Simple linear regression
(STATS201.stat201 20: Simple linear regression)

Anna Helga Jónsdóttir
Sigrún Helga Lund

December 16, 2012

Scatter plot

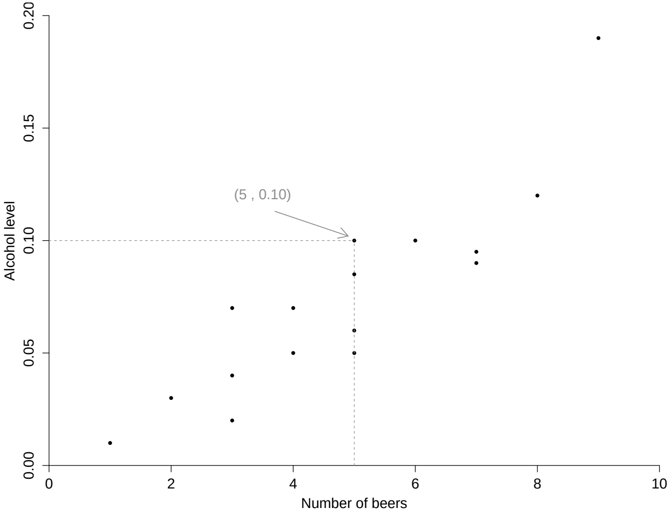
Scatter plot

Scatter plots are used to investigate the relationship between two numerical variables.
The value of one variable is on the y-axis (vertical) and the other on the x-axis (horizontal).
When one of the variable is an explanatory variable and the other one is a response variable, the response variable is always on the y-axis and the explanatory variable on the x-axis.

Response variables and explanatory variables

For every subject, the value of an **explanatory variable** will influence what value the **response variable** receives.

Scatter plot - continuous variables



The equation of a straight line

The equation of a straight line

The equation of a straight line describes a linear relationship between two variables, x and y. The equation is written

$$y = \beta_0 + \beta_1 x$$

where β_0 is the **intercept** of the line on the y-axis and β_1 is the **slope** of the line.

The equation of a straight line

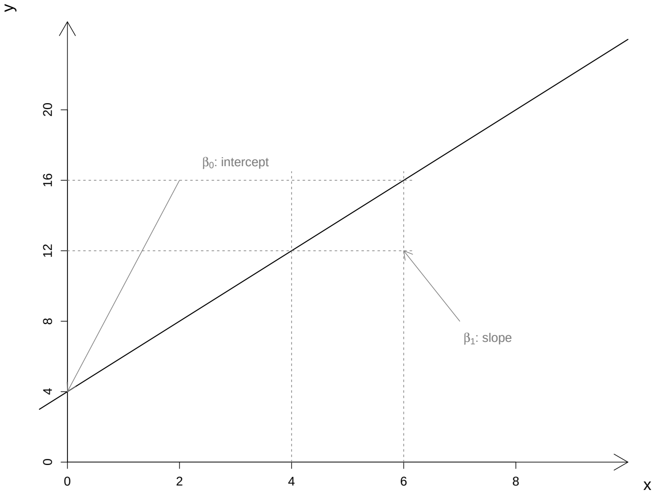


Figure: The equation of a straight line.

Linear relationship

Linear relationship

We say that the relationship between two variables is **linear** if the equation of a straight line can be used to predict which value the response variable will take based on the value of the explanatory variable.

Notice that there can be all sorts of relationship between two variables. For example, the relationship can be described with a parabola, an exponential function and so on. Those relationship are referred to as nonlinear and are not covered in this lecture.

Linear and nonlinear relationship

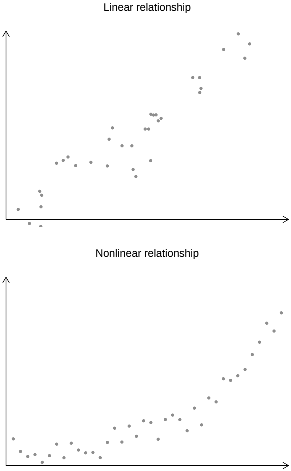


Figure: Scatter plot where the relationship between two variables is linear (above) and nonlinear (below).

