ST404 ASSIGNMENT 2

1. Introduction

It is a sad fact in the medical world that sometimes, a patient with cancer can complete their treatment and be given the all-clear, only for their cancer to return later.

With many forms of cancer, then, patients are monitored after treatment, to ensure that if their illness returns, they can be treated as soon as possible.

The data you be using in this assignment comprises medical information collected from 197 people who were diagnosed with breast cancer. Each patient was given a full course of treatment, and then tested every month for recurrence of the cancer. The test in question could return either return a value of 0, which means the cancer has not *and cannot* return, a value of 1, which means the cancer has returned, or a value of 0.5, which means that the cancer has not returned yet, but still may in the future.

Each of the patients in the data set ultimately received a value of 0 or 1, at which point the time in months between them completing their initial treatment and them receiving a result of 0 or 1 was recorded.

Your group has been contacted by a group of oncologists, who want to better understand how long they should monitor patients who have completed a course of treatment for breast cancer. In particular, they are wondering whether there is a certain time point at which, if a patient's cancer has not yet returned, they can decide to stop testing a patient because it has become sufficiently unlikely that the cancer will return.

This means that they have both an interest both in how to predict monitoring times (so they can tailor their approach to new patients), and in exploring the relationship between monitoring times and the other covariates (as this can help them in the more general goal of undertanding how cancer recurrence operates).

Your task is to use this data to create a model which you believe offers an appropriate balance between:

(a) predictive power (clinicians need to be able to rely on this model's predictions for what effects the amount of time before a conclusive test result);
(b) explanatory power (clinicians want the model to give them a broader understanding of what affects the amount of time before a conclusive test result);

(c) simplicity (clinicians want the model to be simple enough for them to at least remember the basics of the model whilst they're busy testing hundreds of patients every month).

(Note: the data in this assignment is real. I have however slightly tweaked the meaning of some of the variables, to make them more suited to a linear modelling approach. I will discuss the tweaks made and source the original data following the completion of the assignment. Some values have been removed to simulate missing data.)

2. Assignment

Your task is to use the `BC.csv` data set to generate and present a linear regression model which performs well both as a predictive model and an explanatory model for the time value recorded for a patient, whilst using as few covariates as is necessary to achieve those ends. In generating your model you should consider, for example:

(a) Whether the data needs to be cleaned, and if so how;
(b) Whether the data needs to be transformed, and if so how;
(c) Whether there are outliers to be considered, and if so how to deal with them;
(d) How best to test the model for its usefulness in terms of both prediction and explanation;
(e) Whether a penalising function should be applied to the size of the coefficients in the model;
(f) Whether any covariates should be excluded from the model, and if so how these variables are to be identified.

You need present in full only one linear model. *However*, as has been discussed in lectures, the stepwise regression method can often lead to flawed models. Therefore, if you present a model found using stepwise regression, you need to justify why the limitations of stepwise regression have not caused an issue here.

Note that the data contains 31 covariates. The covariate "recurrence" represents the final test value a patient received. The remaining 30 covariates are medical measures that were taken at the start of the patient's initial treatment. These measures are briefly explained in the data dictionary which you can find on the ST404 Moodle. Bear in mind however that we are not expecting you to pretend to be clinicians. Identifying useful predictors as part of a strong model is your goal - trying to justify *why* a given medical measure is useful lies outside the scope of this assignment.

3. Report

You should prepare a report and an oral presentation (with accompanying slides). The report should be structured into three sections:

(1) *Findings (max 4 pages, including figures and tables).* Description of your main findings and recommendations for covariates to focus upon in future, as you would present them to the hosptial managers who decide how to assign funding to the oncologist team. These administrators will not be statisticians, so keep statistical jargon to a minimum, and use figures or tables to support your income predictions or chosen model.

The goal of this section is to provide the administrators with a good understanding of what you did, so they can have faith they can give funding to the oncologists to work on what you have discovered. The report should include:

- A bullet-point list of recommendations for which predictor variables are most important in terms of predicting/explaining the length of time a patient has attached to them.
- A discussion of how the "recurrence" predictor has been handled in your analyis, and why.

