Question 1

The file *LawSalData.txt* is on the GLM Wattle page in the Datasets section. Data were collected in 1977 on 116 employees at a law firm in the USA who were hired between 1969 and 1971 as general office trainees. The following variables were collected for each employee.

- *Gender* (1=Male, 2=Female).
- *Age* - age at hire (in months).
- *EducYrs* - years of education
- *EducLvl* - level of education (1 = Graduate school, 2 = College graduate, 3 = Some college, 4 = High school graduate, 5 = None of the above).
- *WorkExp* - prior work experience (in months).
- *Start* - starting salary (in US dollars).
- *LoS* - length of service (total number of months of employment).
- *Salary* - current salary (in US dollars).

The basic research question of interest is: what are the important determinants of salary?

As *Gender* and *EducLvl* are categorical variables it is appropriate to declare them as factors. This can be done with the commands

```
> g=factor(Gender)
> e=factor(EducLvl)
```

where the factors are stored in `g` and `e`.

(**A**) Before starting on model fitting it is good practice to perform an exploratory data analysis (EDA), i.e. calculate basic summary statistics, obtain box plots and scatterplots of the variables and, ideally, perform other EDA activities. Perform an EDA of the Law Firm dataset. Due to the 10 page restriction you will need to think carefully about which tables, figures and plots to include and which to exclude. Comment briefly on all items that you do include. [One of my mottos is: if a figure/table/plot is not sufficiently interesting to comment on in the text then it is not sufficiently important to include in the report.]

[*12 marks*]

[Comments: My expectation is that you will have done some exploratory data analyses in at least one previous course. However, as this is not something we have discussed explicitly I will mention a few R functions that may be useful: `pairs()`, `summary()` and `boxplot()`. You may well know of or be able to find other R functions (e.g. using google) which are useful for EDA. To use the functions mentioned above, suppose that `x1`, `x2` and `x3` are numerical vectors of the same dimension which have been defined in the R environment. Then type

```
> a=cbind(x1,x2,x3)
> pairs(a)
> summary(a)
> boxplot(a)
```

Some of these functions have useful options which you can explore. Try googling these functions to find out more.]

(**B**) In this part you are asked to perform model selection using AIC and then to provide a more detailed study of the selected model using e.g. residual plots, qqplots and investigating fitted values. As we have not discussed AIC in the lectures in much detail, a brief summary is given here, focusing on the context of the standard linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ given in formula (7) of Chapter 1 of the lecture notes. For a linear model $\mathcal{M}$, AIC is defined by

$$AIC(\mathcal{M}) = 2K + n\log\{SS(res|\mathcal{M})/n\}, \tag{1}$$

where $SS(res|\mathcal{M})$ is the residual sum of squares for the model $\mathcal{M}$ and K is the number of free parameters in $\mathcal{M}$; here $K = p + 1$ when the parameter vector $\boldsymbol{\beta}$ is of dimension p (the 1 in $p + 1$ is for σ^2). There are three important points to note about AIC.

(i) The term $2K$ should be thought of as a penalty for the complexity of model, $\mathcal{M}$, where model complexity is measured by the number of parameters in the model.

(ii) The second term on the right hand side of (1) can be thought of as a measure of the goodness of fit of the model; the smaller its value, the better the fit.

(iii) When performing a search based on AIC for a "best" model among a class of models, we choose the model with the *smallest* value of AIC.

A search for a best model using AIC in R may be performed with the `stepAIC()` function. Information on how to use this function is given below.

Use `stepAIC()` to select a best model for the Law Firm Salary data. In your search you should only consider main effects and not consider possible interactions at this point. Then perform a more detailed study of the selected model by looking at residual plots, qqplots and plots involving fitted values. As far as possible you should check the assumptions in the model. Present selected outputs and summarise your findings in words. [*12 marks*]

[Comment: An example of the use of the function `stepAIC()` is given below. It is assumed that `y` is the response vector and `x1, x2,` and `x3` is each a factor or numerical covariate, and that all these vectors are of the same dimension. The following commands call the function `stepAIC()`.

```
> library(MASS)
> out1=lm(y~1)
> stepAIC(out1,y~x1+x2+x3,direction=c("both","backward","forward"))
```

To extract standardised residuals from the output obtained via out2=lm(), type

```
> r=stdres(out2)
```

and to extract fitted values from the same fit, type

```
> yhat=fitted.values(out2)
```

where `r` contains the standardised residuals and `yhat` contains the fitted values.]

(**C**) Starting with the final model obtained in part (**B**) of this question, try adding interaction terms to the model. Is there a case for including any of the interaction terms in the final model? Use AIC to explore this question. Present your results in the form of a table and comment on your results. [*12 marks*]

[Comment: to extract the AIC value from the output obtained via `out3=lm()`, type

```
> a=AIC(out3)
```

and then a will contain the value of AIC for that model.]

(**D**) This part of Question 1 is deliberately open-ended and it is for you to choose what you want to do. You are required to complete one work package. Possible work packages include:

(i) A study of whether there seem to be any outliers are present in the Law Firm Salary data and if so to provide an assessment of what their effects are on the analysis.

(ii) Exploration of the use of another model selection procedure and comparison of the results with those in part (**B**) where AIC was used. There are many possibilities here; just applying one alternative method would suffice.

(iii) Further investigation of any issues that arose in your EDA in part (**A**) which have not been addressed elsewhere in your response to Question 1 and/or application of further EDA techniques to the Law Firm Salary data.

(iv) Anything else that is broadly of relevance and would be of general interest and/or might help to address the basic research question of interest.

[*12 marks*]

(**E**) In this section you are asked to provide a non-technical summary of your main findings. Your summary should target a **generally well-informed audience that has no knowledge of statistics beyond high-school level.** The summary should be not more than one page including any outputs such as plots, figures or tables. If your summary is too technical then you risk losing marks. [*12 marks*]

Total marks: 60.