Simple linear regression (SLR)

Anna Helga Jonsdottir
Gunnar Stefansson
Sigrun Helga Lund

University of Iceland

Introduction

The following gives a review of:

- Scatter plots
- Correlation
- Simple linear regression - SLR
- Inference in SLR

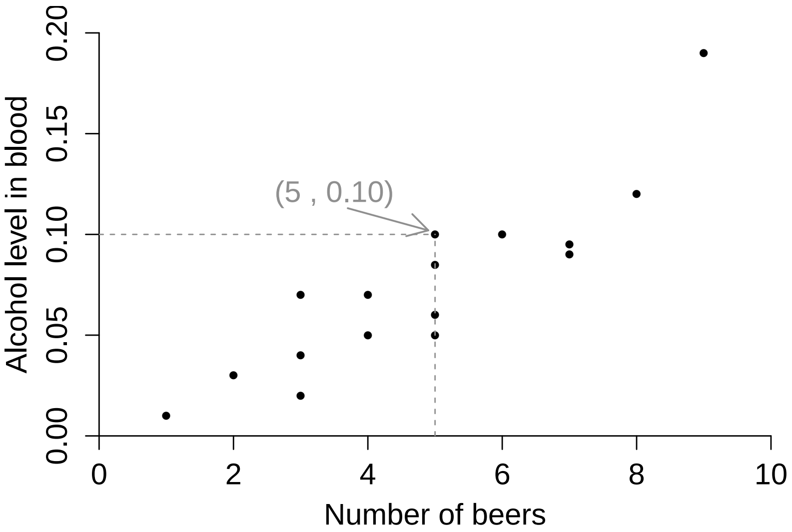
Scatter plot

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.


(5 , 0.10)
Alcohol level in blood
0.00
0.05
0.10
0.15
0.20
0
2
4
6
8
10
Number of beers

The equation of a straight line

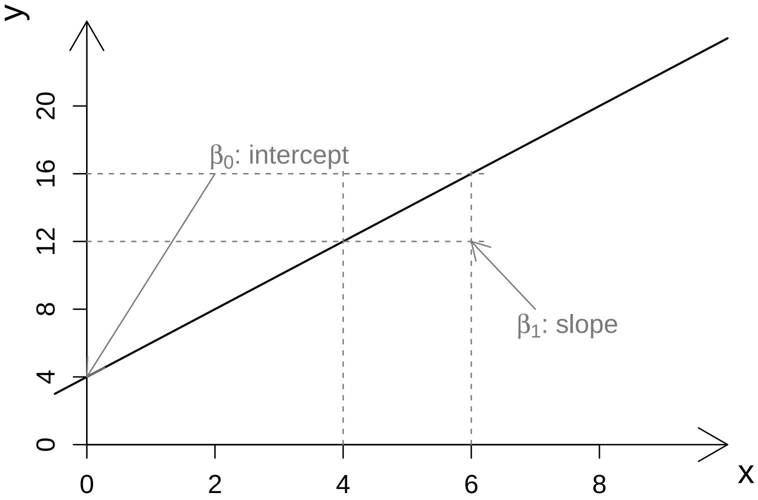
The equation of a straight line

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 straight line



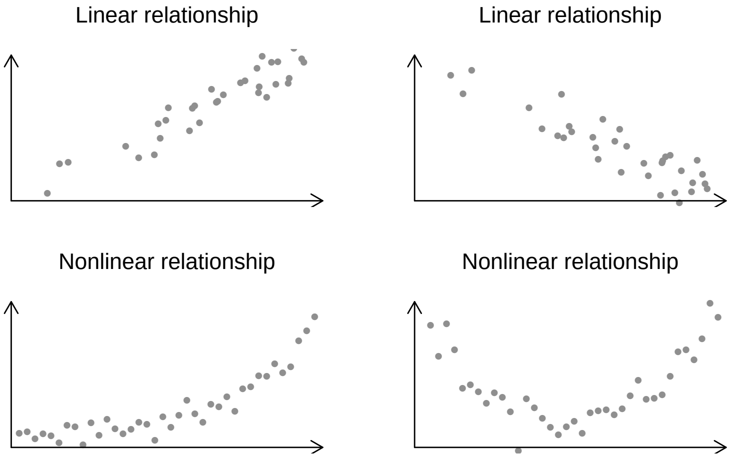
Linear relationship

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.