Sample coefficient of correlation

Sample coefficient of correlation

Assume that we have n measurements on two variables x and y.
Denote the mean and the standard deviation of the variable x with $\bar{x}$ and s_x and the mean and the standard deviation of the y variable with $\bar{y}$ and s_y.
The sample coefficient of correlation is

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right).$$

Be careful! We only use correlation to estimate **linear** relationship!

The size and direction of a linear relationship

The direction of a linear relationship

The sign of the correlation coefficients determines the **direction** of a linear relationship. It is either positive or negative.

- If the correlation coefficient of two variables is positive, we say that their correlation is **positive**.
- If the correlation coefficient of two variables is negative, we say that their correlation is **negative**.

The size of a linear relationship

The absolute value of a correlation coefficient describes the **size** of the linear relationship between the variables.
It tells us how well we can predict the value of the response variable from the value of the explanatory variable.

The size and direction of a linear relationship

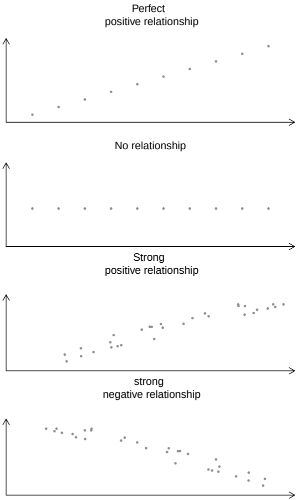


Figure: Scatter plot for various values of r.

Correlation and causation

- **Causation** is when changes in one variable **cause** changes in the other variable.
- There is often strong correlation between two variables although there is no causal relationship.
- In many cases, the variables are both influenced by the third variable which is then a **lurking variable**.
- Therefore, high correlation on its own is never enough to claim that there is a causal relationship between two variables.

The linear regression model

The linear regression model

The simple linear regression model is written

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

when β_0 and β_1 are unknown parameters and ε is a normally distributed random variable with mean 0.

The aim of the simple linear regression is first and foremost to estimate the parameters β_0 and β_1 with the measurements of the two variables, x and Y.

The method we use is called the least squares method.

The least squares method

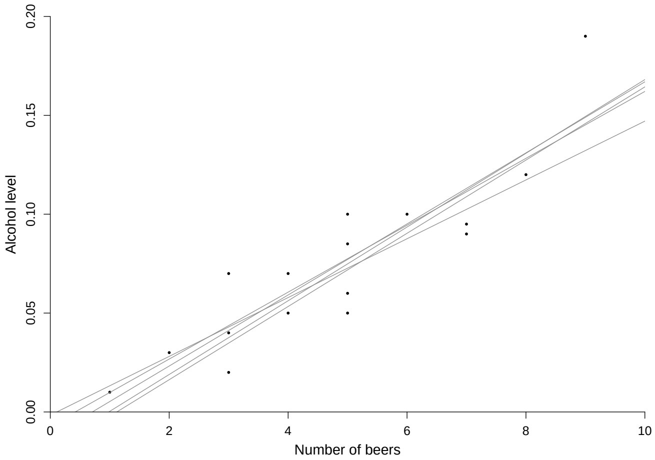


Figure: Many lines, but which one is the best?

The least squares method

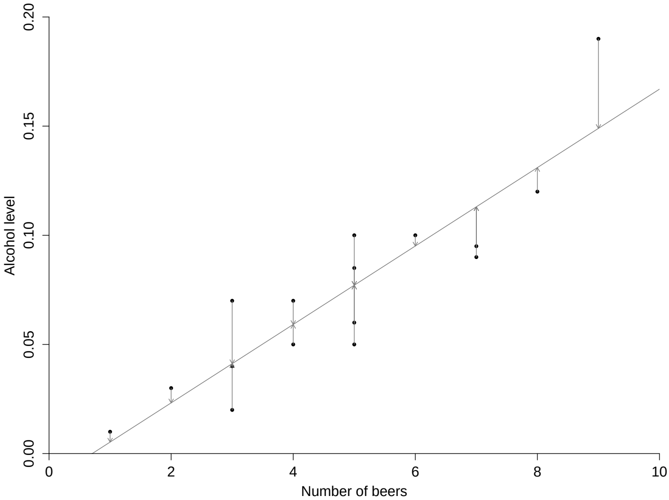


Figure: The least squares method.

