Week 5 Practical

There are three questions in this practical but only question 3 is assessed. It contributes 8.5% to your raw BS1 total mark. Write your answer to question 3 as a report. The deadline for submission is 12 noon Monday week 8 MT 2014 at the Statistics Department reception, 1 South Parks Road.

1 Diamonds

The prices of diamonds are known to depend on Carat: The weight of a diamond stone indicated in terms of carat units, where one carat is the equivalent of 0.2 grams and Clarity: An inclusion is a naturally occurring imperfection found within a diamond, and a diamond with no internal inclusion is classified as IF - internally flawless, a diamond which is very slightly included is classified as VS and finally a diamond which is very very slightly included as VVS.

In the data set “diam.txt” we have information on the caratage and clarity of the stones and their price.

Perform EDA and model selection to develop a parsimonious model which can explain the price (in US dollars) of a diamond stone and model diagnostics to check the validity of your model.

Comment on the scatterplots and box-plots below. Does the relationship between Price and Carats appear to be linear?

```r
plot(Price,Carats)
boxplot(Price~Clarity)
```

There is an apparent abnormality with the relationship between Price and Clarity, where diamonds which are internally flawless seem to have much lower prices. Plot Carats vs Clarity to see if you can explain this.

Choose a tentative model, say the model with the two-way interaction, perform model diagnostics and check if you need any transformations before proceeding to model selection.

```r
diam.fit1 <- lm(Price ~ Carats + Clarity + Carats:Clarity)
```

```r
par(mfrow = c(2,2))
qqnorm(rstudent(diam.fit1))
qqline(rstudent(diam.fit1))
plot(fitted.values(diam.fit1),rstudent(diam.fit1))
plot(hatvalues(diam.fit1));
abline(h = 2*length(diam.fit1$coefficients)/length(Price))
plot(cooks.distance(diam.fit1));
```
We see clear curvature in the residuals vs fitted values plot and a funnel shape. A Box-Cox transformation suggests $\sqrt{\text{Price}}$ as a transformation.

```r
library(MASS)
par(mfrow=c(1,1))
boxcox(diam.fit1)
```

Perform model diagnostics for the model with the transformed response. Are the assumptions of the model satisfied now? Are there any highly influential points?

```r
diam.fit2 <- lm(sqrt(Price) ~ Carats + Clarity + Carats:Clarity)
par(mfrow = c(2,2))
qqnorm(rstudent(diam.fit2))
qqline(rstudent(diam.fit2))
plot(fitted.values(diam.fit2),rstudent(diam.fit2))
plot(hatvalues(diam.fit2));
abline(h = 2*length(diam.fit2$coefficients)/length(Price))
plot(cooks.distance(diam.fit2));
abline(h = 8/(length(Price)-2*length(diam.fit2$coefficients)))
identify(cooks.distance(diam.fit2))
```

Using the transformed response check if you could drop the two-way interaction from the model using the F-test as a criterion.

```r
diam.fit3 <- update(diam.fit2,~.-Carats:Clarity)
anova(diam.fit2,diam.fit3)
```

The interaction terms are not significant at the 5% significance level and should not remain in the model.

Note that there was a highly influential point with a large cook's distance. Remove that observation from the model and repeat the analysis. Do you reach the same conclusion?

An alternative way of performing model selection is by using a model selection criterion, such as AIC. `stepAIC` in package `MASS` fits all possible models automatically and gives the model with the lowest AIC value:

```r
stepAIC(lm(sqrt(Price) ~ Clarity + Carats),
        scope = sqrt(Price) ~ Clarity * Carats, direction="both")
```

In this case we have started from the additive model and have used a stepwise regression process to perform model selection (direction="both"). Check that you understand the output. Does the selected model in this case agree with what you concluded using the F-test?
2 Regular exercise and weight

The headline reads:
“Regular exercise linked to heavier students!”

Oxford University undergraduate students reported their sex, weight in kg and whether they exercise weekly or not. The data are in "students.txt".

The following box-plots of their weights by their exercise habits (=Y if they do exercise weekly and N otherwise) suggest that students who do exercise regularly are considerably heavier than those who do not.

Figure 1: Box-plots of the weight in kg of the students by exercise habits.

Can you think of an explanation for this observed difference in weight between the two levels of variable WeeklyExercise, or is it really that weight and exercise are positively correlated?

Fit normal linear models to justify your answer. Do not forget to check the assumptions of the models and consider possible transformations of your response, if necessary.

```
# read the data
stud.data <- read.table("students.txt")
# change the variable names to something more convenient
names(stud.data)<-c("Gender","Weight","Exercise")
str(stud.data)
# Notice that Gender is a numeric variable - we want categorical
stud.data$Gender<-as.factor(stud.data$Gender)
levels(stud.data$Gender)
levels(stud.data$Gender)<-c("F","M")
# change the levels from "1", "2" to the more meaningful "F","M"
```
head(stud.data)
str(stud.data)
# looks better now

attach(stud.data)
# bring the variables from the data frame stud.data in the workspace

summary(lm(Weight ~ Exercise))
# WeeklyExercise is a significant explanatory variable for weight and the estimated coefficient is positive

summary(lm(Weight ~ Gender))
# Gender is also a significant predictor for weight (as expected)

# Your comments?
3 Assessed

Frog Respiration

Frogs of four species had their oxygen consumption measured at two temperatures and two exercise levels. There were two frogs of each species at each temperature, and each of the two was measured both at rest and during forced exercise. Think of Oxygen Consumption as the response, and the rest of the variables as potentially explanatory.

The data are set out in the following way.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subject</td>
<td>1-16</td>
</tr>
<tr>
<td>Species</td>
<td>1-4</td>
</tr>
<tr>
<td>Temperature</td>
<td>Low or High</td>
</tr>
<tr>
<td>Rest</td>
<td>Oxygen consumption (ml O2/g/hr) at rest</td>
</tr>
<tr>
<td>Exercise</td>
<td>Oxygen consumption during exercise</td>
</tr>
</tbody>
</table>

These data and description are available at

http://www.statsci.org/data/general/frogs.txt

and on the SB1a course website. The data come from


1. Perform exploratory data analysis and give a brief summary of the data.

2. Model the relation between Oxygen Consumption and the available explanatory variables using a normal linear model for the response, or some function of the response. You should consider possible interactions between the explanatory variables. Carry out outlier analysis and model selection, clearly describing and motivating each step.

3. Write the model equation of your selected model from part 2 and state its assumptions.

4. Comment on your findings.

5. Include all the R commands you used to answer these questions in an Appendix.

Your report should be clearly written. There are no marks awarded for presentation but there are marks awarded for clarity. When you make a statistical test do not just report the p-value but also report your conclusion using plain language. The same holds for interpreting models; reporting the coefficients and their standard errors is required but try to link your results to the research question as well. You should use captions for your tables and figures and include your commented R-code, preferably in an appendix.