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.

Linear/nonlinear relationship



Sample correlation coefficient

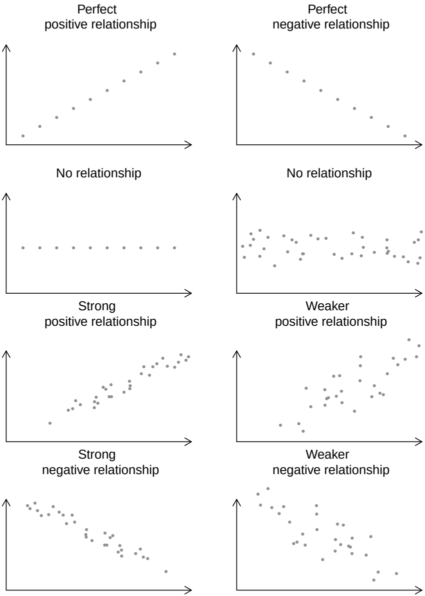
Sample coefficient of correlation

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

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

Warning: The correlation only estimates the strength of a **linear** relationship!


Perfect
positive relationship
Perfect
negative relationship
No relationship
No relationship
Strong
positive relationship
Weaker
positive relationship
Strong
negative relationship
Weaker
negative relationship

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!

Informal regression

Input: Have data as (x, y)-pairs

Suppose a scatterplot indicates a linear relationship

Loosely: Want to "fit a line" through the data

Next: Evaluate the fit

Formal regression

Consider fixed numbers, x_i

Random variables: $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$

or: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$

where $\epsilon_i \sim N(0, \sigma^2)$ is a random error term, independent and identically distributed (i.i.d.)

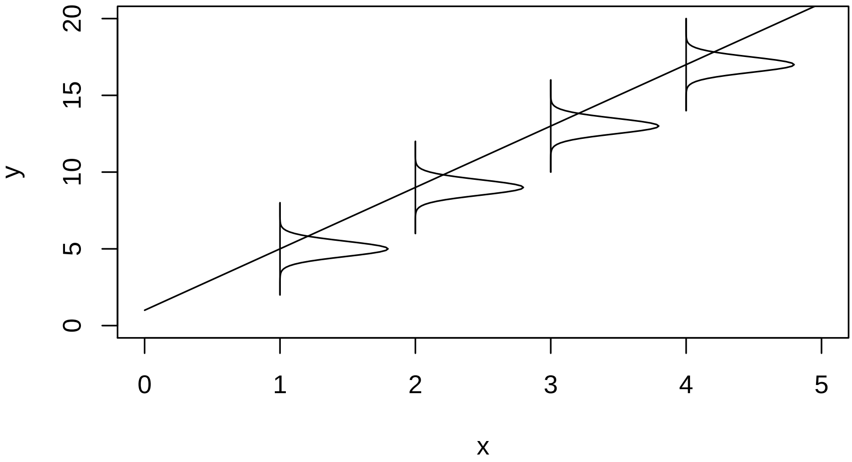
We collect some data on (y_i, x_i) and use the data to estimate β_0 and β_1. We can then predict Y using

$$\hat{y}_i = b_0 + b_1 x_i$$

Then

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

is the ith residual - the difference between the ith observed response and the prediction.



y
0
5
10
15
20
0
1
2
3
4
5
x

The linear regression model

The linear regression model

The simple linear regression model is written

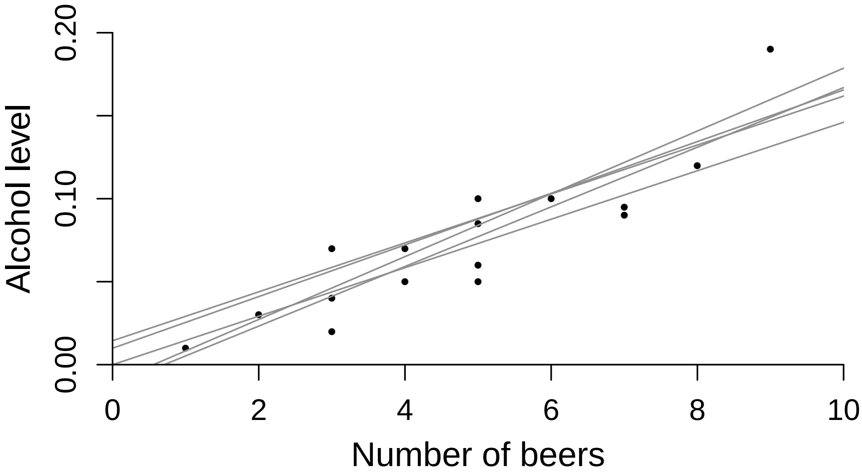
$$Y = \beta_0 + \beta_1 x + \varepsilon$$

where β_0 and β_1 are unknown parameters and ε is a normally distributed random variable with mean 0 and variance σ^2.