The least squares regression line

The least squares regression line

Denote the mean and standard deviation of the x variable with $\bar{x}$ and s_x and the y variable with $\bar{y}$ and s_y and their correlation coefficient with r. Let b_0 denote the estimate of β_0 and b_1 denote the estimate of β_1. Then b_0 and b_1 are given with the equation

$$\hat{\beta}_1 = b_1 = r\frac{s_y}{s_x}$$

and

$$\hat{\beta}_0 = b_0 = \bar{y} - \beta_1\bar{x}.$$

We use the parameters to **predict** the value of y for a given value of x with the least squares regression line

$$\hat{y} = b_0 + b_1 x$$

Residuals

Residuals

The vertical distance from our measurements to the regression line are called the **residuals** and are denoted with e. The size of the residuals can be calculated with

$$e_i = y_i - \hat{y}_i$$

Points above the regression line have a positive residue but points below it have a negative.

Residual plot

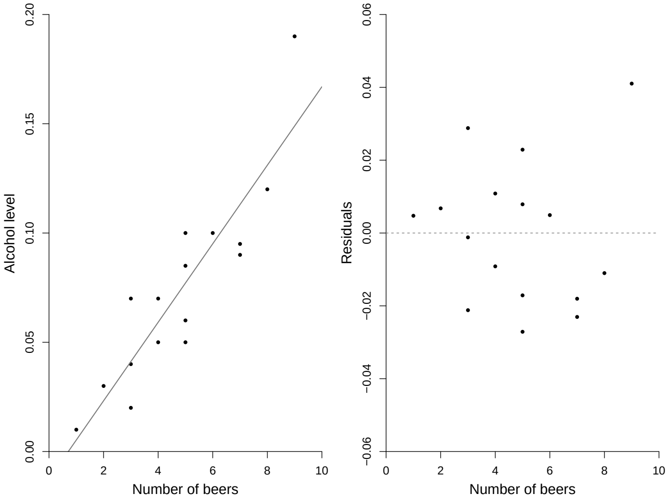


Figure: Scatter plot of the data and a residual plot.

Interpolation

Interpolation

If the regression model is used to predict a value of Y for some value of x which is similar to the x-values that were used to estimate the model is referred to as **interpolating**.

Extrapolation

Extrapolation

Extrapolating is using the regression model to predict a value of Y for some value of x which is far from the x-values that were used to estimate the model.
It can be very questionable to extrapolate!

Coefficient of determination

r^2 in linear regression

If the coefficient of correlation is raised to the second power, r^2, we calculate the coefficient of determination.
r^2 is the proportional variability in Y that can be explained with the variability in x.

Outliers and influential measurements

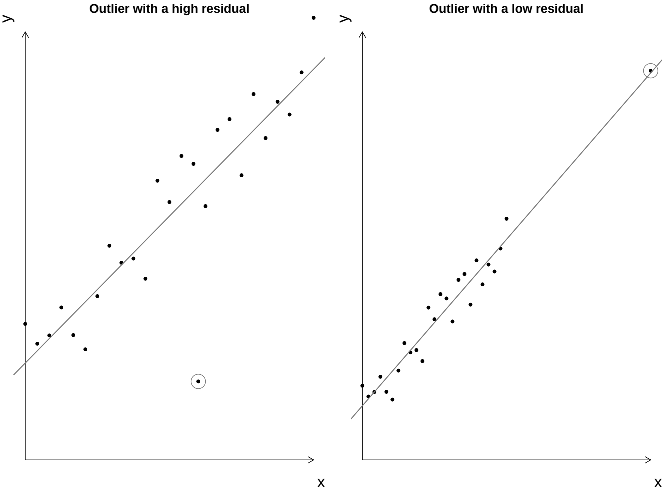


Figure: Outliers and their residuals.

Outliers and influential measurements

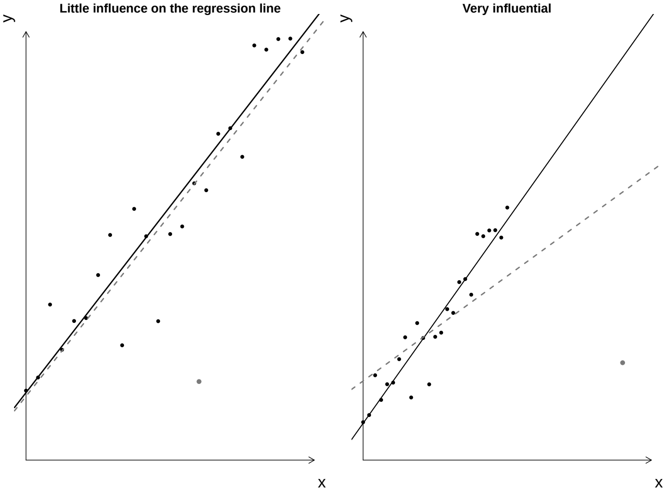


Figure: Influential measurements?

