Rcourse: **Linear model**

Sonja Grath, Noémie Becker & Dirk Metzler

Winter semester 2014-15

Contents

Intruitive linear regression

What is linear regression?

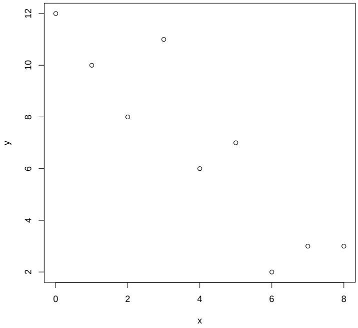
Intruitive linear regression

What is linear regression?
It is the straight line that best approximates a set of points:
y=a+b*x
a is called the intercept and b the slope.

Linear regression by eye

I give you the following points:

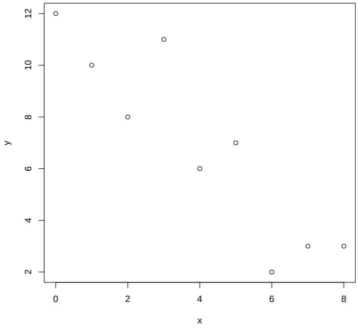
```
x <- 0:8 ; y <- c(12,10,8,11,6,7,2,3,3) ; plot(x,y)
```



Linear regression by eye

I give you the following points:

```
x <- 0:8 ; y <- c(12,10,8,11,6,7,2,3,3) ; plot(x,y)
```

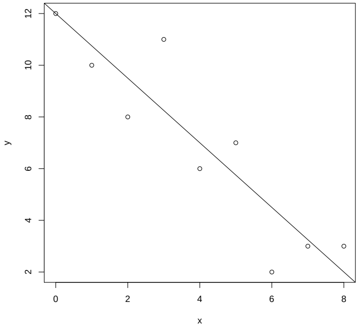


By eye we would say a=12 and b=(12-2)/8=1.25

Linear regression by eye

I give you the following points:

```
x <- 0:8 ; y <- c(12,10,8,11,6,7,2,3,3) ; plot(x,y)
```



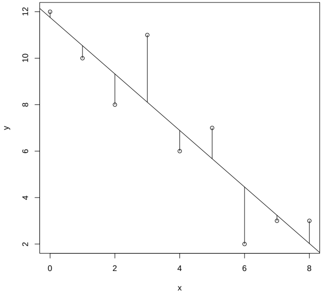
By eye we would say a=12 and b=(12-2)/8=1.25

Best fit in R

y is modelled as a function of x. In R this job is done by the function `lm()`. Lets try on the R console.

Best fit in R

y is modelled as a function of x. In R this job is done by the function `lm()`. Lets try on the R console.



The linear model does not explain all of the variation. The error is called ”residual”.
The purpose of linear regression is to minimize this error. But do you remember how we do this?

Statistics

We define the linear regression

$$y = \hat{a} + \hat{b} \cdot x$$

by minimizing the sum of the square of the residuals:

$$(\hat{a}, \hat{b}) = \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2$$

This assumes that a, b exist, so that for all (x_i, y_i)

$$y_i = a + b \cdot x_i + \varepsilon_i,$$

where all ε_i are independant and follow the normal distribution with varaince σ^2.

Statistics

We estimate a and b, by calculating

$$(\hat{a}, \hat{b}) := \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2$$

Statistics

We estimate a and b, by calculating

$$(\hat{a}, \hat{b}) := \arg\min_{(a,b)} \sum_i (y_i - (a + b \cdot x_i))^2$$

We can calculate $\hat{a}$ und $\hat{b}$ by

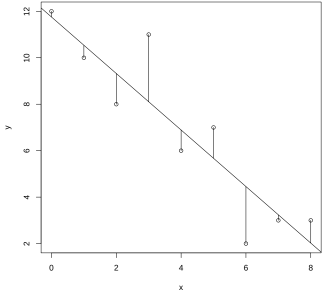
$$\hat{b} = \frac{\sum_i (y_i - \bar{y}) \cdot (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} = \frac{\sum_i y_i \cdot (x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

and

$$\hat{a} = \bar{y} - \hat{b} \cdot \bar{x}.$$

Back to our example

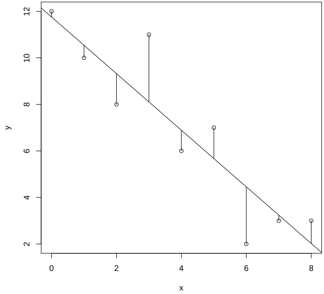
The commands used to produce this graph are the following:




```
regr.obj <- lm(y x)
fitted <- predict(regr.obj)
```