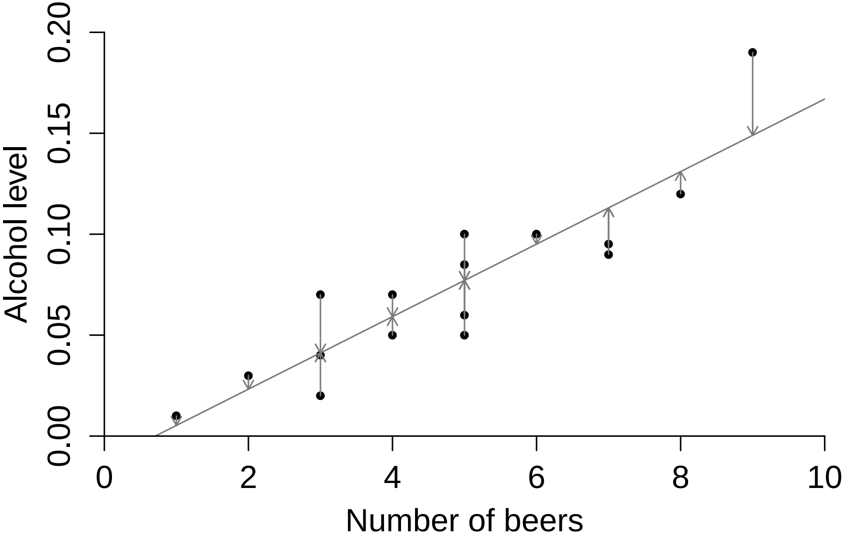
The aim of the simple linear regression is first and foremost to estimate the parameters β_0 and β_1 with the measurements on the two variables, x and Y.

The most common estimation method is through **least squares**.

Which line?



The least squares method



The least squares method

Least squares estimation technique minimizes:

$$S = \sum (y_i - (b_0 + b_1 x_i))^2$$

Maximum likelihood estimation assumes a probability distribution for the data and maximizes the corresponding likelihood function.

In the case of normal distributions the two methods results in the same estimates - we will use least squares.

The least squares regression line

Let 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

$$b_1 = \frac{\sum(x - \bar{x})(y - \bar{y})}{\sum(x - \bar{x})^2} = r\frac{s_y}{s_x}$$

and

$$b_0 = \bar{y} - b_1\bar{x}.$$

These are the least squares estimates of the coefficient of a regression line through the data points (x, y).

Prediction

We often want to use our regression model to predict the outcome of our response variable for some value(s) of the explanatory variable.

Prediction

We can predict the value of Y for some value of x using

$$\hat{y}_h = b_0 + b_1 \cdot x_h$$

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!

Estimating dispersion

A point estimate of σ^2, the variance of the y-measurements, is obtained with

$$s^2 = \frac{\sum_i \left(y_i - (b_0 + b_1 x_i)\right)^2}{n-2}$$

The predicted value of y at a given x is often denoted by $\hat{y} = b_0 + b_1 x$ and therefore

$$s^2 = \frac{\sum_i (y_i - \hat{y}_i)^2}{n-2}$$

Commonly $\hat{\sigma}^2$ is used in place of s^2.

Estimating dispersion

We will commonly use the notation

$$SS_E = \sum_i \left(y_i - (b_0 + b_1 x_i)\right)^2$$

and

$$MSE = SS_E/(n-2)$$

so $s^2 = MSE$.

Correlation and explained variation

Recall the the correlation coeffficient r is always between -1 and 1.

