These are the exercises for the afternoon of Wednesday March 1. You do not need to hand anything in at the end of the session.

1 Linear regression

In sciences one is often interested whether there is a relation between two variables. A first simple approach is to visually check whether there is a linear relation between the variables. The aim of this lab is to introduce the basic statistical methods for finding and interpreting linear relationships. First we consider simple linear regression for which the model is

\[ y_i = a + bx_i + \epsilon_i, \]

where \((x_i, y_i)\) are the observed data. \(a\) and \(b\) are the parameters for our regression line and \(\epsilon_i\) are error terms also known as residuals. It is assumed that the \(\epsilon_i\)'s are independent and normally distributed with some variance \(\sigma\), \(\epsilon_i \sim N(0, \sigma^2)\). All of the model parameters \(a, b, \sigma\) can be estimated (better or worse) from the data \((x_i, y_i)\) with their significance levels. The idea is to find such values \(a, b\) that minimize the sum of the errors at the observed points \((x_i)\) i.e. minimizing,

\[ SS_{res} = \sum_i (y_i - (ax_i + b))^2. \]

The standard mathematical method for finding the best fitting line is “ordinary least squares” but we will not go into the details of the mathematics here. Instead we will focus on performing simple linear regression in R and interpreting the output.

2 Linear regression in R

The basic function used when working with linear models in R is \texttt{lm()}, an abbreviation for “linear models”. The syntax for \texttt{lm()}, which can seem confusing at first, is very general and we will later use it for many different kind of linear models. We will
study simple linear regression and the use of lm() by examining a sample dataset called “hormone.txt”. This data details hormone measurements by two different methods. One old and reliable, denoted Reference, and one new method, denoted Test. The aim is to assess the reliability of the new method by comparing its results to the old, which is assumed to produce correct answers.

1. Locate the dataset “hormone.txt” on the course webpage, download it and read it into R using read.table. Note, this dataset contains extra lines of text before the data starts, find a suitable parameter for read.table that allows you to skip the first few lines. Call the resulting data frame hormone. First assess whether the data seems to follow a linear relationship using a simple scatter plot.

   ```r
   plot(hormone$Reference, hormone$Test)
   ```

   Comment on what you see.

2. Now it is time to fit a linear regression to see how well the two methods compare to each other. Use the lm() function and store the result in an appropriately named object, for example.

   ```r
   myModel <- lm( hormone$Test ~ hormone$Reference )
   ```

   Note the “∼” notation, this is tells R to fit a model that explains Test as a function of Reference. Use summary on your new model and interpret the results. Look at for example the coefficient of determination called R-squared and the results of the tests, what are the hypotheses tested? Find the values for the coefficients, which of these tell us something about the linear relationship between our variables? Also try using the function abline() to add the line you have fit to the previously made scatter plot.

3. Before drawing any conclusions based on our model we need to check that our modeling assumptions hold. Of high importance is the assumption that the residuals are normally distributed. Extract the residuals from your model object and use qqnorm and qqline to check the normality assumption. Hint, you can either extract the residuals using the function residuals() or writing myModel$residuals. In the morning you were introduced to tests for checking normality, apply this to the residuals. What are your conclusions? Can we rely on conclusions drawn from our linear model?

4. Use the logarithm to transform both the Test and Reference values and create a new scatter plot. Can you see any telling differences between this and your previous plot?

5. Fit a new linear model using the log-transformed data. Check the output using summary and recheck the normality assumptions on the residuals for this new model. Can you tell a difference? Which model would you suggest using for comparing the two methods?
6. The aim was to compare the two methods for performing hormone measurements. If they matched perfectly what would we expect the regression parameters to be? Check the confidence intervals for the parameter estimations to see if we can conclude that the methods are different or not, Hint: the function confint().

3 Correlation

An important issue related with linear dependence between two data vectors is the concept of correlation. Formally let \( X \) and \( Y \) be two random variables, their correlation is defined as

\[
\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)} \sqrt{\text{Var}(Y)}}.
\]

\[
\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])].
\]

Intuitively this is a measure of the strength of linear dependence between the data. A coefficient of 1 or \(-1\) means perfect linear dependence, 0 means no linear dependence. Whilst this in itself is very useful one must remember that a correlation of zero does NOT imply independence. We only have that independence implies zero correlation.

We use a simple example to check how correlation can be calculated and tested in R. The data we will use is one of the built in data sets in R called “faithfull” add it to your workspace by writing data(faithfull). Use the help function to see what the data set describes.

1. First we visually inspect the data, produce a scatter plot for eruptions versus waiting. Is there a visual linear dependence?

2. Calculate the correlation coefficient, and also test whether the correlation coefficient is significantly different from zero. Hint, cor() and cor.test(). Are the results unexpected?

Calculating and testing the correlation between two sets of variables is quickly done using R. It is important to keep in mind that there are many common misconceptions regarding correlation. We cannot cover them all here but if you are interested I suggest looking at the Wikipedia page on correlation and dependence to get a deeper insight to what correlation can and cannot tell you about your data.

4 Exercises

These four exercises are based on the exercises at the end of chapter 6 in Dalgaard’s book. They are optional but specifically the first question on using linear models for predicting the response for new data might be of interest to you.

6.1 With the rmr data set, plot metabolic rates versus body weights. Fit a linear regression model to the relation. According to the fitted model, what is the
predicted metabolic rate for a body weight of 70kg? Use the predict() function with the option interval="p". What if we want other than 95% confidence interval? Look at ?predict.lm. The usage of predict() to predict a value is not intuitive, therefore this is some of the relevant code

> attach(rmr)
> lmObject<-lm(metabolic.rate~body.weight)
> newData<-data.frame(body.weight=c(70)) ## here is the important bit:
> # the newData must be a data.frame and have the same column names
> # as the data used in the model
> predict(lmObject,newData,interval="p")

6.2 In the juul data set, fit a linear regression model for the square root of the IGF–I concentration versus age to the group of subjects over 25 years old. What happens if this is done for all of the data?
6.3 In the malaria data set, analyze the log–transformed antibody level versus age. Make a plot of the relation. Do you notice anything peculiar (that’s what Peter Dalgaard asks)?
6.4 In Dalgaard’s book this exercise is about a method to draw from a two–dimensional normal distribution. Instead there is a R package mvtnorm (http://cran.r-project.org/web/packages/mvtnorm/index.html) that has a function rmvnorm() that generates vectors from a multivariate normal distribution with a given mean vector and covariance structure. Look at this package, draw some data points, plot the drawn data and plot density estimates of the marginal distributions.

References