Project 2021-22: MSG500-MVE190 Linear Statistical Models

Deadline: **23:59 on Sunday 9 January**

Instructions

Work in groups of two for writing both (i) the summary of the mini-analyses based on the AirBnB data and (ii) the analyses on cars data and apprentices migration. You should submit all material as a single pdf document, except for the code, to be attached as specified below. The report should be typed (not handwritten).
Write a clear report, **in English**, presenting your approach to the assignments, discussing the methods and results. It should be noted that often there isn't a single "right" answer, but you should motivate your approach in a critical way, by considering the several topics we have covered over the course. Be selective and stay within the allowed number of words.
In addition to the text, you should use figures and tables, with explanatory captions. But do not exaggerate, be selective and go for **quality not quantity**.
Key information may be better summarized in tables than by including the R printouts. Do not paste your R code *in* the report, but you can include some of the R output provided this is nicely formatted and readable. The actual R code produced for your analyses **must** be submitted as detailed below.

Projects will be checked against plagiarism using Urkund.

Form groups

Unless you have so far worked alone, form groups of two students on Canvas, by going to "People→Groups→project groups".
Do not wait the deadline to approach to test this. In practice, first consult with your group partner, so that you do not choose different groups.

Submission

Go to Canvas, "Assignments→project".
Only one person in each group should upload the material on Canvas. You should upload **both the following: a)** a SINGLE pdf document (not .zip) containing both a summary of the mini-analyses + further analyses, and **b)** all the R files you produced for your analyses (do not upload zipped files). In case your .R files are rejected, just modify the file extension from .R to .r. **Deadline for submission: 23:59 on Sunday 9 January**.

1 Summary of mini-analyses

Size limitations: max 2000 words (or max 12000 characters with spaces, whichever is the largest) excluding tables/plots and coverpage. You can use as many tables and figures as necessary, but do not exaggerate. Only put key results. Do not go for quantity. [*Why such limitations?* Well, these are things we have seen and discussed together, accounting for 10% only of the final grade].

Briefly describe the problem and the goals. Then summarize the main results (you can disregard the, often many, residual plots and attempts at finding outliers, which you can instead consider for the other project work.) What were the peculiarities of the data? Anything interesting you learned and that you would communicate to someone interested in analyzing these kind of data? What was

challenging? What were the main findings?

Notice: construct a general summary of your data analysis and modelling with this data set, instead of writing "this is what I did in mini1", "this is what I did in mini2" etc. Try to write a single coherent description of your analyses attempts. Also because, at this point, you may have learned something that make you criticize what you did several weeks ago.

2 Project report of further analyses

[*This accounts for 40% of the final grade*].

Size limitations for the COMBINED cars data and apprentices migration, excluding appendices: max 4100 words (or max 25000 characters with spaces, whichever is the largest) excluding tables/plots and coverpage. You can use as many tables and figures as necessary, but do not exaggerate. Be selective, only put key results. Do not go for quantity. You should put some of the material that is useful for model building, but that is not immediately interesting when communicating the main results, in some appendices (these appendices do not have size limitations). For example, many plots or preparatory analyses that are made during model building can go in appendix (say attempts at transforming variables, outliers detection, leverage values, residual plots).
In the main text you should introduce the problem, discuss if your data have some interesting feature, show your strategy for modelling and the steps you undertake for model construction, and write the main findings by commenting and interpreting the results. This is a report that should be *readable and to the point*, and this is why you should separate the relevant but less-interesting preparatory analyses, from the ones that carry the main messages. In other words, the main document should not illustrate all analyses you ever attempted (also the appendices do not need to report all your attempts). Clearly specify your goals, models and methods that you are using and, most importantly, do interpret the results. **Do interpret the values of the parameter estimates, as long as this is possible (at least for the most interesting findings, not necessarily all estimates), or the effect sizes when interpretation is more difficult**. Notice the **checklist** at the end of this document.

2.1 Cars data

We are considering a dataset providing info on a number of cars that were on the market in the 80's. The list of variables is given below. It is of interest to consider the car price as response variable in multivariate linear regression (MLR).

Data is available on the course webpage. Variables description is at the end of this document.

Points to remember for MLR: Consider the issues of possible multicollinearity among numerical predictors only (checking multicollinearity when including categorical covariates is tricky); possible variables transformation; are the outliers affecting the fit? (outliers are with respect to the chosen model. If you change model, you might have new outliers or previous outliers might no longer be outliers under a different model). Check for addition/removal of covariates using appropriate tests and other procedures. How is the quality of your fit? Is your model good at predicting "unseen observations"? How good? Do you select different models depending on which procedure you use? Which one do you consider as your final model and why? Once you have selected a final model, fit it again on the full dataset and obtain the $\hat{Y}$, then plot the observed Y vs $\hat{Y}$ and comment. Comment also on residuals and other diagnostics for the model you picked. For the picked model, what is the coefficients interpretation when you increase a covariate value (recall the different

interpretations depending on variable transformation). Does the change in the expected response vary significantly? But is the uncertainty around the estimates large (eg via confidence intervals)? Comment on the interpretation of such variability.

Remember **interpretation of results is important!**
Finally, summarise take-home messages.

In this dataset there are many categorical covariates. Do not use too many of them simultaneously with regsubsets, or this may cause possible problems due to the limited number of model possibilities when using the force.in setting. You may first select the most relevant 1-3 categorical covariates as found by other means, then include those into regsubsets with numerical covariates.