Back to our example

The commands used to produce this graph are the following:

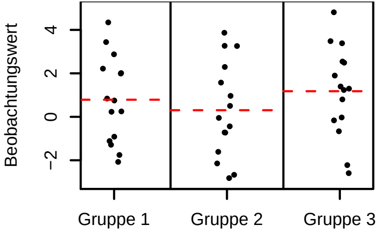



```
regr.obj <- lm(y x)
fitted <- predict(regr.obj)
plot(x,y); abline(regr.obj)
for(i in 1:9)
{
lines(c(x[i],x[i]),c(y[i],fitted[i])
}
```

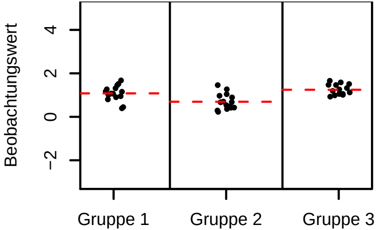
Contents

Reminder: ANOVA

I am sure you all remember from statistic courses:
We observe different mean values for different groups.



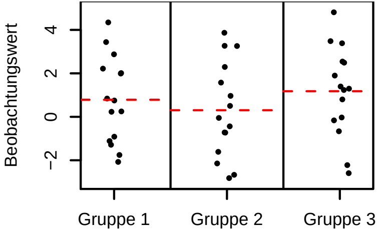
High variability within groups



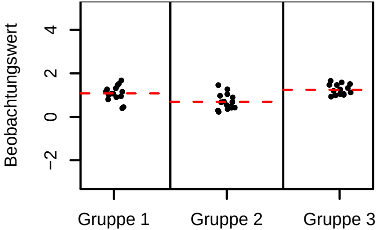
Low variability within groups

Reminder: ANOVA

I am sure you all remember from statistic courses:
We observe different mean values for different groups.



High variability within groups



Low variability within groups

Could it be just by chance?

It depends from the variability of the group means and of the values within groups.

Reminder: ANOVA

ANOVA-Table („ANalysis Of VAriance“)

	Degrees of freedom (DF)	Sum of squares (SS)	Mean sum of squares (SS/DF)	*F*-Value
Groups	1	88.82	88.82	30.97
Residuals	7	20.07	2.87	

Reminder: ANOVA

ANOVA-Table („ANalysis Of VAriance“)

	Degrees of freedom (DF)	Sum of squares (SS)	Mean sum of squares (SS/DF)	F-Value
Groups	1	88.82	88.82	30.97
Residuals	7	20.07	2.87	

Under the hypothesis H_0 „the group mean values are equal“ (and the values are normally distributed)
F is Fisher-distributed with 1 and 7 DF,
$p = \mathrm{Fisher}_{1,7}([30.97, \infty)) \leq 8 \cdot 10^{-4}$.

Reminder: ANOVA

ANOVA-Table („ANalysis Of VAriance“)

	Degrees of freedom (DF)	Sum of squares (SS)	Mean sum of squares (SS/DF)	F-Value
Groups	1	88.82	88.82	30.97
Residuals	7	20.07	2.87	

Under the hypothesis H_0 „the group mean values are equal“ (and the values are normally distributed)
F is Fisher-distributed with 1 and 7 DF,
$p = \mathrm{Fisher}_{1,7}([30.97, \infty)) \leq 8 \cdot 10^{-4}$.
We can reject H_0.

ANOVA in R

In R ANOVA is performed using `summary.aov()` and `summary()`.

These functions apply on a regression: result of command `lm()`.

`summary.aov()` gives you only the ANOVA table whereas `summary()` outputs other information such as Residuals, R-square etc ...

ANOVA in R

In R ANOVA is performed using `summary.aov()` and `summary()`.

These functions apply on a regression: result of command `lm()`.

