Simple linear regression

(STATS544.1: Applied simple linear regression)

Anna Helga Jonsdottir and Gunnar Stefansson

September 23, 2012

Introduction

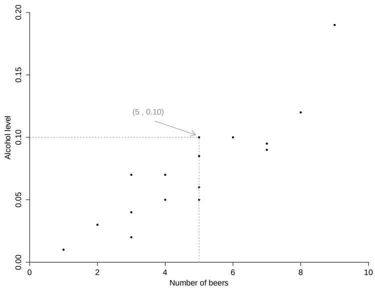
The following section give a review of:

- Scatter plots
- Correlation
- Simple linear regression - 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.



Response variables and explanatory variables

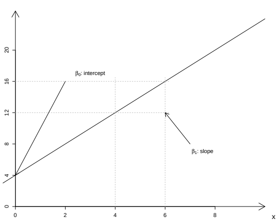
The straight line

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

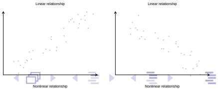
$$y = \beta_0 + \beta_1 x$$

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



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.



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!

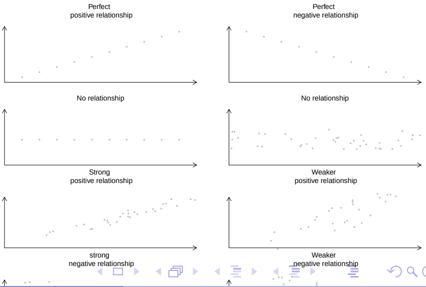
The magnitude and direction of a linear relationship

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 magnitude of a linear



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

Supose a scatterplot indicates a linear relationship

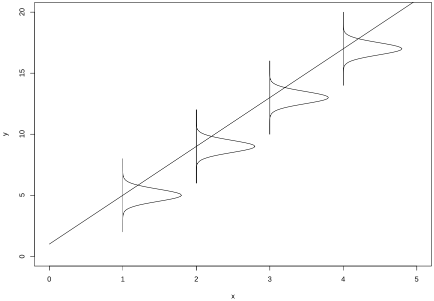
Loosely: Want to "fit a line" through the data

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: $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
$\epsilon_i \sim n(0, \sigma^2)$ independent and identically distributed (i.i.d.)
The data:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$



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.

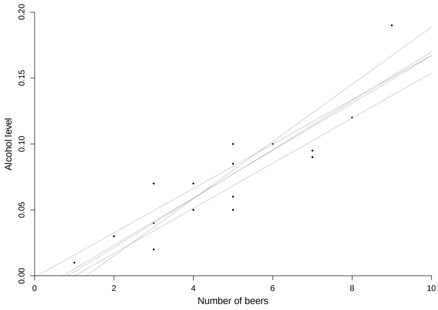


Figure: Many lines, but which one is the best?

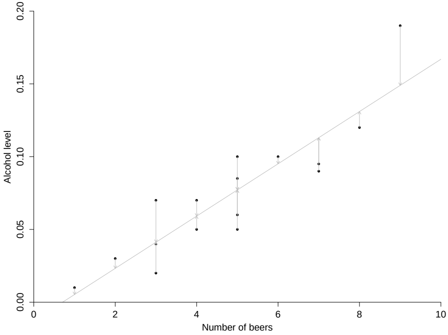
The least squares method

Least squares estimation technique minimizes:

$$S = \sum (y_i - (\beta_0 + \beta_1 x_i))^2$$

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

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

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

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).

Remember: It is assumed that the only errors are in the y-measurements.

Example - using the first expression for b_1: Suppose we have a few measurements, (x, y), to be used in a regression analysis.

	x	y	$x-\bar{x}$	$(x-\bar{x})^2$	$(y-\bar{y})$	$(y-\bar{y})^2$	$(x-\bar{x})(y-\bar{y})$	$\hat{y}$	y -
	1	1.0	-2	4	-5	25	10	2.6	

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

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

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!

On expected values and variances

Expected value:

$$E[Y] = \int y f(y) dy$$

Variance:

$$V[Y] = E\left[(Y - \mu)^2\right].$$

In the regression model:

$$V[Y] = E\left[(Y - (\beta_0 + \beta_1 x))^2\right].$$

Estimating dispersion

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

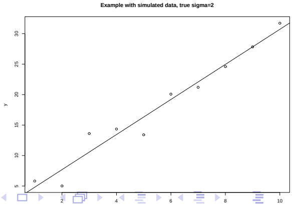
$$s^2 = \frac{\sum_i (y_i - (b_0 + b_1 x_i))^2}{n-2}$$

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, but that is not stricly correct.
Example R summary which gives the variance estimate (simulated data):

```
> x<-1:10
> alpha<-2
> beta<-3
> sigma<-2
> y<-alpha+beta*x+rnorm(10)*sigma
> plot(x,y)
> fm<-lm(y~x)
```



Correlation and explained variation

Recall the the correlation coeffficient r is always between -1 and 1.
Write $SSE = \sum(y-\hat{y})^2$ (sum of squared errors, i.e. error after regression), and $SSTOT = \sum(y-\bar{y})^2$ (total sum of squares, i.e. before regression)

The explained variation

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

$$R^2 = 1 - \frac{SSE}{SSTOT}$$

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)
```

Typical complete interactive R session:

```
> beers<-c(5,2,9,7,3,3,4,5,8,3,5,5,6,7,1,4)
> alcohol<-c(0.1,0.03,0.19,0.095,0.07,0.02,0.07,0.085,0.12,0.04,0.06,0.05,0.1,0.09
>
> fit<-lm(alcohol~beers)
> summary(fit)

Call:
lm(formula = alcohol ~ beers)

Residuals:
```

Interpreting package output

Mathematical model:

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

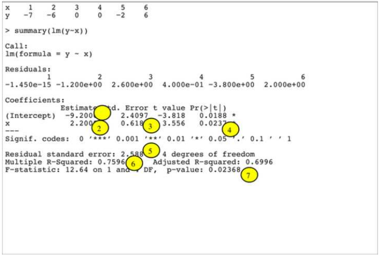
R definition:

$$y \sim x$$

`lm(y~x)`

Storing the output

`fit<-lm(y~x).`



```
x     1    2    3    4    5    6
y    -7   -6    0    0   -2    6

> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
          1          2          3          4          5          6
-1.450e-15 -1.200e+00  2.600e+00  4.000e-01 -3.800e+00  2.000e+00

Coefficients:
            Estimate   d. Error t value Pr(>|t|)
(Intercept)  -9.200      2.4097  -3.818   0.0188 *
x             2.200      0.618    3.556   0.0237
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.588    4 degrees of freedom
Multiple R-Squared: 0.7596    Adjusted R-squared: 0.6996
F-statistic: 12.64 on 1 and   DF,  p-value: 0.02368
```


Figure: Example output from a simple linear model fit of the form y=a+bx. Items (1)-(2) are the estimates of a and b respectively. The estimate of the standard error of b is given by (3). The P-value for testing whether the true (underlying) value of b is zero is in (4). Items (5)-(7) give the MSE, R-squared and P-value for the entire model, respectively.

A sequence:

```
fm<-lm(y~x)          # Fitting the model
summary(fm)          # Traditional summary
```