2.2 GLMs application: apprentices migration in Scotland

This dataset considers the number of individuals in Scotland that moved (migrated) to Edinburgh from several Scottish counties, to be apprentices. Data refer to years 1775 to 1799. It is a reasonably small dataset, where things to be done will be a bit more explicit, since we haven't worked that much with GLMs.
For each of 33 counties, we have the number of migrants that moved to Edinburgh; the size of the counties population; the position of each county with respect to Edinburgh; the level of urbanization in the county; the county distance from Edinburgh.
We are interested in considering the number of apprentices as response variable in either Poisson regression or negative binomial regression. You will notice the population and distance variables are very skewed. Even though this is not linear regression, it is usually better to have covariates that do not vary on widely different scales[1], so you may want to apply some transformation to the variables population and distance.

- run Poisson regression with response `apprentices` and interpret the results. Notice, for the sake of interest of what you will do later: (i) check if the assumption for the Poisson distribution is satisfied; (ii) even if it is not, proceed with Poisson regression as it will be interesting to look at comparing these results with negative binomial regression.
- computer appropriate residuals for this type of regression. What do you deduce?
- now consider negative binomial regression. Check again the residuals and compare with those from Poisson regression. Discuss.
- Produce a likelihood ratio test for testing the Poisson vs the negative binomial regression model. Discuss.
- Did you find any influential observation using diagnostics?
- So what is a summary of conclusions about the chosen model? Which covariates are relevant and what do these tell us regaridng their effect on the response?

3 Variables in the Cars dataset

Some of the variables are obviously categorical, others appear numerical but should be treated as categorical and therefore are explicitly denoted with "(categorical)". The definition of some variables is rather technical (e.g. boreratio, enginetype, fuelsystem) so I leave it up to your interest to find the definitions.

[1] Or otherwise this could give numerical problems when minimizing the negative loglikelihood.

Regarding baseline categories, when explicitly given please use the one indicated in the variables explanation below.

- Car_ID: Unique id of each observation;
- Symboling: an insurance risk rating between -3 and 3[2](or at least that was the range in 1980s). A value of +3 indicates that the auto is "risky", -3 that it is considered "safe".
- carName: Name of the car company (Categorical).
- fueltype: Car fuel type (Categorical, baseline: gas);
- aspiration: Aspiration used in a car, standard or turbo (Categorical, baseline: std)
- doornumber: Number of doors in a car, two or four (Categorical, baseline: four)
- carbody: body of car (Categorical, baseline: hatchback)
- drivewheel: type of drive wheel: front wheel drive (FWD), rear wheel drive (RWD), and 4WD (4 wheel drive). (Categorical, baseline: rwd)
- enginelocation: Location of car engine (Categorical, baseline: front)
- wheelbase: Weelbase of car (Numeric)
- carlength: Length of car (Numeric)
- carwidth: Width of car (Numeric)
- carheight: height of car (Numeric)
- curbweight: The weight of a car without occupants or baggage. (Numeric)
- enginetype: Type of engine. (Categorical, baseline: ohc)
- cylindernumber: number of cylinders in the car (Categorical, baseline: four)
- enginesize: Size of the engine, in cubic inches (??) (Numeric)
- fuelsystem: Fuel system of car (Categorical, baseline: 2bbl)
- boreratio: Boreratio of car (Numeric)
- stroke: Stroke or volume inside the engine (Numeric)
- compressionratio: compression ratio of car (Numeric)
- horsepower: Horsepower (Numeric)
- peakrpm: car peak rpm (rpm = revolutions per minute) (Numeric)
- citympg: Mileage in city per gallon of fuel (Numeric)
- highwaympg: Mileage on highway per gallon of fuel (Numeric)
- price: car price in US$ (Numeric)

[2]Cars are initially assigned a risk factor symbol associated with its price. Then, if it is more risky (or less), this symbol is adjusted by moving it up (or down) the scale. Actuarians call this process "symboling".

4 Variables in the dataset `apprentice`

Download the dataset `apprentice` from the course webpage. Load the dataset into R, then run the following to give names to the columns:

```
data<-setNames(apprentice,c("county","distance","apprentices","population","urbanization",
        "direction"))
```

- county: name of the county;
- distance: distance of the county in km (using some hypothetical centroid in the county) from Edinburgh;
- apprentices: number of apprentices for the given county who moved to Edinburgh.
- population: county population in thousands;
- urbanization: percentage of county urbanization (that is percentage of county population living in urban settlements);
- direction: (categorical) direction of the county compared to Edinburgh, 1=North, 2=West, 3=South (I guess East is not considered since Edinburgh is on the east coast). Use North as baseline.

Checklist

Make sure your work complies to the list below before submitting it:

a) Name and surname of all authors should appear on the coverpage.

b) pages should be numbered.

c) Briefly describe the methods used. Be brief - don't repeat what's in the notes, just the key elements.

d) Discuss your results. Results without discussion are not graded.

e) Divide the text into paragraphs and structure it with clear and suitable section headings

f) Include only the crucial plots and graphs, don't go for quantity.

g) key information may be better summarized in tables than by including the full R printouts (e.g. it may be enough to give regression coefficients and p-values without all the accompanying information provided by R).

h) Label all plots and graphs. Reference to those in the text so the report is understandable and readable.

i) ask yourself if tables/plots are readable (example, the output of the pairs() function with many variables is usually difficult to read. Perhaps report only the most relevant associations).

j) Conclusions: what is the take-home message.

k) Do not paste the code in the report. The code should be uploaded separately.