MATH 204 - EXERCISES 4

These exercises are not for assessment

Use General Linear Modelling, and stepwise selection, to find the best fitting model for each of the following data sets that are available from the website

`www.math.mcgill.ca/~dstephens/204/Data/`

1. The data in the SPSS data set **Mercedes.sav** are taken from the advertising pages of the London Sunday Times, presenting Mercedes cars for sale in the UK. The asking prices (in pounds sterling) are compared against various factors (type/model of car, age of car in six-month units based on date of registration), recorded mileage, and vendor)

   Interest lies in explaining the variation in price, due to these factors, identifying *outliers* (non-typical prices), and predicting prices. Data columns are:

   - Car number $1, \ldots, 54$.
   - Asking **price** in pounds.
   - **Type/Model** of car: this is a discrete **factor** taking five levels (0=model 500, 1=450, 2=380, 3=280, 4=200).
   - **Age** of car in six-month units, based on registration date; this is a continuous **covariate**.
   - Recorded **mileage** (in thousands); this is a continuous **covariate**.
   - **Vendor**; this is a discrete **factor** taking five levels (0,1,2,3 are different dealerships, 4 is "sale by owner").

   Using these data, use ANOVA and regression techniques to identify which factors are influential in explaining the variation in car price. Are there any anomalous car prices ?

2. **Oysters Data Set** (**OYSTERS.SAV**): Four replications (**rep**) are done in each of five experimental conditions (**trt**); the continuous covariate is measured weight at the beginning of the experiment (**initialwt**), and the response is measured weight at the end of the experiment (**finalwt**).

3. **Oranges Data Set** (**ORANGES.SAV**): This data set records the sales per customer of two varieties of oranges (**Q1** and **Q2**), at six stores (**store**), over six days (**day**), at prices (**P1** and **P2**) for the two orange varieties that changed from time to time at each store.

   For this data set, it is also possible to treat the two sales prices separately in two analyses, or together in a single analysis with variety as another binary factor predictor. To perform this analysis, the data need to be reformatted so that (**Q1** and **Q2**) are stacked on top of one another, with each row corresponding to a

   variety x **store** x **day**

   design.

4. **Cotton Data Set (COTTON.SAV)**: The dependent variable is the weight of useable lint (**lint**); one treatment is the variety of cotton (**variety**); another treatment is the distances in the spacing of planting (**spacing**), which takes two levels, 30 or 40). There are two replications plants per factor predictor level combination. The continuous covariate **bollwt** measures the total boll weight which can be used to predict the response variable.

5. **Doses Data Set (DOSES.SAV)**: An trial is carried out on different doses of two drugs (**type**), conducted across four hospitals sites (**bloc**). For each drug, three levels of **dose** (1 unit, 10 units, 100 units) are used. The response variable is an improvement measure (y).

   In this example, it is possible to treat **dose** first as a **continuous covariate**, on the original or log scale, perhaps using quadratic and higher, order terms, and then as a **factor predictor** taking three levels.