Data analysis with R

Lecture 7

Introduction to formal analysis

Jouni Junnila

Statistical models

- Statisticial models rely on probabilistic forms of description that have wide application over all areas of science.
- Often consists of a deterministic component as well as a random component.
  - The random component attempts to account for variation that is not accounted for by a law-like property.

Statistical models (2)

- Models should be scientifically meaningful, but not at the cost of doing violance to the data.
- As seen in the previous lectures, consideration of a model stays somewhat in the background in initial efforts at exploratory data analysis.
- In formal analysis the choice of model is of crucial importance!
- The choice may be influenced by previous experience with comparable data, by subject area knowledge and of course by exploratory analysis of the data.

Model components

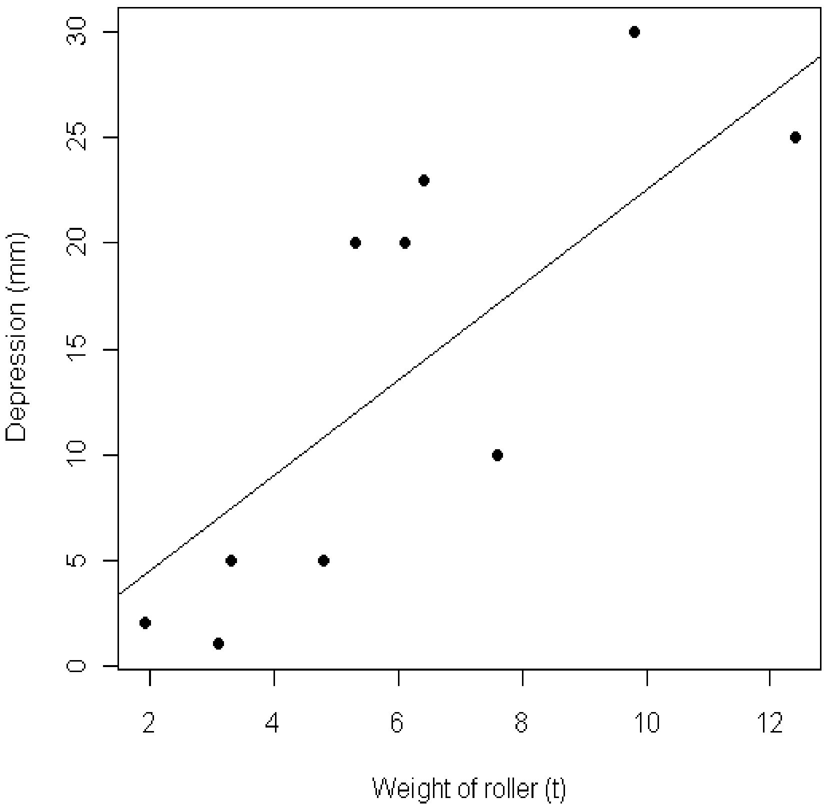
- Statistical models typically include at least two components. One component describes law-like behavior i.e fixed effects. The other is random, often thought as ”noise”, i.e subject to statistical variation.
- Usually we assume that the elements of random component are uncorrelated.
- Also in many cases we assume that the random components have mean of zero.

Example

- Let's consider an example where different weights of roller were rolled over different parts of lawn and the depression noted.
- We would expect the depression to be proportional to the roller weight.
- Drawing a scatter plot of the data shows is this really true.
- `plot(depression~weight, data=roller, xlab="Weight of roller (t)",`
- `ylab="Depression (mm)", pch=16)`
- `abline(0, 2.25)`

Example continues

- A slope of 2.25 seems to fit the data quite well.
- However we see, that the observations vary quite a bit and are not on the line.
- That's why we need a random variable to model the differences of the observations from the line.



Example model formula

- Our model formula would be
  - ✓ Depression = b *x* weight + error
- In the model *b* is constant, and that is the one we want to estimate. The error is different for each part of the lawn.
- If the error would be zero, all the observations would lie on the line and we could ignore it.
  - However, this is never the case in real-life.

Model formula

- In general we write a basic statistical model as follows:
  - *observed value = model prediction + statistical error*
  - *Or mathematically:* $Y = \mu + e$
- As said, the *e* tells us how much the actual observations differ of that what our fitted model estimates.
- This can be thought as the accuracy of the prediction.
- For assessing the accuracy we need residuals.
  - The more noise there is in the data, the more difficult is to conduct an accurate prediction.
  - Residuals are what is ”left over” after fitting the linear model. They are the estimate of the noise in the data.

Constructing a model

- The first duty of any model is to be useful. Model must yield inferences that, for its intended use, are acceptably accurate.
- Intended use can be eg. prediction, model parameters or in many cases both.
- Statistical model should reflect the data structure as good as possible and be also scientifically meaningful.

Model formula in R

- R's modeling functions use model formulae to describe the role of variables and factors in models.
- A large part of data analyst's task is to find the model formula that will be effective for the task in hand.
- By default, R-modelling functions fit the model with an intercept term added on.
  - This can be changed, though.

Example model formula

- Let's consider the previous example again. To fit the straight line to the data we can use a function called *lm (linear model).*
  - *lm(depression ~ model, data=roller)*
- Above will fit the model with the intercept. To remove the model formula is:
  - *lm(depression ~* ***-1*** *+ model, data=roller)*

Model assumptions

- Common model assumptions are normality, independence of the elements of the error and homogeneity of variance.
- There are some assumptions whose failure is unlikely to compromise the validity of analyses.
  - We say that the method used is *robust* against those assumptions.
- Other assumptions matter a lot.
- There are few hard and fast rules to decide is the assumption important or not.

