for men than women ***of the same abdominal and arm circumference*** by 1.39 units.

Multiple Linear Regression

Dichotomous predictor: sex.

Variable male = 0 for a female, = 1 for a male.

BMI =	–6.44 +	0.18 × AC +	0.64 × MUAC	– 1.39 × male
	–8.49 to –4.39	0.14 to 0.22	0.50 to 0.78	–1.94 to –0.84
		P<0.001	P<0.001	P<0.001

When we have continuous and categorical predictor variables, regression is also called **analysis of covariance** or **ancova**.

The continuous variables (here AC and MUAC) are called **covariates**.

The categorical variables (here male sex) are called **factors**.

Regression in clinical trials

Used to adjust for prognostic variables and baseline measurements.

An example: specialist nurse education for acute asthma

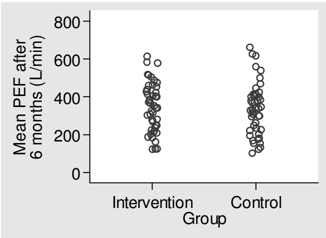
Measurements: peak expiratory flow and symptom diaries made before treatment and after 6 months.

Outcome variables: mean and SD of PEFR, mean symptom score.

Levy ML, Robb M, Allen J, Doherty C, Bland JM, Winter RJD. (2000) A randomized controlled evaluation of specialist nurse education following accident and emergency department attendance for acute asthma. *Respiratory Medicine* 94, 900-908.

Regression in clinical trials

An example: specialist nurse education for acute asthma

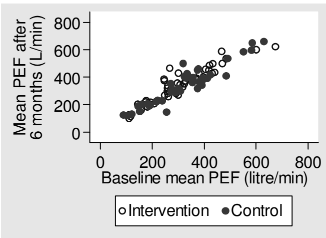


Means: 342 338 litre/min

95% CI (intervention – control) –48 to 63 litre/min, P=0.8.

Regression in clinical trials

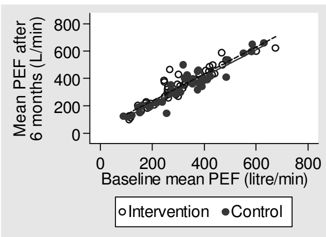
An example: specialist nurse education for acute asthma



If we control for the baseline PEF, we might get a better estimate of the treatment effect because we will remove a lot of variation between people.

Regression in clinical trials

An example: specialist nurse education for acute asthma



PEF@6m = 58.4 + 0.986 × PEF@base + 20.1 × intervene

	P<0.001	P=0.046
95% CI	0.907 to 1.064	0.4 to 39.7

Regression in clinical trials

Advantages

Reduces variability between subjects and so increase power, narrows confidence intervals.

Removes effects of chance imbalances in predicting variables.

Is adjustment cheating?

It can be if we keep adjusting by more and more variables until we have a significant difference.

We should state before we collect the data what we wish to adjust for and stick to it.

Should include any stratification or minimisation variables, centre in multi-centre trials, any baseline measurements of the outcome variable, known important predictors of prognosis.

Types of regression

Ordinary least squares regression is one types of regression

There are many other types for different kinds of outcome variable:

- Logistic regression (dichotomous)
- Cox regression (survival analysis)
- Ordered logistic regression (ordered categories)
- Multinomial regression (unordered categories)
- Poisson regression (counts)
- Negative binomial regression (counts with extra variability)