G12SMM Statistical Models and Methods

Linear Models, Assessed Coursework 2 — 2016/2017

Please hand in your work by 3.00pm on Wednesday 5 April to a student services centre. Please remember to complete and attach a coursework submission sheet to your work. Your solutions should contain all relevant plots and `R` output needed to justify your answers/arguments, together with appropriate discussion, but please do not include pages of irrelevant plots/output which you do not discuss. The easiest way to include `R` output is to use R Markdown to produce your solutions, but you do not have to do so. You do not need to include your `R` code, though you can include it if you wish. If you are using R Markdown, and do not wish to include your `R` code, then you can suppress the `R` code using the `echo = FALSE` argument, i.e. enclose the code in an {`r, echo=FALSE`} environment in the Markdown file.

There will be a Moodle forum specifically for answering queries about the coursework, so you may post questions and I will answer them there so that everyone receives the same assistance. Please be careful to not inadvertently give away parts of your answer if you do post a question. Note that as this is assessed work, I can only answer queries relating to clarification, and I will only answer queries via the forum so that everyone can see my responses.

Unauthorised late submission will be penalised by 5% of the full mark per day. Work submitted more than one week late will receive zero marks. You are reminded to familiarise yourself with the guidelines concerning plagiarism in assessed coursework (see the student handbook), and note that this applies equally to computer code as it does to written work.

You can answer the questions one-by-one as you would normally answer an exercise sheet, and in particular there is no need to produce a formal "report". However, you should explain carefully your conclusions about the practical questions being asked, and how they follow from the models you fit, and include all relevant output to support your conclusions.

The Data

You work as a risk analyst for a bank, making lending decisions on loan applications. Data are available on the credit score of 1000 individuals, together with the values of 20 explanatory variables for each individual. Interest lies in modelling credit score using these explanatory variables, where the credit score is a measure of credit worthiness based on an individual's credit history. These models can then be used to assign a credit score to a new loan applicant, to form the basis of a lending decision.
The data, which come in two parts, are available on Moodle. They are

`Train.txt`	Training data, which will be used to build models.
`Test.txt`	Test data, which will be used to make and assess predictions using models built on the training data.

They can be read into R (after saving the file in your working directory) using

```
Train = read.table("Train.txt",header = T)
Test  = read.table("Test.txt",header = T)
```

A description of the variables can be found at the end of this document.
After reading in the data, first check that R is treating the variables as desired (see variable description at the end of this document), and change if necessary.

The Tasks

(a) **Using only the TRAINING data**, investigate models to explain the relationship between `CreditScore` and the other variables. That is, `CreditScore` (or transformations of it) is to be the response variable, and all other variables are potential explanatory variables. **[25]**

(b) Use your chosen "best" fitted model from (a) to predict the responses (credit scores) for the individuals in the **TEST** data set. That is, for each of the individuals in the Test data, use the values of the explanatory variables as input to your fitted model equation from (a) to obtain fitted/predicted responses (credit scores) for these individuals. Compare your predicted responses with the known observed responses from the observations in the Test data, using suitable plots/numerical summaries. What is the mean-squared error of the predictions, and how does this compare to the mean-squared error of the "full" model (i.e. the model with all 20 explanatory variables included additively)?
NOTE: you should **NOT** fit a new model to the TEST data. The idea is to use the training data to build a model which is useful for predicting new responses (which in real world applications would be unknown) — but when evaluating models we can use the known responses in test data to evaluate our model's predictive power. **[15]**

(c) It is now of interest to classify the individuals from the TEST set into two groups ("bad" risks and "good" risks), using your model's predicted credit scores. The bank considers "bad" risks to be those with a credit score below 500. Use the following classification rule, based on the **predicted** credit scores from part (b), to classify the individuals in the TEST set as bad/good risks.

```
Risk = 1   if predicted credit score is less than 500
Risk = 0   otherwise,
```

i.e. 0 corresponds to being classified as "good" risk (won't default) and 1 corresponds to being classified as "bad" risk (will default).
Then, use the true credit scores for the individuals in the TEST set to determine their true risk class. What proportion of the individuals are correctly classified by your model?

