MAS programmes - Statistical Analysis (Autumn Term)

Exercises 4

1. Recall the `oil` dataset which has been analysed extensively in Lectures 3 and 4. The *chosen* model uses the explanatory variables `distil` and `endpoint` to predict the response `spirit`.

(a) In Lecture 4 we saw that observation number 32 had the highest value of *Cook's distance*, so that this data point exerts the greatest influence on the model (having both moderately high values of *leverage* and *standardized residual*). By examination of the data, explain why this point has high leverage.

The following command can be used to identify the values of the explanatory variables for observation number 32.

```
> oil[32,]
```

[*Hint*: you may also find it useful refer back to the exploratory analysis you undertook in Exercises 3].

(b) To illustrate what influence this observation actually exerts on the model, it is possible to refit the model omitting it, and noting any differences. This can be achieved using, for example

```
> oil232.lm<-lm(spirit~distil+endpoint,data=oil,subset=c(-32))
```

What is the fitted model with observation number 32 omitted? Recalculate the predicted value for the yield of petroleum spirit (`spirit`), when the two explanatory variables `distil` and `endpoint` take the values 200 and 400 respectively.

2. Carry out appropriate an appropriate *diagnostic* analysis for the simple linear regression model of `lconsump` on `price` for the `sugar` data set from Exercises 2/3. Does your analysis suggest any cause for concern?

[You might consider as part of your analysis the use of appropriate diagnostic plots of *standardized residuals*, *leverages* and *Cook's distances* as shown in Section 4.4 of the Lecture Notes].

3. In an investigation of air pollution, data on seven variables were collected for 41 American cities over the years 1969 to 1971. The data set is available as a csv file `pollution1.csv` from Moodle. In order to load the data into R use the `read.csv()` function.

Descriptions of the variables

variable name	description
SO2	annual mean concentration of sulphur dioxide (in micrograms per cubic meter)
temp	average annual temperature
manufact	no. of manufacturing enterprises employing 20 or more workers
popul	population size in thousands
wind	average annual wind speed in miles per hour
precip	average annual precipitation in inches
pdays	average number of days with precipitation per year
city	name of city

A multiple regression analysis is to be carried out with the variable `SO2` as the response variable and some or all of the other six variables as regressor variables.

(a) Carry out the regression of `SO2` on all six regressor variables and comment on whether any of the regressor variables can be dropped from the regression.

(b) Carry out two stepwise regression procedures, using the function `stepAIC`, one (i) with no regressor variables present and another (ii) with all regressor variables present. Write down the fitted equations of the models that are suggested in each case.

(c) Obtain and list the values of R^2, s and the AIC for each of the suggested models and for the full model with all regressor variables present.

(d) Which of these models would you use if you gave preference to (i) the most compact model, (ii) the model that minimized the AIC.