Assignment 2: For 10% module credit (MSc group);
25% Module credit (PhD group)
Submission deadline: 1.00pm Monday 27th February 2017 - hard-copy only!

1. **Introduction**

   Researchers in the Radiation and Environmental Science Centre are conducting research into the effect of solar radiation on the mortality of human skin cells. Colonies of human skin cells are placed in medium and transferred to wells on experimental plates. These plates are then exposed to radiation in a solar simulator for various amounts of time from 0 (control) to 3.5 minutes minutes. After exposure, the number of live cells in a sample are counted under a microscope and the total number of live cells in the colony is extrapolated from the sample result. The purpose of this experiment is to assess the effect of the varying times of exposure to radiation on cell death. Due to inherent variability in the response of cells, the researchers replicate their experiments a number of times. Their is also concern that environmental conditions within the lab from day to day may have an impact on the results. Therefore, the experiment is repeated over a number of days. Initially the experiment is designed with complete balance; i.e. the same number of replications at each exposure time on each day. But due to experimental loss, complete balance was not reflected in the final data (e.g. some plates were contaminated and had to be discarded).

   The dataset is skincells.xls and is available on Webcourses. The variables are:

Variable	**Description**
day	The day (number code) that observation was recorded.
time	The amount of radiation exposure in minutes the colony was exposed to in the solar simulator.
logcells	The logarithm (base 2) of the number of live cells in the colony extrapolated from the sample result under the microscope.

   (a) Using an R programme read these data into a data.frame and obtain basic descriptive statistics/plots of the data. Present salient features from this exploratory data analysis in your report.

   (b) Using an R programme, analyse these data considering the two predictors, i.e. time (continuous) and day (categorical) and the response logcells. You should consider the possibility of interactions between the predictors. Submit your complete R programme code with suitable explanatory in-code comments.

   (c) Report your findings (5 pages maximum): summarise the relationship between predictor(s) and the response(what are the average effects, confidence intervals, results of tests of hypotheses etc.). Present plots where appropriate.

   NB: Reports exceeding the 5 page maximum will be penalised. See an example report on the Turkey data.

   [50 marks]

2. **Underfitting:** Imagine a true regression model is represented by a partitioned X matrix and β vector:

$$Y = \left[\; X_A \; \middle| \; X_B \; \right] \left[\frac{\beta_A}{\beta_B} \right] + \epsilon$$

where X_A is a $n \times p$ regression matrix, β_A is $p \times 1$, X_B is a $n \times q$ regression matrix and β_B is $q \times 1$. An underfitted model is fitted to the data:

$$Y = [X_A][\beta_A] + \epsilon$$

Show that:

$$\begin{aligned} E[\hat{\beta_A}] &= \beta_A + A\beta_B \neq \beta_A \\ E[s^2] &= E\left[\frac{1}{n-p} Y'[I - H_A]Y \right] \\ &= \sigma^2 + \beta_B'[X_B - X_A A]'[X_B - X_A A]\beta_B > \sigma^2 \end{aligned}$$

where $A = (X_A'X_A)^{-1}X_A'X_B$ is the Alias matrix.

[20 marks]

3. **Overfitting:** Imagine the true regression model is:

$$Y = [X_A][\beta_A] + \epsilon$$

where X_A is a $n \times p$ regression matrix, β_A is $p \times 1$.

The overfitted model is:

$$Y = \left[\; X_A \; \middle| \; X_B \; \right] \left[\frac{\beta_A}{\beta_B} \right] + \epsilon$$

where X_B is a $n \times q$ regression matrix and β_B is $q \times 1$.

Show that:

$$E[\hat{\beta_A}] = \beta_A$$

$$E[s^2] = \sigma^2$$

[30 marks]

Hint: The following result (easily proved!) for the inverse of a partitioned matrix is helpful: Let A be the following $m \times m$ partitioned symmetric matrix:

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right]$$

where A_{11} is of dimension $p \times p$; A_{22} is of dimension $q \times q$; A_{12} is of dimension $p \times q$; and $m = p + q$. The inverse of such a partitioned matrix is:

$$A^{-1} = \left[\begin{array}{c|c} A_{11}^{-1}(I + A_{12}FA_{21}A_{11}^{-1}) & -A_{11}^{-1}A_{12}F \\ \hline -FA_{21}A_{11}^{-1} & F \end{array}\right]$$

where $F = [A_{22} - A_{21}A_{11}^{-1}A_{12}]^{-1}$.