Write $SS_E = \sum(y - \hat{y})^2$ (sum of squared errors, i.e. error after regression), and $SS_{TOT} = \sum(y - \bar{y})^2$ (total sum of squares, i.e. before regression)

The explained variation

The explained variation, often called the coefficient of determination, is calculated with

$$R^2 = 1 - \frac{SS_E}{SS_{TOT}}$$

Note:

$$R^2 = 1 - \frac{\sum(y - \hat{y})^2}{\sum(y - \bar{y})^2} = \ldots = r^2$$

SLR in R

It is easy to perform linear regression in R using the `lm()` function. For simple linear regression the syntax is

```
fit <- lm(x ~ y, data=nameofdataset)
```

The results can then be looked at using the `summary()` function

```
summary(fit)
```

Inference in the linear regression model

Recall that if we have n paired measurements $(x_1, y_1), \ldots, (x_n, y_n)$, 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 where $\epsilon \sim N(0, \sigma^2)$ and

$$\hat{\sigma}^2 = \frac{\sum (y - \hat{y})^2}{n - 2}$$

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

Estimating slope and intercept accuracy

The standard error of the slope is:

$$\hat{\sigma}^2_{b_1} = \frac{\hat{\sigma}^2}{\sum(x - \bar{x})^2}$$

and the standard error of the intercept is:

$$\hat{\sigma}^2_{b_0} = \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right)\hat{\sigma}^2$$

where

$$\hat{\sigma}^2 = \frac{\sum(y - \hat{y})^2}{n - 2}$$

Elements of inference in simple linear regression

- Basic inference: Test hypotheses and generate confidence intervals for slope and intercept.
- Results on the estimators can be used to make inference on the true slope and intercept.
- The first question raised is whether there is any relationship between the x and y measurements, i.e. whether the slope is zero. This can be phrased as a general hypothesis test for the slope.
- Although hypothesis tests are important, they give no information if the hypothesis can not be rejected and hence confidence intervals tend to be more informative in general.
- Both hypothesis tests and confidence intervals can be derived for the intercept as well as the slope, although inference for the intercept tends not to be as commonly used.

Testing hypotheses concerning the slope

Hypothesis test for β_1

The null hypothesis is:

$$H_0 : \beta_1 = \beta_{1_0}$$

The test statistic is:

$$t = \frac{b_1 - \beta_{1_0}}{\hat{\sigma}_{\hat{\beta}_1}}$$

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:
$\mathrm{H}_1 : \beta_1 < \beta_{1_0}$	$t < -t_{1-\alpha}$
$\mathrm{H}_1 : \beta_1 > \beta_{1_0}$	$t > t_{1-\alpha}$
$\mathrm{H}_1 : \beta_1 \neq \beta_{1_0}$	$t < -t_{1-\alpha/2}$ or $t > t_{\alpha/2}$

Confidence interval for β_1

Confidence interval for β_1

The lower bound of$1-\alpha$ confidence interval for β_1 is:

$$b_1 - t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}$$

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

$$b_1 + t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_1}$$

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

Confidence interval for β_0

Confidence interval for β_0

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

$$b_0 - t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0}$$

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

$$b_0 + t_{1-\alpha/2,(n-2)} \cdot \hat{\sigma}_{\hat{\beta}_0}$$

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

Confidence interval for a point on the regression line

Confidence interval for a point on the regression line

The lower bound of $1 - \alpha$ confidence interval for $\hat{Y}_h$ is:

$$(b_0 + b_1 x_h) - t_{1-\alpha/2,(n-2)} \cdot s_{y_h}$$

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

$$(b_0 + b_1 x_h) + t_{1-\alpha/2,(n-2)} \cdot s_{y_h}$$

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

$$s_{y_h} = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_j (x_j - \bar{x})^2} \right)}$$

Predicting a new observation

Notice, that as a prediction for a future point, includes two sources of variation or error, first due to the measurement errors in the original data through variation in the parameter estimates and secondly through the future measurement errors at this point.

Prediction interval for a new observation

Prediction interval for a new observation

The lower bound of $1-\alpha$ prediction interval for $\hat{Y}_h$ is:

$$(b_0 + b_1 x_h) - t_{1-\alpha/2,(n-2)} \cdot s_{pred}$$

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

$$(b_0 + b_1 x_h) + t_{1-\alpha/2,(n-2)} \cdot s_{pred}$$

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

$$s_{pred} = \sqrt{\hat{\sigma}^2 \left(1 + \frac{1}{n} + \frac{(x_h - \bar{x})^2}{\sum_j (x_j - \bar{x})^2}\right)}.$$