Tutorial Problems

Question 1

To determine the effect of exhaust index (in seconds) and pump heater voltage (in volts) on the pressure inside a vacuum tube (in microns of mercury), three exhaust indices and two voltages are chosen at fixed levels. It was decided to run two trials for each combination of index and voltage so in all 12 trials were run. A completely randomised design was deemed appropriate. Assume that we use the following model for the response

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}, \quad \epsilon_{ijk} \sim N(0, \sigma^2)$$

where $i = 1, 2$ indexing the voltage level, $j = 1, 2, 3$ indexing the exhaust index and $k = 1, 2$ indexing the replicate.

The results of the 12 experiments are given below.

Pump Heater	Exhaust index			
Voltage	60	90	120	Totals
127	48, 58	28, 33	7, 15	189
220	62, 54	14, 10	9, 6	155
Totals	222	85	37	344

(a) Calculate the interaction sum of squares and hence complete the below analysis of variance table. Test for an interaction effect. Note that the sum of squares of observatons ($\sum_{i=1}^{2}\sum_{j=1}^{3}\sum_{k=1}^{2} Y_{ijk}^2$) is 14,988. (Lecture 18 and 19)

Source of Variation	df	Sum of Squares
Voltage(V)		
Exhaust Index (E)		4608.17
Interaction(V $\times$ E)		
Residual	6	139.00
Total		

(b) Calculate a 95% confidence interval for the average difference in pressure at the two voltage levels when the exhaust index is set at 90 using a two-way ANOVA interaction model. (New material)

```
qf(c(0.95, 0.975), 2, 6)
```

```
## [1] 5.143253 7.259856
```

```
qt(c(0.95, 0.975), 6)
```

```
## [1] 1.943180 2.446912
```


Computer Problems

Question 1

A very famous data set is (Fisher's or Anderson's) iris data set, which gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris setosa, versicolor, and virginica. The data set is already available in R and can be called by typing `iris`

(a) Store the 50 observations for the setosa species in a new data frame called `dat`.
(b) For setosa, draw boxplots in a single frame for the four variables `Sepal.Length`, `Sepal.Width`, `Petal.Length` and `Petal.Width`.
(c) For setosa, draw Q-Q-plots in a single frame for the four variables and showing the variable name as a heading for each of the four plots.
(d) Use the recommended `shapiro.test` to assess the normality assumption for those four variables for the setosa species.
(e) Explore the other normality tests shown in Lecture 20. Do you get the same answers as in (d)?

Question 2

The aim of this question is to give you a real-life experience of what it means to analyse a data set with the aim to produce a model that is optimal from a non-standard point of view. Consider the data `phys.txt`.

(a) Read the data into R and store it in a data frame called `dat`. What are the variable names (given in the first row of the txt file)?
(b) Fit a full model, i.e. `lm(Mass ~ .,data=data)` and note that the adjusted R-squared is 0.9565, i.e. already large.
(c) Check the diagnostic plots for the full model and report any possible violations of the standard assumptions for the errors in the linear regression model. If there are no violations, explain why there are none.
(d) Find that model that gives you largest adjusted R-squared value by using exactly two variables. To do so efficiently, write one or more loops and comment the code, so that it can be easily understood by your peers.
(e) Create two new variables, `z1` and `z2`, that are functions (not necessarily linear functions) of any combination of the explanatory variables such that `lm(y ~ z1 + z2)` has larger adjusted R-squared value than the model in (d).