Treatment of outliers and influential measurements

- Outliers and influential measurements shall always be viewed carefully.
- If a mistake has been made, the measurement shall be eliminated.
- If it cannot be shown that a mistake has been made it is often good to show estimates with and without these measurements.
- In some cases it is more appropriate to use the estimates without the outliers/influential measurements.
- In these cases, it shall be pointed out that the model does not it data outside the range of the measurements used for estimating the model.

The linear regression model

If we have n paired measurements $(x_1, y_1), \dots, (x_n, y_n)$, the regression model can be written as

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i.$$

- β_0 is the true intercept (population intercept) that we do not know the value of.
- β_1 is the true slope (population slope)
- ϵ_i are the errors.

β_0 and β_1 are therefore statistics, that we both want to estimate and make inference on.

We do that by applying the least squares method to our data.

The random variable ε

ε describes the uncertainty in our measurements of Y.

We assume that ε_i are independent and identically distributed random variables that follow a normal distribution with mean 0 and variance σ^2.

Estimating σ^2 in simple linear regression

The estimate of σ^2 in simple linear regression is denoted with s_e^2 and calculated with

$$s_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}.$$

This is the same equation as for the "normal" standard deviation, but now we divide with $n-2$ but not $n-1$.

Confidence interval for β_0

Confidence interval for β_0

The lower bound of a $1-\alpha$ confidence interval for β_0 is:

$$b_0 - t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{s_x^2 \cdot (n-1)}}$$

The upper bound of $1-\alpha$ confidence interval is:

$$b_0 + t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{\frac{1}{n} + \frac{(\bar{x})^2}{s_x^2 \cdot (n-1)}}$$

where b_0 is calculated the same way as usual, n is the number of paired measurements, $\bar{x}$ is the mean of the explanatory variable, s_x is the standard deviation of the explanatory variable and $t_{1-\alpha/2,(n-2)}$ is in the table for the t-distribution.

Confidence interval for β_1

Confidence interval for β_1

The lower bound of1 $-\alpha$ confidence interval for β_1 is:

$$b_1 - t_{1-\alpha/2,(n-2)} \cdot s_e \frac{1}{\sqrt{s_x^2 \cdot (n-1)}}$$

The upper bound of $1-\alpha$ confidence interval is:

$$b_1 + t_{1-\alpha/2,(n-2)} \cdot s_e \frac{1}{\sqrt{s_x^2 \cdot (n-1)}}$$

where b_1 is calculated the same way as usual, n is the number of paired measurements, s_x is the standard deviation of the explanatory variable and $t_{1-\alpha/2,(n-2)}$ is found in the t-distribution table.

Prediction interval

Prediction interval

The lower bound of1 $-\alpha$ prediction interval for Y is:

$$(b_0 + b_1 x_0) - t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_x^2(n-1)}}$$

The upper bound of $1-\alpha$ prediction interval is:

$$(b_0 + b_1 x_0) + t_{1-\alpha/2,(n-2)} \cdot s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{s_x^2(n-1)}}$$

where b_0 and b_1 are calculated the same way as usual, n is the number of paired measurements, s_x is the standard deviation of the explanatory variable and $t_{1-\alpha/2,(n-2)}$ is found in the t-distribution table.

Hypothesis test for the correlation coefficient

Hypothesis test forρ

The null hypothesis is:

$$H_0 : \rho = 0$$

The test statistic is:

$$T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

If the null hypothesis is true the test statistic follows the t distribution with n-2 degrees of freedom or $T \sim t(n-2)$.

Alternative hypothesis	Reject H_0 if:
$\mathrm{H}_1 : \rho < 0$	$T < -t_{1-\alpha}$
$\mathrm{H}_1 : \rho > 0$	$T > t_{1-\alpha}$
$\mathrm{H}_1 : \rho \neq 0$	$T < -t_{1-\alpha/2}$ or $T > t_{\alpha/2}$