STAT 210
Applied Statistics and Data Analysis
Problem list 8
(Due on week 9)

Exercise 1

In this exercise we will use the data set `iris`.

(i) Extract the data corresponding to species `setosa` to a separate data frame. Plot the numerical variables for this set in a matrix of plots.

(ii) Use the function `scatterplot` from the `car` package to plot `Sepal.Width` as a function of `Sepal.Length`. Comment on the graph.

(iii) Fit a linear regression model for `Sepal.Width` as a function of `Sepal.Length`. Produce a table using `summary` and discuss the results.

(iv) Find the R^2 and verify that for simple linear regression, this coefficient is equal to the square of the correlation between the two variables.

(v) Write down the equation for the regression line and interpret the parameters.

(vi) Do the diagnostic plots for this model and comment.

(vii) In this case, the diagnostic plots give sufficient information about the normality assumption. However, if we wanted to test this assumption, we could use the Shapiro-Wilk test. Do this test and comment on the result.

(viii) The assumption of uniform variance is not so clear from the plots, particularly from the Scale-Location graph. The test we used for analysis of variance does not work here, because we do not have grouped data. A test that can be used in this situation is the Score Test, proposed by Cook and Weisberg (1983) and described in Applied Linear Regression by S. Weisberg, Wiley. This test is available in the `car` package as `ncvTest`. Do this test and comment on the results.

Exercise 2

For this exercise we will use the data set `birdsdiet.csv`, downloaded from https://wiki.qcbs.ca/r_workshop4. The set has the following variables

- `Family` Common name of family
- `MaxAbund` The highest observed abundance at any site in North America
- `AvgAbund` The average abundance across all sites where found in NA
- `Mass` The body size in grams
- `Diet` Type of food consumed. Categorical variable with five levels: `Plant; PlantInsect; Insect; InsectVert; Vertebrate`.
- `Passerine` Is it a songbird/ perching bird.
- `Aquatic` Is it a bird that primarily lives in/ on/ next to the water.

Load packages `MASS`, `car` and `alr4`.

(i) Load the dataset `birdsdiet.csv` and do an exploratory data analysis.

(ii) Focus on numerical variables and graph a scatterplot matrix.

(iii) Do a simple regression of `MaxAbund` on `Mass`. Look at the summary for this model and interpret the results.

(iv) Graph a scatterplot of the variables with the regression line. Comment.

(v) Do the diagnostic plots (standard and `residualPlots()`). Comment

(vi) Do histograms of response and regressor. Comment

(vii) Perform a Shapiro-Wilk test for normality for the two variables. Interpret the result.

(viii) Use `boxcox` to select a transformation.

(ix) Transform the data using `log10`, plot histograms and do Shapiro-Wilk again.

(x) Fit a new model with this data.

(xi) Plot the model on a scatterplot of the transformed data. Plot the diagnostic graphs.

(xii) Can you write down the equation for your model?

Exercise 3

For this exercise we will use the data set `ais` in the package `DAAG`, that has information about Australian athletes. Look up the help for this set and get familiar with the variables in it. To load the data set you need to load the library and run the command `data(ais)`.

(i) Plot a scatterplot of `wt` as a function of `ht`. Add the corresponding regression line.

(ii) Fit a regression to this data. What are the estimated values for the intercept and slope? Write down the regression model in this case and interpret the meaning of the coefficients.

(iii) What are the results of the t-tests in this example?

(iv) What would be the predicted value according to this model for the weight corresponding to a height of 1.95 cm?

(v) Describe the sampling distribution for the estimated parameters in the previous regression.

(vi) Give a confidence interval at a confidence level of 98% for the parameters of the regression.

(vii) Find the value for the R^2 and comment on its meaning.

(viii) Draw diagnostic plots and discuss the results.

(ix) Produce an anova table for the regression and interpret the test. How does this compare to the results of the summary table of the regression?

Exercise 4

For this exercise we will use the data set `ais` in the package `DAAG`, that has information about Australian athletes. Look up the help for this set and get familiar with the variables in it. To load the data set you need to load the library and run the command `data(ais)`.

(i) Draw a scatterplot of `bmi` against `lbm`. 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. Write down explicitly the model that you get and interpret the meaning of the coefficients.

(iii) Use the function `summary` on the output of the regression. Interpret the t-tests in the table. Are the parameters different from zero?

(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.

(vi) Draw a scatterplot of the data and include the regression line with confidence bands for the mean and predicted values.

(vii) Draw diagnostic plots and discuss the results.

(viii) Produce an anova table for the regression and interpret the test. How does this compare to the results of the summary table of the regression?