STAT 210
Applied Statistics and Data Analysis:
Problem list 7
(Due on week 8)

Exercise 1

The data for this problem is in the file `P1725F_P1.csv`, which has three variables, namely `weight`, `height`, and `species`. Read the data into a data frame named `q1_data`. The data correspond to measurements taken on samples of two species of draxen.

(a) Do a scatterplot of `weight` as a function of `height`, including the regression line. Fit a regression and print the summary table. Interpret the output in the table. State explicitly the assumptions that underlie this model. Write down the equation for the regression line, and give an interpretation of the parameters. Predict the `weight` for a draxen with `height` = 62.8 cm and include a confidence interval at the 98% level. Include your comments on every step that you take.

(b) There are two species of draxen in the file, denoted `A` and `B`, and this characteristic is available in the categorical variable `species`. If this variable was not read as a `factor`, transform it before you continue. Do a scatterplot of `weight` as a function of `height` and color the dots by species. Comment on what you observe. Fit a different model for each species and add the corresponding regression lines to the scatterplot. Compare the three models that you have fitted and comment. Write down equations for the two new models and compare with the previous one. Predict the value of the `weight` for draxen of both species having height 62.8 cm, including confidence intervals at the 98% confidence level. Compare with the previous prediction. Include your comments on every step that you take.

Exercise 2

For this problem we use the data set `cars`, available in `R`. The data give the speed of cars and the braking distances. Note that the data were recorded in the 1920s.

(a) Plot the data. Fit a regression model and add the regression line. Print a summary of the regression and interpret the output. Write down the equation for the regression line, and give an interpretation of the parameters.

(b) An important step after fitting a model is to check whether the assumptions on which the model was built are satisfied. This is a topic that will be discussed in next week's videos, but as an advance you will use two statistical tests for normality and homogeneous variances, respectively. The first test is the Shapiro-Wilk test that we have used previously. You should use this test on the standardized residuals, which can be obtained with the command `rstandard(mod)`, where `mod` is the name of the model you fitted using the `lm` function. The Levene test for homogeneous variances that we used for ANOVA is not suitable for a regression model because it requires that the data be grouped, which is not the case here. We use the `ncvTest` function available in the `car` package. The argument for this function is the name of the model that you fitted. Use both tests for the model fitted in (a) and comment. Do you think the assumptions are satisfied?

(c) Fit a new model for the square root of `dist` as a function of speed. Do a scatterplot including the regression line. Print a summary of the regression and interpret the output. Write down the equation for the regression line, and give an interpretation of the parameters. Use the tests introduced in (b) on the new model and compare with the previous results. Comment.

Exercise 3

For this exercise we will use the data set `Cars93` in the `MASS` package, that has information about 93 cars on sale in the USA in 1993.

(i) Draw a scatterplot of `MPG.city` against `Weight`. For this, use the function `scatterplot` in the `car` package. This function draws the points and also a simple regression line for the two variables. Moreover, it also plots a broken line that represents a local smoother function for the points as well as confidence bands for the smoother. The function also graphs boxplots for both variables on the corresponding axes. How would you interpret the differences between the regression line and the local smoother function that you see on the graph?

(ii) Use the function `lm` to fit a regression line to this data. Use the function `summary` on the output of the regression. Interpret the t-tests in the table. Are the parameters different from zero?

(iii) Write down explicitly the model that you get and interpret the meaning of the coefficients.

(iv) Describe the sampling distribution for the estimated parameters in this regression.

(v) Give confidence intervals at a confidence level of 98% for the parameters of the regression.

Exercise 4

For this question we will use the data set `cats` in the library `MASS`. which has the heart and body weights for a sample of cats.

(i) Draw a scatterplot of `Hwt` against `Bwt`. Color the points according to sex. Comment on what you see on the graph. Count how many data points for each sex there are.

(ii) Use the function `lm` to fit a regression line to this data. Add the regression line to the plot. Use the function `summary` on the output of the regression. Interpret the t-tests in the table. Are the parameters different from zero? Write down explicitly an equation for the model that you get and interpret the meaning of the coefficients.

(iii) Fit separate models for each sex. Plot a scatterplot of the data with points colored by sex and add the two regression lines. Comment.

(iv) Print summary tables for the two models you fitted in (iii) and discuss the results.

(v) Write down the equations for the two models fitted in (iii). Comment.