An important point to include in this section are any criticisms or limitations of the data or the analysis that you just performed. Your healthy criticism may give directions for future research, which would be very valuable to the oncologist team going forwards.

(2) *Statistical methodology (max 7 pages, including figures and tables).* Description of the methods you used. This should indicate any strategies for outlier removal, outcome/predictor transformations, variable selection strategies, analysis of the residuals, and model diagnosis. You should consider at least one selection or penalized likelihood strategy from the following list:

(a) Stepwise regression with AIC
(b) Ridge regression
(c) LASSO regression

Here you should discuss why you ended up choosing one of these approaches over the others (see part 2 for a comment on using stepwise regression), and provide any necessary evidence. A statistical explanation of how you arrived at the recommendations given in the previous section should be included here, along with any additional discussion of how the "recurrence" predictor was handled and why suggest any improvements/alternatives to your approach for future work

A major goal of this section is to give enough details so that if another statistician attempted to reproduce your results, they could do so without having to guess at any stage about what decisions you made and processes you followed - it is *not* enough to

simply include all code used in the appendix and expect someone to read through it without explanation.

(3) *Appendix (max 4 pages).* Here you should include annotated `R` code and any additional figures or results supporting statements in Sections (1)-(2) but not included there. Do not put any `R` code in Sections (1)-(2).

The oral presentation must be no more than 12 minutes long (groups with five people in it can have up to 15 minutes if they'd like), and all students in a group should spend a roughly equal amount of time speaking.

Both the report and your presentation slides should be submitted electronically in Moodle by 11am on 27/02/2019. The slides should contain a brief description of your methodology and your findings, and be as visually appealing as possible. Presentations will be held during lecture hours in weeks 9 and 10 - further details will be given nearer the time. The report is worth 60% of the final mark of Assignment 2 (see marking criteria below) and the presentation the remaining 40% (again, see below).

4. Marking Criteria (Total 100 points)

Findings

(1) Clarity and accurateness of overview of data and the description and interpretation of model;
(2) Quality and relevance of numerical and graphic output;
(3) Quality of recommendations provided;
(4) Appropriateness, clarity, and correctness of language. **[20 marks]**

Statistical Methodology

(1) Relevance and quality of numerical and graphical evidence;
(2) Soundness and justification of modelling decisions;
(3) Depth of critical evaluation of the final model;
(4) Structure and clarity, appropriate use of terminology, correctness of English. **[35 marks]**

Appendix Appropriately annotated and complete. **[5 points]**

Presentation Slides: Marked as a group.

(1) Layout, structure and visual appeal;
(2) Accuracy and relevance of content. **[20 points]**

Oral Presentation: Marked individually.

(1) Fluidity;
(2) Persuasiveness;
(3) Appropriate use of language;
(4) Response to targetted questions, where appropriate. **[20 points]**

Layout: The report should be written in a font size 11 or higher with a 1.5 spacing between the lines. Margins should be appropriate. All figures and tables should be numbered and have captions.

Penalties:

Late submission (-5% per working day)
Over page limit (-5%)
Not using prescribed layout (-5%)

The second part of this assignment and the delivery during the oral presentation will receive an individual mark. The report itself and the slides used in your presentation will receive a group mark. This group mark will be distributed across team members using the weighting algorithm described below.

Each team should decide how to distribute the group mark by allocating to each team member a share of $n \times 100\%$ where n is the number of students in the team. This will act as a weighting factor to convert the group mark into an individual mark. For example, suppose the

group mark is 70% and a team of 5 students decides to allocate 100% to each team member, then each member receives the mark of 70%. On the other hand if the team decides to allocate 108% to one team member and 98% to the other four team members, then the former receives a mark of 75.6% and the latter four team members receive the mark 68.6%. The maximum weighting factor that can be awarded is 110%., the minimum is 90%. The module leader reserves the right to moderate the weighting factors, impose equal weighting factors, and/or request further evidence.