DATA 303/473 Assignment 2

Due 1159pm Thursday 31 March

Instructions

- Prepare your assignment using Rmarkdown
- Submit your solutions in two files: an Rmarkdown file named e.g. `a1.Rmd` and the PDF file named `a1.pdf` that results from knitting the Rmd file. The Rmarkdown file (`tut1sol.Rmd`) used to create the Tutorial 1 Solutions is provided in the Tutorial Solutions section of Blackboard as an example for you to follow should you wish.
- The YAML header of your Rmarkdown file must contain your name and ID number in the author field, and should have the output format set to `pdf_document`. For example:

```
---
title: "DATA 303/473 Assignment 2"
author: "Nokuthaba Sibanda, 301111111"
date: "Due: 31 March 2022"
output: pdf_document
---
```

- While you are developing your code you may find it easiest to have the output set to `html_document`, but change it to `pdf_document` when you submit.
- In your submission, embed any executable R code in code chunks, and make sure both the R code and the output is displayed correctly when you knit the document.
- If there are any R code errors, then the Rmarkdown file will not knit, and no output will be created at all. So if you can't get your code to work, but want to show your attempted code, then put `error=TRUE` in the header of the R code chunk that is failing.

~~~
```{r, error=TRUE}
your imperfect R code
```
~~~

- **You will receive an email confirming your submission. Check the email to be sure it shows both the Rmd and PDF files have been submitted.**
- Title each question answer with its question numbers as `Q1.`, `Q2`,... instead of `1.`,`2.`,....
- Where you are asked to perform a hypothesis test, state the hypotheses being tested and give the test statistic, p-value and conclusion.

Assignment Questions

Q1.(35 marks) In a 2015 article comparing technological advancement of hybrid electric vehicles (HEV) in different market segments, authors Lim et al. collected data on prices and other features for 154 HEV models (*Lim et al. 2015. Technological Forecasting and Social Change. vol 97, pages 140-153*). We will use regression analysis to explore the factors that influence price. The dataset is in the file `hybrid_reg.csv` and contains the following variables:

- `carid`: Vehicle ID
- `vehicle`: Make of vehicle
- `year`: Model year
- `msrp`: Manufacturer's suggested retail price in 2013 (US dollars).
- `accelrate`: Acceleration rate in km/hour/second
- `mpg`: Fuel economy in miles/gallon
- `mpgmpge:` Max of mpg and mpge (mpge is miles per gallon equivalent for plug-in HEVs to take into account the all electric range, with mpge = $\frac{33.7*driverange}{batterycapacity}$.
- `carclass`: Model class. C = Compact, M = Midsize, TS = 2 Seater, L = Large, PT = Pickup Truck, MV = Minivan, SUV = Sport Utility Vehicle
- `carclass_id`: Index representing model class

The variables `carid` and `vehicle` are vehicle identifiers and will not be used in the analysis. Likewise `carclass_id` will not be used as it is a numerical form of the variable `carclass` and does not provide any additional information.

a. **(3 marks)** Read the dataset into R. Prepare the data for analysis by adding the new variables below to the dataset. Give the number of observations in each year group of the new variable `yr_group`.:

- `yr_group`: group `year` as follows "1997-2004", "2005-2008", "2009-2011", "2012-2013".
- `msrp.1000`: convert `msrp` from US$ to US$1000 by dividing `msrp` by 1000.

b. **(3 marks)** Use the `ggplot2` package to plot `msrp.1000` against each of the predictor variables, `yr_group`, `accelrate`, `mpg`, `mpgmpge` and `carclass`. Are there strong indications of non-linear relationships with any of the numerical predictors? If so, which ones?

c. **(3 marks)** Create pairwise scatterplots of the numerical predictors. Is there any indication of potential multicollinearity among these predictors?

d. **(4 marks)** Fit a linear model with all predictors (`yr_group`, `accelrate`, `mpg`, `mpgmpge` and `carclass`) included in the model. Calculate the VIF statistic for the predictors. To check for evidence of multicollinearity we will use a different threshold defined by

$$VIF_{model} = \frac{1}{1 - R^2_{model}},$$

where R^2_{model} is the R^2 value for the model that includes all predictors. Using this threshold identifies predictors that have stronger relationships with other predictors than the response variable has. It is a more stringent way of identifying multicollinearity. If $GVIF^{(1/(2\times Df))} > VIF_{model}$, then this is evidence of severe multicollinearity. Calculate VIF_{model} for your fitted model. Is there evidence of severe multicollinearity? Are you surprised by the result?

e. **(3 marks)** Fit a generalised additive model to the data including all predictors, using a smooth spline for each numerical predictor. Present the RSE, R^2 and adjusted R^2 values in a table.

f. **(3 marks)** Print the results for the significance of smooth terms in a table. Which of the numerical predictors have a significant non-linear effect on `msrp.1000`? Justify your answer briefly.

g. **(4 marks)** Perform a diagnostic check of regression assumptions and adequacy of basis functions for the model you fitted in part (e). What conclusions do you draw from your results? (Note: ensure your diagnostic plots fit on a single page).

h. **(4 marks)** Calculate and print a table of AIC values for the model in part (e) (Model 1) and each of the following models:

- Model 2: excludes `mpg` only from Model 1
- Model 3: excludes `mpgmpge` only from Model 1
- Model 4: excludes `mpg` and `mpgmpge` from Model 1

i. [**3 marks**] What do your results in part (h) indicate about whether both `mpg` and `mpgmpge` should be included in the model? Explain your answer briefly. What regression pitfall does this point to?

j. [**2 marks**] Are you surprised by your conclusions in part (i) given your findings in part (d)? Explain your answer briefly.

k. [**3 marks**] Calculate and print a table of BIC values for Models 1 to 4. Based on these results, which model would you choose as your preferred model? Explain your answer briefly.

2. **Q2. (5 marks)** Suppose we have a data set with five predictors:

- $X_1 =$`GPA`
- $X_2 =$`IQ`
- $X_3 =$`Gender`(0=female, 1=male)
- $X_4 =$Interaction between `GPA` and `IQ`
- $X_5 =$Interaction between `GPA` and `Gender`.

The response variable, Y, is starting salary after graduation (in thousands of dollars). Suppose we get the following regression coefficient estimates:

$$\hat{\beta}_0 = 5, \quad \hat{\beta}_1 = 8, \quad \hat{\beta}_2 = 0.2, \quad \hat{\beta}_3 = 10,$$
$$\hat{\beta}_4 = 0.05, \quad \hat{\beta}_5 = 2$$

a. **(1 mark)** Write down the estimated model equation in terms of $\hat{Y}$, X_1, X_2 and X_3.

b. **(3 marks)** Which one of the following statements is correct and why? Show any working you do.

   i. For a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

   ii. For a fixed value of IQ and GPA, females earn more on average than males.

   iii. The difference in expected salary between males and females increases as GPA increases.

   iv. An increase in `IQ` by one point is associated with a reduction in expected salary, provided GPA is high enough.

c. **(1 mark)** True or False: Since the coefficient for the GPA:IQ interaction term is very small, there is very little evidence of an interaction effect. Justify your answer.

Assignment total: 40 marks

‘