ST404 Applied Statistical Modelling
Assignment 2. Linear modelling

1. Introduction

Assignment 2 accounts for **40%** of your final module mark. It consists of ***two*** components:

1. An individual piece of reflective writing, details of which can be found on the other assignment brief on Moodle (**5%**)
2. A piece of group work analysing a second dataset. (**35%**)

This document outlines the second of these two components.

2. Group Task

For this assignment you will continue to analyze the US Crime data. In this fictitious scenario, your group is working in a consulting role for a US congressional committee. Your task is to analyze the US Crime data to reveal patterns in the crime rate. The committee is interested in a number of questions, such as:

- What are the major determinants of high crime rates?
- Are there areas of the US with unusually low or high crime rates that do not conform to the general pattern?
- Are the causes of violent and non-violent crime similar, or are there important differences?

The expectation when working with policy makers is to provide information that allows governments to formulate their own policy, but **not** to make policy recommendations yourself. So, for example, it is acceptable to describe the major determinants of variation in crime rate between counties, and to describe clear differences where these occur. But it is not expected that you would describe to the client what the government should do to decrease crime rates.

Your model choice strategy combines predictive and explanatory goals. Your model should have good predictive power in order to support your conclusions about the determinants of crime rates in the USA. However, it should not be a “black box” that gives good predictions but gives no understanding of why crime rates are high (or low) in certain counties. Hence your task is to use these data to create a model which you believe offers an appropriate balance between:

a) predictive power;
b) explanatory power;
c) simplicity (the clients are not experts in statistics, they need to understand it and be able to use it).

Your task is to use the *US Crime* data set to generate and present a linear regression model which performs well both as a predictive model and an explanatory model for the crime rate in a given county. You will need to generate separate models for violent and non-violent crime.

In generating your model you should consider the lessons learned from the Exploratory Data Analysis in Assignment 1. Since you are working in different groups, your first task will be to combine the ideas you have from your first groups. In your report, you can refer to your EDA findings without having to repeat them. We note that some groups included regression diagnostics from simple linear models, which went beyond the requirements of assignment 1. Model diagnostics can and should be repeated as you build your models.

Factors to consider are:

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 penalty 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 model for each of the two outcome variables. **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.

3. Required Submission Format

You should prepare a report and a poster presentation.

3.1 Report

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 predictors to focus upon in future, as you would present them to the client. The client is not a statistician, so keep statistical jargon to a minimum, and use figures or tables to support your predictions or chosen model.

The goal of this section is to provide the client with a good understanding of what you did, so they can take an evidence based approach to future policy making. Your report should

reach clear conclusions about what are the major determinants of violent and non-violent crime.

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 implementations, which would be very valuable to the client 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 and/or BIC
   b) Ridge regression
   c) LASSO regression

   Here you should discuss why you ended up choosing one of these approaches over the others (see above 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 limitations of the data and suggestions of 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).

3.2 Poster Presentation

In addition to the report, you should prepare a poster. Posters are a standard way for early career scientists to communicate the findings of their research at conferences, especially for work in progress. In a poster session, conference attendees are free to come and go, to read the contents of the poster and discuss them with the authors.

Normally we organize a live poster presentation session, but due to the COVID restrictions this will not be possible this year. Instead, you will present your poster in a similar way that you presented the slides for assignment 1.

The target audience for the poster is the same as for the report, i.e., you should aim the main messages of your poster at policy makers. However, you should be ready in the poster session to defend your chosen approach to a fellow statistician.

The poster should be of A1 size (594 × 841 mm). It should contain a brief description of your methodology and findings which should be visually appealing to a non-technical audience.

4. Marking Criteria

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.

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.

Appendix

Appropriately annotated and complete.

Poster: Marked as a group.

1) Layout, structure and visual appeal;
2) Accuracy and relevance of content.

Oral Presentation: Marked individually.

1) Fluidity;
2) Persuasiveness;
3) Appropriate use of language;
4) Response to targeted questions, where appropriate.

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. Do not include raw output from R in your report.

Penalties:

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

Peer review and weighting

The report itself and the poster will receive a group mark. This group mark will be distributed across team members using the weighting algorithm described below. The delivery during the oral presentation will receive an individual mark, as with assignment 1.

Weighting for Group Members

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.