`summary.aov()` gives you only the ANOVA table whereas `summary()` outputs other information such as Residuals, R-square etc ...

Lets see a couple of examples with self-generated data in R.

Contents

Model checking

When you perform a linear model you have to check for the pvalues of your effects but also the variance and the normality of the residues. Why?

Model checking

When you perform a linear model you have to check for the pvalues of your effects but also the variance and the normality of the residues. Why?

This is because we assumed in our model that the residues are normally distributed and have the same variance.

Model checking

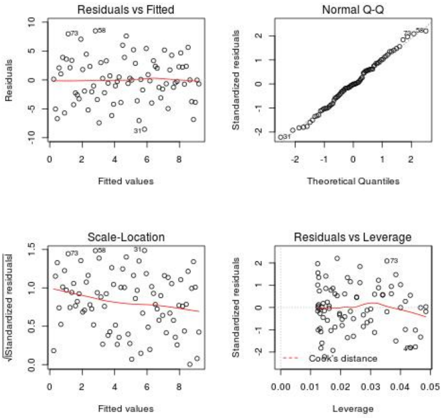
When you perform a linear model you have to check for the pvalues of your effects but also the variance and the normality of the residues. Why?

This is because we assumed in our model that the residues are normally distributed and have the same variance.

In R you can do that directly by using the function `plot()` on your regression object.
Lets try on one example. We will focus on the first two graphs.

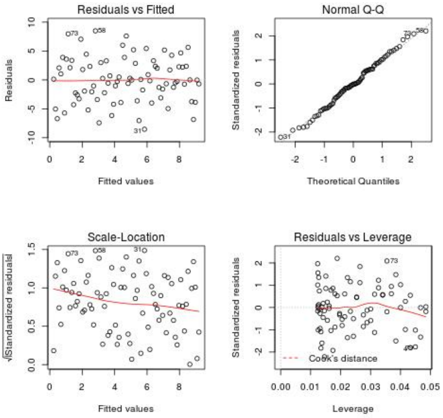
Model checking: Good example

This is how it should look like:



Model checking: Good example

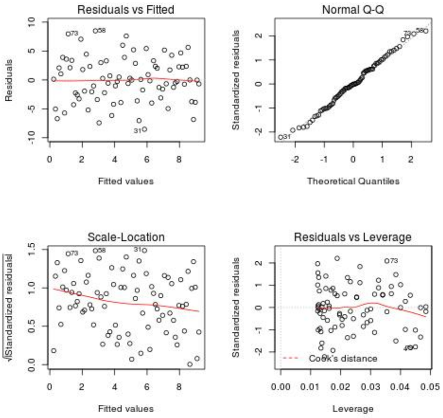
This is how it should look like:



- On the first graph, we should see no trend (equal variance).

Model checking: Good example

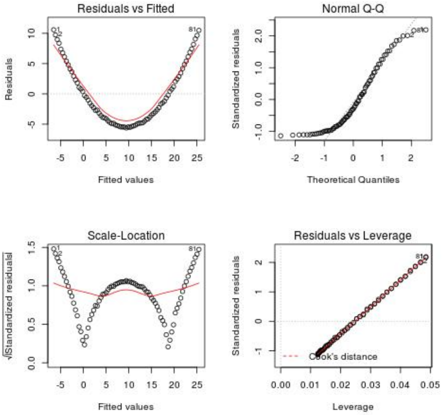
This is how it should look like:



- On the first graph, we should see no trend (equal variance).
- On the second graph, points should be close to the line (normality).

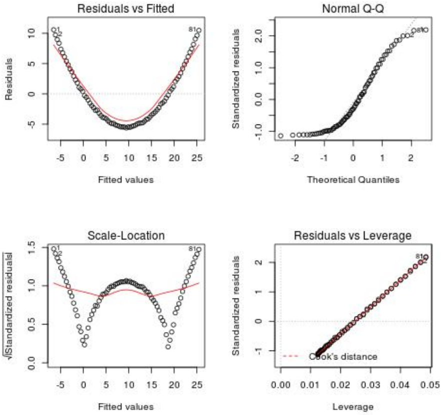
Model checking: Bad example

This is a more problematic case:



Model checking: Bad example

This is a more problematic case:



What do you conclude?