STATS 3001 Statistical Modelling III
PROJECT
Due: noon Monday 4th June (Week 13), 2018

IMPORTANT : Projects must be submitted with a signed Assessment Cover Sheet. All members of the group should sign the form. These forms are available on MyUni/Canvas under `Project`→`Project cover sheet`. Please note that Project marks cannot be counted for your assessment unless a signed declaration is received.

Please self-select yourselves into groups of between three and five people, and let me know your group membership by email. If necessary, I will re-allocate students to ensure there are at least three people in each group.

Each group member will receive the same mark that is awarded to the group for the Project. The Project contributes 10% towards your final mark for this course.

Check off the following prior to submitting your Project:

- ☐ Your project has been completed using Rmarkdown, Latex or Word and the final document is a single pdf file.
- ☐ Sufficient working has been provided in response to each question to satisfactorily demonstrate to me that you understand the required concepts and steps in the analysis.
- ☐ All R output and plots to support your answers are included where necessary. These can be included in the Project text where appropriate, or included as an Appendix. Ensure that all output, figures and tables are appropriately labelled and cross-referenced in the text.
- ☐ One Project as a pdf file per group is to be submitted electronically to MyUni (details to follow).

Professor Patty Solomon

PART A: Multiple linear regression

Heart catheterisation is sometimes performed on children with congenital heart defects. In this procedure, a catheter is inserted into a major vein or artery in the femoral region and pushed into the heart to obtain information about heart physiology and function. An experiment to investigate whether it is possible to predict the required length of catheter was conducted. The variables Height, Weight and catheter Length were recorded for each of 12 children, and are given in the table below.

Child	Height (cm)	Weight (kg)	Length (cm)
1	108.7	18.14	37.0
2	161.29	42.41	49.5
3	95.25	16.10	34.5
4	100.33	13.61	36.0
5	115.57	23.59	43.0
6	97.79	7.71	28.0
7	109.22	17.46	37.0
8	57.15	3.86	20.0
9	93.98	14.97	33.5
10	59.69	4.31	30.5
11	83.82	9.53	38.5
12	147.32	35.83	47.0

(1) Enter the data into R. Examine the relationships amongst the variables in the usual way and comment.

(2) Fit the following sequence of models:

```
lm(Length~Height+Weight, data=catheter)
lm(Length~Height, data=catheter)
lm(Length~Weight, data=catheter)
```

(3) The multiple regression model in (2) can be stated as

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + e_i,$$

where $e_1, e_2, \ldots, e_n$ are independent $N(0, \sigma^2)$ realizations of the true errors $\mathcal{E}_i$. List the model assumptions that can be checked using the residuals.

(4) Obtain appropriate diagnostic plots of the residuals for the multiple regression model. Making explicit reference to the diagnostic plots, explain whether each of the above assumptions is reasonable in this case.

(5) Consider the two simple linear regression models fitted in (2) above.

(a) Compare the estimated regression coefficients for `Height` and `Weight` from the simple linear regressions to those obtained jointly in the multiple regression model, in terms of numerical value and also statistical significance.

(b) Discuss briefly the different interpretations of the coefficient of `Weight` in the two contexts and give an explanation for the differences.

(6) The multiple regression model can be described in terms of two subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, relating to the two simple linear regressions, one with `Height` as the predictor variable, and one with `Weight` as the predictor variable. Let $\boldsymbol{x_1}$ denote the vector of `Height` values and let $\boldsymbol{x_2}$ denote the vector of `Weight` values.

   (a) Specify the two subspaces $\mathcal{L}_1$ and $\mathcal{L}_2$, and also $\mathcal{L}_1 \cap \mathcal{L}_2$.

   (b) Specify the two subspaces $\mathcal{L}_1 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}$ and $\mathcal{L}_2 \cap \{\mathcal{L}_1 \cap \mathcal{L}_2\}^{\perp}$.

   (c) Noting that the subspaces in (b) are one-dimensional, calculate the angle between the two spaces and comment in light of the preceding statistical analyses.

(7) Select a regression model for the data, carefully justifying your choice using diagnostics and appropriate statistics.

[Total: 35]

PART B: Logistic regression

Mammography is the most effective method available for breast cancer screening. However, the low predictive value of breast biopsy resulting from mammograms leads to approximately 70% of unnecessary biopsies with benign (non-malignant) outcomes. To reduce the high number of unnecessary breast biopsies, several computer-aided diagnosis (CAD) systems have been proposed in recent years. These systems help physicians in their decision about whether to perform a breast biopsy on a suspicious lesion seen in a mammogram, or to perform a follow-up examination instead.

The dataset `mammo.txt`[1] contains data on the true status of 961 mammographic mass lesions given by the variable `Severity`, where 0 =benign is not cancer and 1 =malignant is cancer, the patient's age in years (`Age`), and three BI-RADS attributes which are described below:

- `Shape`: round= 1, oval= 2, lobular= 3, irregular= 4;
- `Margin`: circumscribed= 1, microlobulated= 2, obscured= 3, ill-defined= 4, spiculated= 5;
- `Density`: high= 1, iso= 2, low= 3, fat-containing= 4.

BI-RADS stands for Breast Imaging Reporting and Data System and was established by the American College of Radiology. The dataset also contains an assessment of the masses by physicians that have been identified on full field digital mammograms (the variable `BI.RADS`). This variable is not a predictor variable however and is excluded from the present analysis.

Your Project task is to firstly, obtain the best predictive model for mammographic mass severity using the available predictor variables, and secondly, to obtain predicted probabilities of mass severity from your final model that can used by clinicians in their decision-making.

Your Project report should include the following sections:

- Introduction (description of data and purpose of the analysis). (5 marks)
- Data entry and data cleaning. (5 marks)
- Data visualisation and data summaries. (10 marks)
- Model fitting and model selection. (5 marks)
- Justification for choice of final model. (5 marks)
- Interpretation of parameters from final model. (5 marks)
- Predicting probabilities and interpretation. (10 marks)

[Total: 45]

[1] Donated by M. Elter, Fraunhofer Institute for Integrated Circuits, Image Processing and Medical Engineering Department, Erlangen, Germany; `https://archive.ics.uci.edu/ml/machine-learning-databases/`

PART C: GROUP WORK

Write a brief summary (not less than a paragraph and no more than one page) of how your group work was undertaken. For example, by face-to-face meetings, email or social media, etc. State when you commenced the Project, how often you met, or communicated (approximately), give a timeline for work completed, etc.

Each group member should also include a brief statement of their contribution to the work.

[Total: 10]

PROJECT MARKS SCHEME

- **Part A**: 35 marks
- **Part B**: 45 marks
- **Part C**: 10 marks
- **Presentation**: 10 marks
  - Your report should have a title page with Project title, names of authors and the date of completion.
  - Plots and tables should have captions, and be correctly cross-referenced in the text.
  - Expression and spelling will be taken into account where appropriate.
  - R code may be included in the text where appropriate to show results or to demonstrate an analysis, but most of it should be put into an appendix.
  - All mathematics should be formatted and displayed correctly.
- **TOTAL**: 100 marks.

April 25, 2018