[10]

Here are a couple of useful R commands you might wish to investigate for part (c). However, there are also many other ways to do the required computations.

If `x` is a vector, then the command `ifelse(x<10,1,0)` sets the elements of `x` to be `1` if they are less than `10`, `0` otherwise.
If `x1` and `x2` are vectors of the same length, then `sum(x1 == x2)` will count the number of elements of `x1` and `x2` which are the same.

Notes

- As with any analysis, the first step should be to do some exploratory analysis using any relevant plots and numerical summaries.
- For the model fitting, you can use any of the techniques we have covered this semester to investigate potential models — the automated methods of Chapter 6/Case Study 9 will be useful to avoid manually checking lots of models, but you can still use, for example, F-tests to compare two potential (nested) models, e.g. if two different automated methods/criteria give different answers, or for checking significance of a single additional variable. The task is deliberately open-ended: as this is a realistic situation with real data, there is not necessarily one single correct answer, and different selection methods may suggest different "best" models — this is normal. Your job is to investigate potential models, and provide a summary of what they tell us about the problem we are trying to solve. The important point is that you correctly use the relevant techniques in a logical and principled manner, and provide a concise but insightful summary of your findings and reasoning. (Note however that you do not have to produce a report in a formal "report" format.)
- You should pay attention as to whether the model assumptions are being met, for example using suitable diagnostic plots, and consider any transformations of the numerical variables if appropriate. Also consider whether your conclusions depend on a few outlying or influential points.
- You should (briefly and concisely) interpret your model(s) and consider whether they make sense in the context of the problem, for example via interpreting the fitted parameters.
- You do not need to include all your R output, as you will generate lots of output when experimenting with the model fitting. However, you should include the output which is relevant to the arguments that you make when describing the logical developments of your model fitting, and any diagnostic plots which justify changes you make in order to meet the modelling assumptions. Finally, at all stages please remember to explain your reasoning and describe (concisely but accurately) the action you take and why, along with the relevant output.

An explanation of the variables in each data set is given below, with (F) signifying factor and (C) signifying a continuous/numerical quantity.

`Status (F)`. Status of current account balance. Levels: "Negative","Small", "Large","None" (no current account or unknown).

`Duration (C)`. Duration of requested loan in months.

`History (F)`. Status of previous loan history. Levels: "A" (none, or all paid back in full), "B" (all at this bank paid in full), "C" (ongoing loans fully paid so far), "D" (late payments in past), "E" (critical delays/defaults in past).

`Purpose (F)`. Purpose of loan. Levels: "NewCar","UsedCar","Other","Furniture","Television", "Domestic","Repairs","Education","Training","Business".

`Amount (C)`. Amount requested in Euros.

`Savings (F)`. Balance of savings account. Levels: "Low","Medium","Large","VeryLarge","Unknown".

`Employment (F)`. Time in current employment. Levels: "Unemployed","Short","Medium", "Long","VeryLong".

`Disposable (C)`. The monthly repayment installments as a percentage of annual disposable income.

`Personal (F)`. Personal status. Levels: "M:DivSepMar" (Male, Divorced/Separated/Married), "F:DivSepMar" (Female, Divorced/Separated/Married) ,"M:Single" (Male, Single), "F:Single" (Female, Single).

`OtherParties (F)`. Other parties with an interest. Levels: "None","Coapp" (another co-applicant), "Guarantor" (a guarantor).

`Residence (C)`. Full years in current residence.

`Property (F)`. Most valuable significant asset. Levels: "House","Savings","Car","None".

`Age (C)`. Age of applicant.

`Plans (F)`. Other current loan plans. Levels: "Bank","Stores","None".

`Housing (F)`. Ownership status of accommodation. Levels: "Rent","Own","RentFree".

`Existing (C)`. Number of existing credits at this bank.

`Job (F)`. Level of current job. Levels: "Unemployed","Unskilled","Skilled","Management:Self".

`Dependants (C)`. Number of dependants.

`Telephone (F)`. Does the applicant have a registered phone in their name? Levels: "No","Yes".

`Foreign (F)`. Is the applicant a foreign worker? Levels: "Yes","No".

`CreditScore (C)`. Credit score of the applicant (higher is better).