Project 2019-20: MSG500-MVE190 Linear Statistical Models

Deadline: **16:00 on Friday 10 January**

Instructions

Work in groups of two for writing both (i) the summary of the mini-analyses based on the NY AirBnB data and (ii) the analyses on car prices and infections data. 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

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, creating problems for the other students trying to log-in simultaneously and not finding available 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)** summary of mini-analyses + further analyses as a single pdf document, and **b)** the R code you produced for your analyses either as a single file `.R` file, or as a zipped folder containing all codes (in case your .R files are rejected, just modify the file extension from .R to .r). **Deadline for submission: 16:00 on Friday 10 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: try to construct a general summary of your data analysis and modelling with this data set. That is, instead of writing separately "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 previously did.

2 Project report of further analyses

[*This accounts for 40% of the final grade*].
Size limitations for the COMBINED cars data and infections analyses, 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. 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 predict the cars price using 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. Based on what is included in your final model, you may produce confidence intervals for $E(Y)$ based on some values of the covariates. And the same for prediction interval for future observations. Comment.

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. See first what happens when you have only one or two categorical covariates, then possibly enlarge if it gives no problems.
You can choose yourself the baseline categories. And again, generally, it is important that you try to interpret the parameter estimate values (at least for the most interesting findings, not necessarily all estimates).

2.2 Bird, human and equine infections

The so-called West Nile Virus (WNV) was introduced into the United States in 1999. Some north american bird is highly susceptible to WNV and it is very important to monitor birds mortality because this typically precedes human or equine WNV infection. Areas with greater human population densities are more likely to have dead birds reported (due to an increased probability of people finding them) and are more likely to have an increased number of human cases (due to a larger number of people at risk of infection). We are interested in predicting the *rate* of WVN-positive equine in a county: however since the number of horses per county is not available, we may use the number of farms as a proxy for horse density in a given county, therefore allowing us to model instead the # of WVN-positive equines per farm.

(i) following the hints above, write a Poisson regression model to fit the rate of WVN-positive equine. [tip: there is no need to use all given variables. Read the text above carefully]

(ii) check if assumptions are satisfied or not.

(iii) interpret the parameter values

(iv) construct a new variable for the observed rates of WPN-positive equine, that is the ratio `equine_cases/farms`. Then produce a scatterplot comparing such new variable versus the fitted rate of WVN-positive equine. Add a line having slope equal to 1[1]. Comment.

(v) Now use a negative binomial model, and produce the same plot as in (iv).

(vi) Using a likelihood-ratio test, to check if a negative binomial model is suggested over the Poisson one.

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.

- Car_ID: Unique id of each observation;
- Symboling: an insurance risk rating. A value of +3 indicates that the auto is risky, -3 that it is considered safe (Categorical)
- carName: Name of the car company (Categorical).
- fueltype: Car fuel type (Categorical);
- aspiration: Aspiration used in a car (Categorical)
- doornumber: Number of doors in a car (Categorical)

[1] You can use `abline(0,1)`.

- carbody: body of car (Categorical)
- drivewheel: type of drive wheel: front wheel drive (FWD), rear wheel drive (RWD), and 4WD (4 wheel drive). (Categorical)
- enginelocation: Location of car engine (Categorical)
- 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)
- cylindernumber: number of cylinders in the car (Categorical)
- enginesize: Size of the engine, in cubic inches (??) (Numeric)
- fuelsystem: Fuel system of car (Categorical)
- 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)

4 Variables in the Nile dataset

- county: county name
- bird_cases: number of birds found WPN-positive in the county
- equine_cases: number of horses found WPN-positive in the county
- farms: number of farms in the county
- area: county area (square miles)
- popul: human population size
- human_density: the ratio population/area
- birdrate: the ratio (bird_cases/population)$\times 10,000$, that is the number of positive birds per 10,000 inhabitants

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.