MATH2011 Statistical Models and Methods

Linear Models, Assessed Coursework — 2020/2021

The deadline for the work is **Thursday 13th May, 15:00**. The grace period means that you may submit up until **Thursday 20th May, 15:00** without any penalty, but submissions after this receive a mark of zero unless you have been granted extra time due to extenuating circumstances.

Since this work is assessed, your submission must be entirely your own work. Please make sure you are fully aware of the University's policy on academic misconduct.

The submission should be produced and uploaded electronically via the submission box on Moodle. Your work should contain 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 — this will detract from the analysis. 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. Please post any academic queries about the work on the Piazza forum, so that everyone receives the same assistance. If you have any issues relating to your own personal circumstances, then please email me.

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. The credit score has been determined independently by experts, without reference to the covariates. Interest lies in using these data to build models to explore any association between credit score and the available explanatory variables, and to predict credit scores based on these variables. The models could then be used to assign a credit score to a new loan applicant, to form the basis of a lending decision given the applicant's data.
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

See also the "Notes" section below for further guidance.

(a) **Using only the TRAINING data**, investigate models to explain the relationship between `CreditScore` and the other variables. That is, `CreditScore` is to be the response variable, and all other variables are potential explanatory variables. Briefly interpret your models in terms of the association between the explanatory variables and credit score. **[25]**

(b) Use your chosen "best" model(s) from (a) to predict the responses (credit scores) for the individuals in the **TEST** data set. 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)? **[10]**

NOTE: you should **NOT** fit a new model to the TEST data. The idea is to use the fitted model(s) from part (a) to make predictions, then see how they compare to the true responses.

(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?

[5]

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

- Part (a) covers all the modelling part of the analysis. This means you should cover
  - Exploratory analysis
  - Model fitting/selection
  - Interpretation
  - Diagnostic checks
- Parts (b) and (c) should be conducted after all model fitting is completed. These measures (mean-squared error/classification rate) are not intended to inform the model selection process. The objective in these parts is not to try and find the best possible answer in

terms of these measures, it is simply to use the model(s) from (a) to see how well they actually perform. Do not worry if it turns out other models exist which do better! Simply report what you find, and briefly discuss it.

- For the model fitting in (a), you can use any of the techniques we have covered this semester to investigate potential models — the automated methods of Chapter 6/Case Study 7 will be useful to avoid manually checking lots of models. (You will likely find that best subsets regression takes an impractical amount of time with this number of covariates.) However, you could still use hypothesis tests too, 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.
- You do not need to (and should not) 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. 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.

The Variables

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 (considered less risky).