MSBA7002 Business Statistics - HW 1

Name: ________________________ Student ID: ____________________________

28 October 2023

Overview / Instructions

This homework will be *due on 11 November 2023 by 11:55 PM* via Moodle.

You are required to submit 1) original R Markdown file and 2) knitted HTML or PDF file. Please provide comments for R code wherever you see appropriate. Nice formatting of the assignment will have extra points.

In general, be as concise as possible while giving a fully complete answer. All necessary data are available in Moodle.

Remember that the Class Policy strictly applies to homework. You are encouraged to discuss with fellow students. However, each student has to know how to answer the questions on her/his own. Note that the final exam is individually based.

Question 0

Review the lectures.

Question 1: Manager Rating (Bootcamp Lecture)

We discussed this example in class using the following dummy variable

Origin[Internal]=1, if Origin=``Internal”; =0, otherwise,

and considered the following interaction model:

$$Rating = \beta_0 + \beta_1 Origin[Internal] + \beta_2 Salary + \beta_3 Origin[Internal] * Salary + \varepsilon.$$

Now define another dummy variable

Origin[Internal]=1, if Origin=``Internal”; =-1, otherwise,

And consider the following model

$$Rating = \alpha_0 + \alpha_1 Origin[Internal] + \alpha_2 Salary + \alpha_3 Origin[Internal] * Salary + \varepsilon.$$

Please derive the relationships between $\{\alpha_0, \alpha_1, \alpha_2, \alpha_3\}$ and $\{\beta_0, \beta_1, \beta_2, \beta_3\}$.

Question 2: Production Time Run

ProdTime.dat contains information about 20 production runs supervised by each of three managers. Each observation gives the time (in minutes) to complete the task, Time for Run, as well as the number of units produced, Run Size, and the manager involved, Manager.

Which manager performs the best?

Question 3: Auto Data from ISLR

The original data contains 408 observations about cars. It has some similarity as the data CARS that we used in our lectures. To get the data, first install the package ISLR. The data Auto should be loaded automatically. We use this case to go through methods learnt so far.

You can access the necessary data with the following code:

~~~~
```{r, eval = F}

# check if you have ISLR package, if not, install it

if(!requireNamespace('ISLR')) install.packages('ISLR')

auto_data <- ISLR::Auto

```
~~~~

Get familiar with the data first. You can use `?ISLR::Auto` to view a description of the data.

Q3.1

Explore the data, with particular focus on pairwise plots and summary statistics. Briefly summarize your findings and any peculiarities in the data.

Q3.2

What effect does time have on MPG?

i. Start with a simple regression of mpg vs. year and report R's `summary` output. Is year a significant variable at the .05 level? State what effect year has on mpg, if any, according to this model.

ii. Add horsepower on top of the variable year. Is year still a significant variable at the .05 level? Give a precise interpretation of the year effect found here. Include diagnostic plots with particular focus on the model residuals and diagnoses.

iii. The two 95% CI's for the coefficient of year differ among i) and ii). How would you explain the difference to a non-statistician?

iv. Do a model with interaction by fitting `lm(mpg ~ year * horsepower)`. Is the interaction effect significant at .05 level? Explain the year effect (if any).

Q3.3

Note that the same variable can play different roles! Take a quick look at the variable `cylinders`, try to use this variable in the following analyses wisely. We all agree that larger number of cylinder will lower mpg. However, we can interpret `cylinders` as either a continuous (numeric) variable or a categorical variable.

i. Fit a model, that treats `cylinders` as a continuous/numeric variable: `lm(mpg ~ horsepower + cylinders, ISLR::Auto)`. Is `cylinders` significant at the 0.01 level? What effect does `cylinders` play in this model?
ii. Fit a model that treats `cylinders` as a categorical/factor variable: `lm(mpg ~ horsepower + as.factor(cylinders), ISLR::Auto)`. Is `cylinders` significant at the .01 level? What is the effect of `cylinders` in this model?
iii. What are the fundamental differences between treating `cylinders` as a numeric or a factor? Use `anova(fit1, fit2)` to help gauge the effect. Explain their difference.

Question 4: Crime Data

We use the crime data at Florida and California to study the prediction of the number of violent crimes (per population). Use the following code to load data.

```
crime <- read.csv("CrimeData_sub.csv", stringsAsFactors = F, na.strings = c("?"))
crime <- na.omit(crime)
```

Our goal is to find the factors/variables which relate to violent crime. This variable is included in crime as crime$violentcrimes.perpop.

Q4.1

Divide your data into 80% training and 20% testing. Run the ordinary least square regression with all the variables and with the training data. Get RMSE and R2 for both the training and testing data and see if there is a difference.

Q4.2

Use LASSO to choose a reasonable, small model, based on the training data you created. Re-fit an OLS model with the variables obtained. The final model should only include variables with p-values < 0.05. Note: you may choose to use lambda 1se or lambda min to answer the following questions where apply.

i. What is the model reported by LASSO? Use 5-fold cross-validation to select the tuning parameter.
ii. What is the model after refitting OLS with the selected variables? What are RMSE

and R2 for the training and testing data? Compare them with results in Q4.2.

iii. What is your final model, after excluding high p-value variables? You will need to use model selection method to obtain this final model. Make it clear what criterion/criteria you have used and justify why they are appropriate.

iv. Try Ridge regression with 5-fold CV to select the tuning parameter. Compare its training and testing RMSE and R2 with the previous models.