Assignment 4, ST2304

Problem 1 Additative versus multiplicative effects:

1. Reanalyse the data in assignment 3 by first logtransforming the response variable. Again, based on the anova table you may want to fit a reduced model in which non-signficant terms are omitted. Based on the adjusted R^2 value shown in the output from `summary( )`, does this alternative model give a better fit to the data?

2. This alternative model is based on the assumption that the factors have an additative effect on log flighttime. The effects on flighttime itself are therefore multiplicative. The effect of a given factor level relative to the control level may thus be to increase the untransformed expected response by for example 20% regardless of the value of other explanatory variables. Compute the estimated effect of attaching a clip (in percent) on flighttime. Also compute an estimate of how much faster (in percent) a small helicopter falls to the ground relative to a large helicopter.

3. Compute confidence intervals for the parameter estimates by applying the function `confint` on the fitted model object. This gives you a matrix of confidence interval limits for the different parameters representing the effects on log flighttime.

   Transform these confidence intervals to confidence intervals for the corresponding percentwise increase or decrease in flighttime.

   Hint: Recall that if (A, B) is a $(1-\alpha)$-confidence interval for some parameter θ, and f is a strictly increasing function, then $(f(A), f(B))$ is a $(1-\alpha)$-confidence interval for the parameter $\theta' = f(\theta)$.

Problem 2 In this exercise we will estimate a model relating your grades in MA0001 Mathematical Methods A to various explanatory variables such as the number of hours you spent per week on assignments, lectures etc. You should enter data about yourself anonymously in this google docs spreadsheet. Make sure that you log out of your google account (from the pulldown-menu in the top right corner) before entering any data, otherwise your google identity will be visible in the revision history of the file. Fill in an optional id known only to yourself in the first column. Preferably everyone should have filled in data about themselves by friday evening.

The data set includes students from previous years. The grades for a random subset of 20 of these students have been given missing values in the spreadsheet (represent by NA-values). The objective of the exercise is to estimate and select a model using the non-missing part of the dataset (the training part), and then use your estimated model to predict the grades for the missing cases (the validation part of the data set). The details about how to this follows below.

Once the dataset has been completed, download the file in csv-format and load in into R and split it into two parts using the commmands (this depends on where you store the file)

```
grades <- read.csv("~/Downloads/grade prediction data 2015 - Sheet1.csv",skip=2)
trainingset <- grades[complete.cases(grades),]
validationset <- grades[complete.cases(grades[,-2])&is.na(grades$grade),]
attach(trainingset)
```

This way we use only cases which are complete to estimate the model and only cases for which grade is missing but which are otherwise complete as the validation set.

Descriptions of all variables are given in the heading in the spreadsheet.

1. First make a scatter plot of all variables using `pairs(grades[,-1])`. Also compute the correlation matrix between the variables using `cor( )`. You will need to exclude non-numerical variables and missing cases by using the command

   ```
   cor(grades[,sapply(grades,is.numeric)],use="complete.obs")
   ```

   Some of the explanatory variables measures overlapping aspects of each student and may thus be correlated. Does this seem to be the case for any of the variables?

2. You may use several strategies or combinations of strategies to select a model.

   - First, you may want to make some decision as to which variables you consider relevant. For example, facebook may only have an indirect effect by reducing the amount of time studying in the course. In a model including relevant variables representing the amount of time actually studying, it is hard to imagine how the additional time spent on facebook should have any effect. Similar arguments can perhaps be made for the variables fbfriends, training, gaming and partner although some research indicate that physical exercise aswell as gaming has a positive (albeit small?) effect on mental capabilities. Some variables may also perhaps, a priori, be consider irrelevant altogether.
   - Once you have decided which variables you consider for inclusion you may start by first fitting a model with all those variables present using, for example, the command

     ```
     mymod <- lm(grade ~ course +  lectures + assignments + reading
       + training + alcohol + work + age + sleep)
     ```

     You are free to include transformations of any of the variables if you think this makes sense.

     Then use the command `drop1(mymod, test="F")` and make a decision about which variables to exclude from the model. Then remove this variable from the model, refit the model using the above command.
   - Alternatively, you may start by fitting an model with only an intercept term using the command

     ```
     mymod <- lm(grade ~ 1)
     ```

     and then use

     ```
     add1(mymod,.~. + course +  lectures + assignments + reading
     + alcohol + work + age + sleep, test="F")
     ```

     repeatedly to test if additional terms should be added to the model. Only terms not already present in the model are considered for additon. The second argument specifies which variables are considered for inclusion, again, you may want to make some decision a priori about this.

     You may also use `drop1()` and `add1()` in combination, e.g. to check if any terms should be reconsidered for addition after deletion of other terms.
   - Yet another alternative is to use an automatic model selection procedure based on Akaike's information criterion (AIC), see `?step()`. This authormatic procedure can start with any initial model, e.g. the full model or a model only containing an intercept. Variables you want to consider for inclusion are specified with second scope-argument. For example, depending on which variables you want to consider for inclusion, you may use the command

```
fit <- lm(grade ~ 1)
fit2 <- step(fit,.~.+lectures+assignments+nassign+sleep)
```

Note that this procedure typically selects more complex models than the above approach based on hypothesis testing.

- It is useful to always examine the parameter estimates under different alternative models using `summary( )`. If some estimates goes in surprising directions, this may be an effect of chance, especially if the effect is only marginally statistically significant (p-value close to 0.05), and inclusion of the variable should be viewed with sceptisism.

Describe briefly how you arrived at your selected model. Examine the estimated parameters of your selected model using `summary( )` and comment on whether you think the model makes sense.

3. When you think that you have arrived at a reasonable model, store the fitted model object in an object called `mymod`. Using the command

```
predict(mymod,validationset)
```

you will get then predictions for the students in the validationset for which the response is unknow (to you). Comment on whether you think your predictions makes sense. Keep in mind that the predictions ought to be on a scale from 1 to 6.

If you format the output printed to the screen by predict using the command

```
cat(predict(mymod,validationset),sep="\n")
```

you can easily copy your predictions into this Google docs spreadsheet as follows. First select the numbers printed in the R console, press ctrl-C (cmd-C on Mac), and paste the numbers into an emtpy column in the Google docs spreadsheet using Ctrl-V. If you overwrite something by accident, undo the changes using Ctrl-Z.

Based on the deviations between your predicted values $\hat{y}_i$ and the observed values Y_i in the validation set known only by the lecturer, your predictions will be judged based on the root mean squared deviation

$$\sqrt{\frac{1}{20}\sum_i (Y_i - \hat{y}_i)^2} \tag{1}$$

between the observed and predicted values.