DATA 303/473 Assignment 2

Due 1159pm Friday 31 March

Assignment Questions

Q1.(20 marks) We'll continue to use the CarDekho data from Assignment 1. As a reminder the variables in the `cardekho2.csv` dataset are:

- `price`: Selling price in thousand Indian Rupees (INR)
- `make`: Car make grouped into eight categories: `Ford`, `Honda`, `Hyundai`, `Mahindra`, `Maruti`, `Tata`, `Toyota`, `Other`
- `kms`: Kilometres driven (x 1000)
- `fuel`: Fuel type: `Diesel` or `Petrol`
- `seller`: Seller type: `Dealer`, `Individual` or `Trustmark Dealer`
- `tx`: Transmission type: `Automatic` or `Manual`
- `owner`: Current owner is: `First`, `Second` or `Third or above` owner
- `mileage`: Fuel economy in kilometres per litre (kmpl)
- `esize`: Engine size in cubic centimetres (CC)
- `power`: Maximum engine power in brake horse power (bhp)

The residual diagnostic plot showed evidence of non-linear relationships between `price` and some predictors, non-normality and non-constant variance. To address non-constant variance, use `log(price)` as the response variable for this assignment.

a. **(3 marks)** Fit a model with `log(price)` as the response variable and include all predictors without transformations or interactions. Use the plot function to carry out residual diagnostics for your fitted model. Based on these plots, are there any observations you might consider excluding from further analysis? Explain your answer briefly.

Some data cleaning is done and a new dataset, `cardekho3.csv`, (available on Canvas) is created. Use this new dataset to answer the rest of Question 1.

b. **(3 marks)** Read in dataset `cardekho3.csv` and fit the same model as in part (a). Plot the residuals from your fitted model against each of the numerical predictors `kms`, `mileage`, `esize` and `power`. Is there an indication of a non-linear relationships with `log(price)` for any of these predictors? If so, which ones?
c. **(3 marks)** Based on the model fitted in part (b), calculate and give an interpretation for the difference in **price** for a petrol car compared to a diesel car when all other predictors are held constant?
d. **(4 marks)** Based on the dataset and model in part(b), provide two plots that give graphical evidence that a log transformation is the most appropriate transformation for `kms` in a model for `log(price)`. Explain your reasoning briefly.
e. **(3 marks)** Apply stepwise regression based on the AIC criterion for the model in part (b). Are there any predictors you would exclude from the model? Explain your answer briefly.

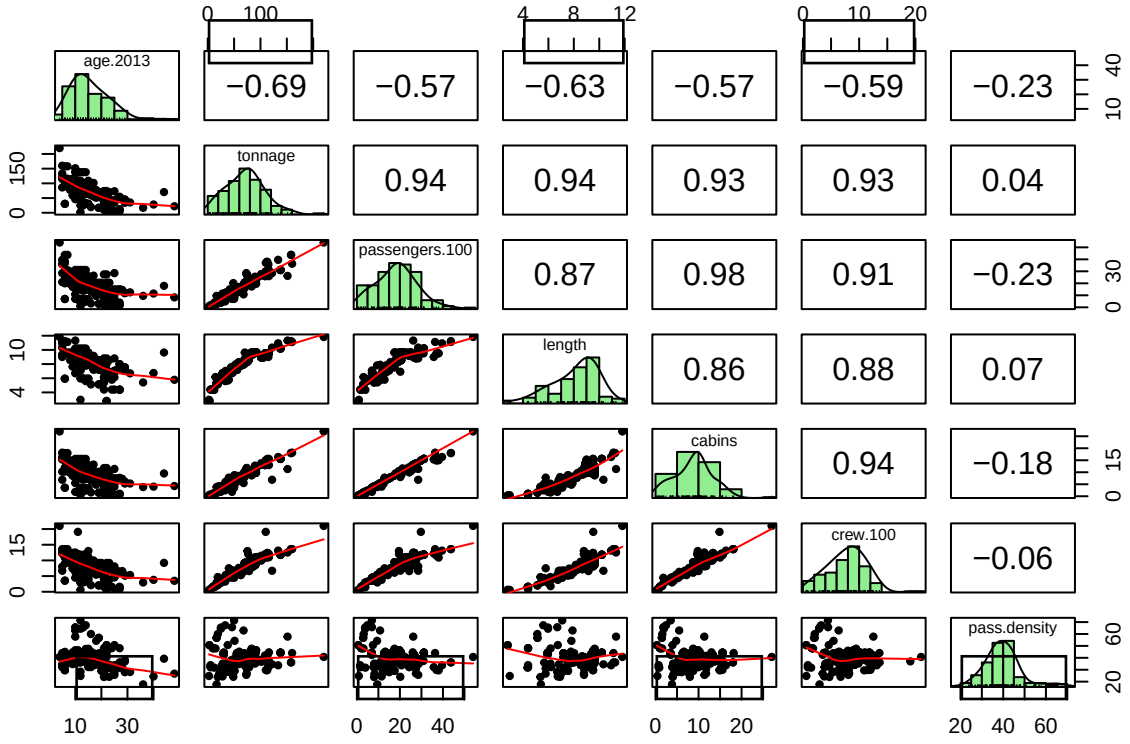
f. **(4 marks)** Fit a model you would use to investigate whether the effect of `mileage` on `log(price)` depends on the value of `tx`. Based on your model, give the change in $E(log(price))$ associated with a unit increase in `mileage` for a car with:

(i) Automatic transmission
(ii) Manual transmission.

Q2.(20 marks) Data were collected on 158 cruise ships in operation around the world in 2013. Complaints had been raised by customers about overcrowding on cruises and there was interest in investigating whether there was a trend of overcrowding on certain types of ships. As part of the investigation, a regression analysis was carried out to explore the connection between passenger density (no. of passengers per unit area) and ship characteristics. The variables in the dataset were:

- `name`: Ship Name
- `line`: Cruise Line
- `line_grp`: Cruise Line grouped
- `age.2013`: Age (as of 2013)
- `tonnage`: Weight of ship (1000s of tonnes)
- `passengers.100`: Maximum no. of passengers (100s)
- `length`: Length of ship (100s of feet)
- `cabins`: No. of passenger cabins (100s)
- `pass.density`: Passenger density (no. of passengers per square foot)
- `crew.100`: No. of crew member (100s)

The data are available in the file `cruise_ship.csv`. The dataset was imported into R and the scatterplot matrix below was obtained.



The scatterplot matrix indicates severe multicollinearity among the predictors `tonnage`, `passengers.100`, `length`, and `crews.100`. These four predictors are all related to the size of a ship, so only a subset will be used.

a. [**8 marks**] Fit a model for `pass.density` using the predictors `line_grp`, `age.2013`, `passengers.100` and `length`. Using residual diagnostic checks, determine whether any transformations of the predictors or response variable are necessary. Explain your answer, including identification of which predictors you may need to transform. Provide output of any graphical checks or hypothesis tests you perform.

For the rest of the question use `log(pass.density)` as the response variable.

b. [**3 marks**] Fit a model with `log(pass.density)` as the response variable including all the predictors in part (a) without any transformations. Apply stepwise regression based on the BIC criterion. Are there any predictors you would exclude from the model? Explain your answer briefly.
c. [**3 marks**] Fit a GAM for `log(pass.density)` and smooth terms for each of the predictors `age.2013`, `passengers.100` and `length`. Comment on the non-linearity and significance of smooth terms.
d. [**2 marks**] Is there evidence that more basis functions are required for any of the smooth terms? Explain your answer briefly.
e. [**3 marks**] Use the `gam()` function to fit a model for `log(pass.density)` with linear terms for all three predictors. Calculate BIC for this model and for the model with smooth terms in part (c). Print the results in a table and state which of the models is preferred. Explain your answer briefly.
f. [**1 mark**] Explain briefly why it is valid to make the comparison in part (f) using BIC.

Assignment total: 40 marks

‘