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 question use the dataset `PL825FQ2`.

This dataset has information on fuel efficiency, measured in miles per gallon, and seven other variables for 80 different car models. There are two variables related to fuel efficiency, `City.Mpg` and `Highway.Mpg`. We will only consider `City.Mpg`, and we will work with the reciprocal of this variable, 1/`City.Mpg`, which we will call `City.fc` for fuel consumption. We want to explore the relation between `City.fc` and the car's weight (`Weight`).

(i) Read the data and define a new variable called `City.fc` in the data frame equal to the reciprocal of `City.Mpg`. Draw a scatterplot of `City.fc` as a function of `Weight`. Fit a simple linear regression for `City.fc` as a function of `Weight` and add the line to the plot. Comment. Obtain a summary of the regression and comment.

(ii) Draw the diagnostic plots. Do you identify any point as an outlier? If you do, which point is this? Can you identify this point in the initial scatterplot? Can you find a reason why this point is different from the rest?

(iii) Fit a new regression model excluding the outlier(s) you identified in the previous section. Draw a scatterplot with both regression lines. Compare the summary tables. Draw the diagnostic plots and comment.

(iv) Run the Shapiro-Wilk test on the residuals for both models and compare the results.

Exercise 3

For this question use the data set `PL825FQ3`.

The data for this question come from an experiment to determine the relation between the volume of a gas and the pressure. The file has two variables, `Height` and `Pressure`. Height corresponds to the height of a cylindrical container with a fixed circular base with a movable top that allowed changing the volume of the container. `Height` was measure in inches. `Pressure` is measured in inches of mercury as in a barometer. We want to study the relation between these two variables.

(i) Read `data2` and plot `Pressure` as a function of `Height`. Fit a simple linear regression for `Pressure` as a function of `Height` and add the regression line to the plot. Comment. Obtain a summary for the regression and draw the diagnostic plots. Comment on the results

(ii) Use the function `boxcox` on the package `MASS` with the argument set to the model you fitted in (i). If the maximum value in the graph is close to an integer value, use a power transformation with exponent equal to the integer value for `Pressure` and fit a new model. Obtain a summary of the new regression and compare with the previous one. Draw the diagnostic plots and compare with the previous results.

(iii) If the p-value for the intercept is large, fit a model without intercept by adding `+ 0` at the end of the regression equation in the call to the `lm` function. Use this model to write down an equation for the relation between pressure and height for a gas. What would be the predicted `Pressure` for a point with `Height` $= 32$? Draw a scatterplot of `Pressure` against `Height` and add the regression line for the first model and the curve you obtained with the second regression.

Exercise 4

For this problem use the data set `PL825FQ4.txt`.

(i) Read the data and plot `yval` as a function of `xval`. Fit a simple linear regression for `yval` as a function of `xval` and add the regression line to the plot. Comment. Obtain a summary for the regression and draw the diagnostic plots. Comment on the results

(ii) Use the function `boxcox` on the package `MASS` with the argument set to the model you fitted in (i).

(iii) If the confidence interval in the graph includes zero, use a logarithmic transformation for `yval` and fit a new model. Obtain a summary of the new regression and compare with the previous one. Draw the diagnostic plots and compare with the previous results.

(iv) Write down the final model in terms of the original variables. Draw a scatterplot of `yval` against `xval` and add the regression line for the first model and the curve you obtained with the second regression.
