Exercises for the course

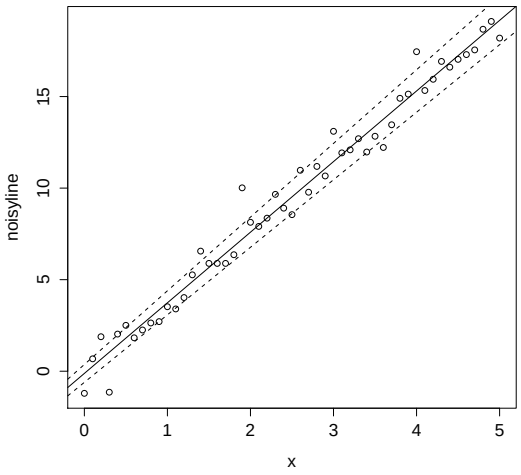
"An introduction to R"

Sheet 07

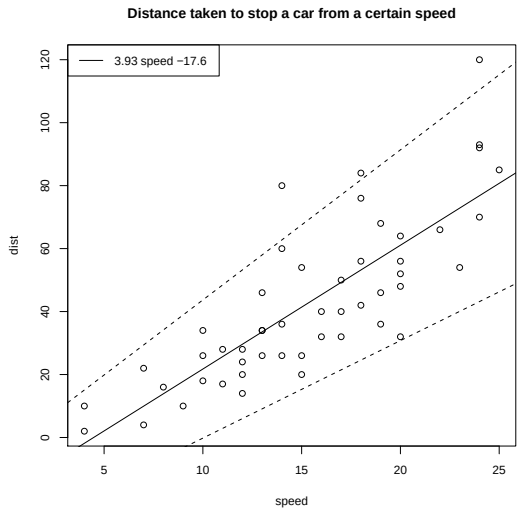
Exercise 37: *Supporting a hypothesis*
Apply a suitable test:

- The built-in data set `sleep` shows the effect of two soporific drugs (increase in hours of sleep compared to the control group) on 10 patients. Are the increases in sleeping hours of the two groups significantly different?
- Recall `heartbeats` from Exercise 12. Is the increase in weight of the heartbeat group compared to the control group significant? Then answer this question for every weight class separately.
- The common red spider (Bohnenspinnmilbe) is a vermin of agricultural plants to which plants react with the production of toxic substances. Do plants "remember" a former attack? Two groups of each 20 cotton plants are being infected with the mite. One group has never encountered this mite before. The second group has already survived a former attack. After a certain time, the mites on the plants are counted. You find the result in the file `mite.txt` on the web page. Is there a significant difference between the two groups? Compare the two groups visually (produce an appropriate plot) and apply a suitable test. Formulate an answer!

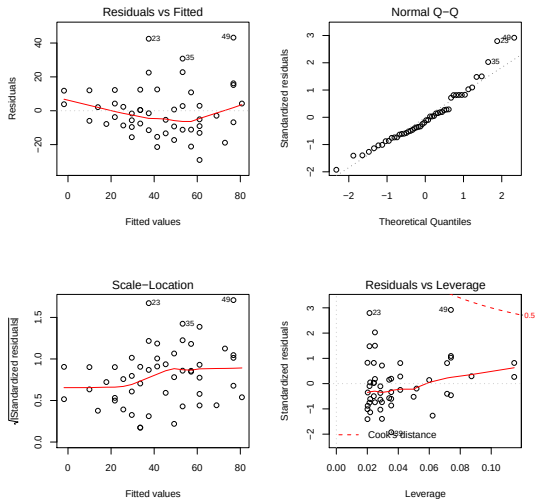
Exercise 38: Set the seed to 1234 to get the same picture. Define two vectors `x <- seq(from=0,to=5,by=0.1)` and `noisyline <- 4*x + rnorm(length(x))`. Explain the variable `noisyline` with a linear model of the independent variable `x`. Use the command `lm()` for this. Denote the returned object as `regr`. What is the Anova table of this regression? Read off the intercept and the slope of the fitted line with `coef()`. Extract the p-values of the intercept and of the slope from `summary(regr)`. What is the fraction of the total variation in `noisyline` that is explained by the regression? This fraction is called 'r squared' and is printed by `summary(regr)` under 'Multiple R-squared'. Check that this value is equal to $\big(\mathrm{cor}(\mathrm{x}, \mathrm{noisyline})\big)^2$. In order to visualize the linear model, plot `noisyline` against `x`. Add the regression line with `abline(regr)`. Next calculate confidence intervals of the fitted parameters with `confint()` applied to `regr`. Denote the object returned by this command as `cf`. Enter `cf` to view it. The two columns of `cf` define two lines. Add these two lines to the plot with line type "dashed". Your result should resemble the following figure.



Exercise 39: Recall the data set `cars` from Exercise 21. Plot `speed` as a function of `dist`. The plot suggests that `speed` depends linearly on `dist`. Find this linear dependence with a linear regression in which `dist` is the response variable. Store the returned regression object into the variable `regr`. What is the Anova table of this regression? Read off the p-values of the intercept and of the slope from `summary(regr)`. What is the fraction of the total variation in `dist` that is explained by the regression, that is, what is r-squared? Then add the regression line to the plot. Extract intercept and slope of the regression line from `regr` and round both values to 3 significant digits. These values are used for the legend of the following figure. In addition add the two lines which you get from the confidence intervals for the intercept and for the slope of the regression line.



Finally start a 2 by 2 multi-figure and enter `plot(regr)` to obtain the following 4 figures which can be used to check the linear model. (4 points)



Exercise 40: *Get a feeling for the reliability of t-test*
The command `t.test(x,mean=m)` tests the null hypothesis H_0: The sample x is a sample from a distribution with mean m.

- Sample 100 independent, standard normally distributed values. Use the command `t.test(,mean=0)` to test whether this data comes from a distribution with mean 0 (which it obviously does).
- Let us approximate the probability that the t-test rejects the null hypothesis $H_0 : \mu = 0$ on the significance level 5% although it is true. Repeat the last test with a `for()`-loop 10.000 times (100 standard normally distributed values in each loop). Use `t.test()$p-value` to count how often the t-test rejects the null hypothesis. Then divide this count number by 10.000. What do you think of the result?
- Repeat the last part (10.000 loops) but now with sample length 10 instead of 100. What is the result now?

Exercise 41: *Get a feeling for the power of the t-test*

- Sample 100 independent values from a normal distribution with mean 1 and standard deviation 1. Use the command `t.test(,mean=0)` to test whether this data comes from a distribution with mean 0 (which it obviously doesn't).
- Let us approximate the probability that the t-test rejects the null hypothesis $H_0 : \mu = 0$ on the significance level 5% if it is true. Repeat the last test with a `for()`-loop 10.000 times (100 values from a normal distribution with mean 1 in each loop). Count how often the t-test rejects the null hypothesis. Then divide this count number by 10.000. What do you think of the result?
- Repeat the last part (10.000 loops) but now with sample length 10 instead of 100. What is the result now?
- Now try the same but with true mean $\mu = 0.5$, $\mu = 0.1$, $\mu = 0.01$ and $n = 100$ and $n = 10$. Try also other values for the true mean, the true standard deviation and the sample size. Find out when the t-test has only little power to reject the null hypothesis $H_0 : \mu = 0$.