TMA4315 Generalized Linear Models Fall 2008
Assignment 2: Normal Linear Models

Deadline: 20.10.08

To be delivered in the mailbox of Hjartan Rimstad (Dep. of Mathematical Sciences, 7th floor, SB II) or sent electronically (pdf) to `rimstad@math.ntnu.no`.

The exercise consists of two parts, A and B. Only solution to part B is to be submitted.

Part A

Exercise 1: Model specification

Assume the true models for the relationship between gas consumption and mean external temperatures for one family before and after cavity-wall insulation was installed are:

$$y_i^B = 6.8 - 0.4t_{i_B} + \epsilon_i^B, \quad (1)$$
$$y_i^A = 4.7 - 0.27t_{i_A} + \epsilon_i^A \quad (2)$$

where y^A refers to gas consumption before insulation was installed and y^B to the same measure after installation of insulation, and $\epsilon_i^B, \epsilon_i^A$ are independent and Gaussian distributed with zero mean and variance $\sigma^2 = 0.1$.
The file `Insul.txt` contains simulated datasets from the above models. The aim of the exercise is to assess the effect of the insulation on gas consumption. To read the data use

```
data <- read.table("Insul.txt")
```

The **R** command `lm` fits a linear model and is the main command you are going to use in this exercise. Note that you may as well use the command `glm` for which "gaussian" is the default distribution family.

The main arguments to `lm` are

```
lm(formula,data,weights,subset,na.action)
```

where

`formula` is the model formula and the only argument which is always necessary. Note that the intercept term is included by default in the regression model, if you want to exclude it use the command `lm(y~a-1)` where `a` is the covariate you want to include

`data` name of the data frame (optional)

`weights` vector of positive weights for the data, only if non-uniform weights are needed

`subset` an index vector specifying a subset of the data to be used, by default all data are used

`na.action` specify how missing values should be handled

For more information on `lm` or any other **R** command you can type `help(lm)` or `?lm` in **R**.

a) Plot the gas consumption against the temperatures before and after insulation was installed. Is a linear model appropriate?
Hint: There are several ways to plot data in **R**. The simplest way to plot vector `x` against vector `y` is the command `plot(x,y)`

b) Fit two separate linear models for gas consumption and temperatures before and after insulation.
Hint: use the option `subset` in `lm` to select the right subset of data. Analyse the results, there are several commands in **R** which allows you to extract results from a fitted model, the most common are `summary`, `coef`, `resid`, `fitted`, `deviance`, `anova`, `predict`, `plot`.

c) Fit both regressions in the same "`lm`" model object, use the command

```
gas.lm<-lm(Gas~Insul/Temp, data=data)
```

Note: the terms `a/x` where a is a factor can be thought of as "separate regression models of type `1+x` within the level of `a`". Compare the results with what you got in **b)**

c) Fit again the model in **c)** this time excluding the intercept term (`Gas ~ Insul/Temp-1`). How does the parameter interpretation change?
Hint: check the model matrix by `model.matrix(gas.lm)`.

d) Verify that a different way to fit the model in **c)** is

```
gas.lm<-lm(Gas~Insul*Temp-1, data=data)
```

(the term `a*b` expands in `1+a+b+a:b`) what happens if you do not include the interaction term `a:b` ? Is this term necessary? What happens when you include the intercept term?
Hint: the `anova` command when used with two or more nested models gives the analysis of variance for those models.

Exercise 2: Model selection

The data set `Statedata.txt` contains data on 50 states in the US, the variables are: population estimates as of July 1975; per capita income; illiteracy in percentage of population; life expectancy in years; murder rate per 100.000 population; percent high-school graduates; mean number of days with temperatures <32 degree in capital or large cities; land area in square miles.This data set has been slightly modified from the original one collected from the Bureau of Census. We will consider life expectancy as the response and the remaining variables as predictors.

a) Construct pairwise scatter plots for each variable. Look at the relationship between the variables, compute the correlation matrix and point out obvious collinearity problems if any.

b) Fit a full model where all covariates are included. Are all the covariates significant?

c) Select a significant set of covariates using a stepwise procedure
Hint: you can update a model using the command

```
update(fullmodel,.  ~ .-Var)
```

where `fullmodel` is the previously fitted model and `Var` is the variable you want to exclude.

d) Check for possible outliers and influential points in the final model. Plot for example the residuals, the refitted values vs the observed ones, the residuals against the fitted values and the leverage for each point
Hint R gives you the possibility to compute the leverage by first extracting the model matrix `x<-model.matrix(model)` and then compute the diagonal of the H matrix by `lev<-hat(x)`, a "rule of thumb" is that the leverage is high if it is more than $2p/n$ where n is the number of points in the model and p the number of parameters.

e) If you find outliers or very influential points in the data set, try to fit the model again without those points and compare the results.

Part B

Exercise 3

The purpose of this exercise is to use a linear normal model to analyse a data set. Unlike in the first two exercises, you are here not given a step-by-step procedure for doing the analysis. Instead you should adapt the procedures used in the two previous exercises to this new data set.

Download the data file `Cars.txt`. It contains information about 84 makes of car on sale in the USA. For each make there are 6 variables recorded:

`MPG.city` City MPG (miles per US gallon by EPA rating)

`Type` the type of car : Small, Sporty, Compact, Large

`Origin` the origin of the car: USA or non-USA

`EngineSize` Engine size (litres)

`Horsepower` Horsepower (maximum)

`Fuel.tank.capacity` Fuel tank capacity (US gallons)

Consider the MPG as a response variable and the rest as predictors. Suppose that there is no interaction between the origin of the car and the type of the car and between the origin and the other technical feature of the cars. Propose a linear model for the city MPG as a function of all the other variables. Choose a parametrisation of the factors. Start from the full model where different regression lines with different intercepts are considered for each model type and select the best set of predictors for MPG. Give an interpretation of the parameter values. Verify your model, check for possible outliers and high influential points and if any consider the possibility of fitting the model again without the influential points and compare it with the previous one.