Random sampling assumptions

- Usually, data analyst has a sample of values, that will be used as a window into a wider population.
- Almost all standard statistical methods require that all population values are chosen with equal probability,independently of the other sample values.
- However, often the sample is chosen at random, for example a survey can bee conducted in a shopping center, which results in bad quality of data.

Random sampling assumptions (2)

- In practice, analyst may make the random sampling assumption, eventhough the selection mechanism does’t guarantee randomness.
  - Inferences which are made based on this kind of data are less secure, than with random samples.
  - Random selection avoids the conscious and unconscious biases in the results.

Random sampling assumptions (3)

- Failing in inpendence assumption is a common reason for wrong statistical inference.
  - It is quite hard to detect, though.
- Data should be gathered so that the independence assumption is guaranteed.
- Because of the importance of independence, randomization in designed experiments and random sampling in sample surveys are so important

Checks of normality

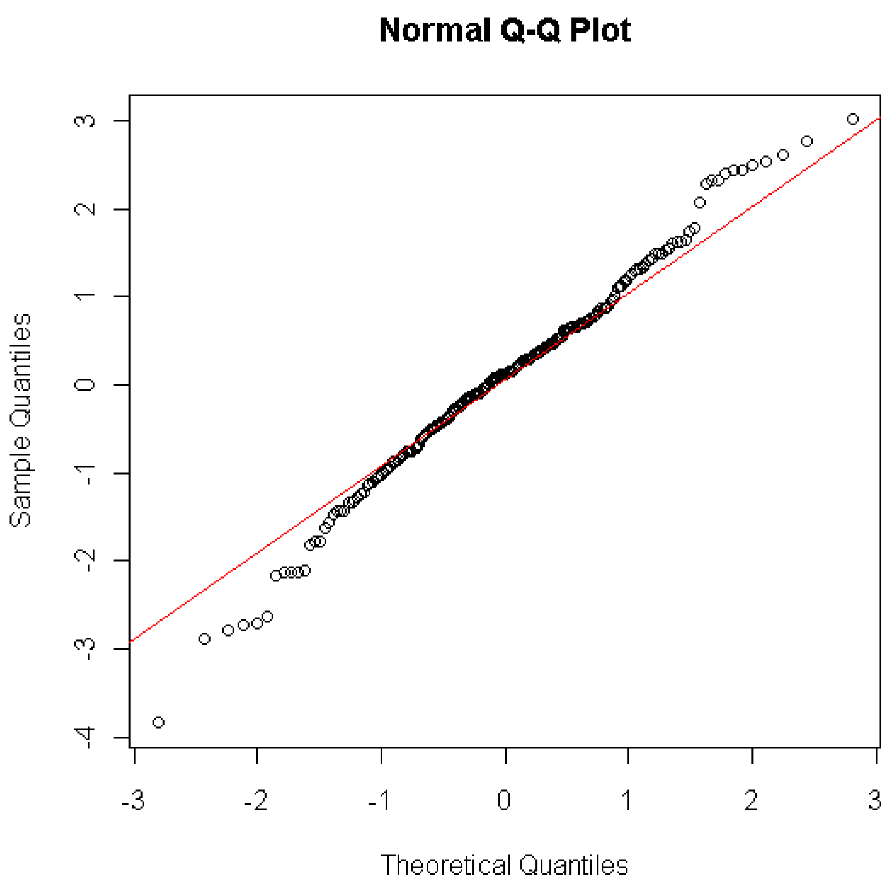
- Many data analysis methods rest on the assumption that the data are normally distributed.
- The question is how much departure from normality can we tolerate.
- Histograms and density plots are one way of checking this, but maybe more accurate possibility is to draw normal probability plot (QQ-plot)
- If the data are from a normal distribution, the QQ-plot should approximately be a straight line.

Normal probability plot

- To compare data with the normal distribution we can use a function called *qqnorm*. (Normal QQ-plot)
- In the graph we compare the quantiles from the data to the theoretical quantiles from the normal distribution.
- With *qqline* we can draw a line to the graph, where the observations should be, to make our investigation easier.

Normal probability plot; example

- > y <- rt(200, df = 5)
- > qqnorm(y)
- > qqline(y, col = 2)



Formal statistical testing for normality

- There are several statistical tests for normality.
- Problem with these tests is that normality is difficult to rule in small samples, while in big samples the normality assumption is practically always accepted.
- So we should rely also in something else (ie. graphs) in addition of the formal tests.
- Most common tests for normality are Shapiro-Wilk test (*shapiro.test)* and Kolmogorov-Smirnov test (*ks.test*).

Checking other assumptions

- As stated before, exploratory data analysis play an important role when checking assumptions before the formal analysis.
- Following the formal analysis, investigating the residuals of the model, is a good way to go, to make sure that everything is ok.
- You may find evidence of outliers, increase/decrease of standard deviation in the data or most importantly identify data points, that have high influence in the model estimates.

Non-parametric methods

- Classical non-parametric methods don’t have as much assumptions as does the parametric methods.
- However these methods might not be the answer if our parametric assumptions fail.
- If we ignore some structures of the data that we would consider with parametric approach, we loose valuable insights of the data.
- Non-parametric methods often assume too little, and that’s why these models are often unsatisfactory.