DATA 303/473 Assignment 3

Name:xxxxxxxx, ID:xxxxxxxx

Due: 12 May 2022

Intructions

- Prepare your assignment using Rmarkdown
- Submit your solutions in two files: an Rmarkdown file named `assignment3.Rmd` and the PDF file named `assignment3.pdf` that results from knitting the Rmd file.
- 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 Assignment 3"
author: "Nokuthaba Sibanda, 301111111"
date: "Due: 12 May 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
```
```

- Where appropriate, make sure you include your comments in the output within the Rmarkdown document.
- **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.**

Assignment Questions

Q1

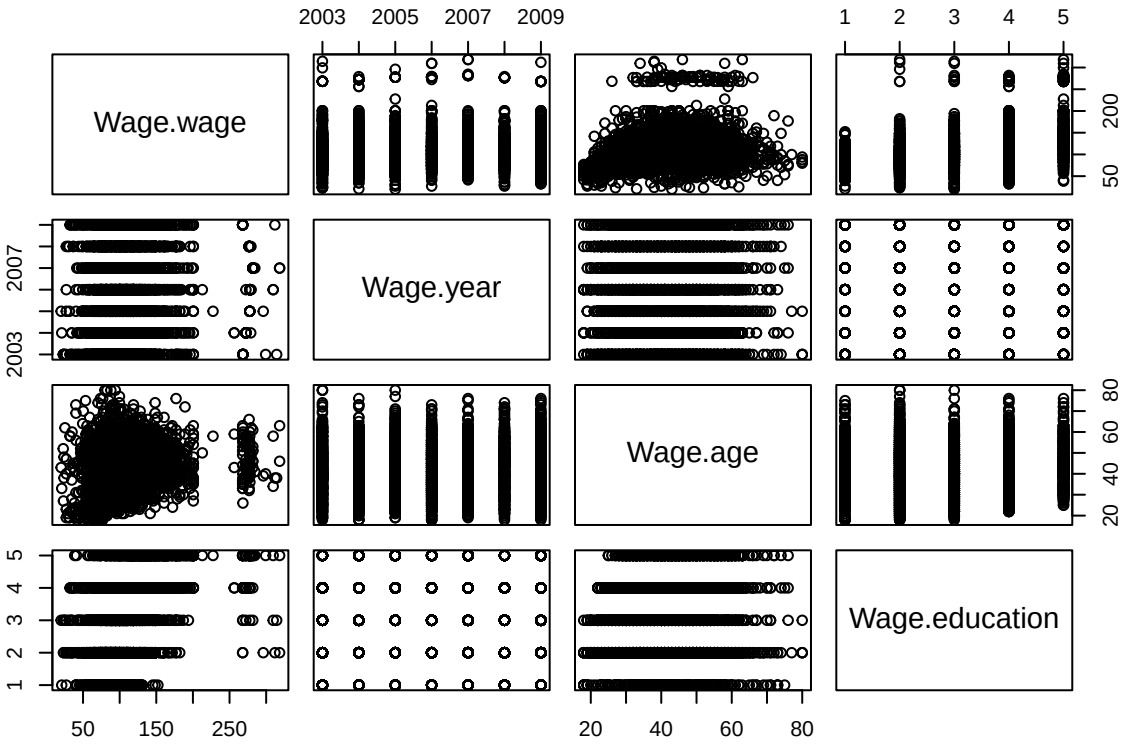
We use `Wage` data set which is in the library `ISLR2`. The `Wage` data set contains the following variables.

```
library(ISLR2)
#head(Wage)
summary(Wage)
```

```
##       year           age                     maritl           race
##  Min.   :2003   Min.   :18.00   1. Never Married: 648   1. White:2480
##  1st Qu.:2004   1st Qu.:33.75   2. Married      :2074   2. Black: 293
##  Median :2006   Median :42.00   3. Widowed      :  19   3. Asian: 190
##  Mean   :2006   Mean   :42.41   4. Divorced     : 204   4. Other:  37
##  3rd Qu.:2008   3rd Qu.:51.00   5. Separated    :  55
##  Max.   :2009   Max.   :80.00
##
##                education                     region               jobclass
##  1. < HS Grad      :268   2. Middle Atlantic   :3000   1. Industrial :1544
##  2. HS Grad        :971   1. New England       :   0   2. Information:1456
##  3. Some College   :650   3. East North Central:   0
##  4. College Grad   :685   4. West North Central:   0
##  5. Advanced Degree:426   5. South Atlantic    :   0
##                           6. East South Central:   0
##                           (Other)              :   0
##             health      health_ins      logwage           wage
##  1. <=Good     : 858   1. Yes:2083   Min.   :3.000   Min.   : 20.09
##  2. >=Very Good:2142   2. No : 917   1st Qu.:4.447   1st Qu.: 85.38
##                                      Median :4.653   Median :104.92
##                                      Mean   :4.654   Mean   :111.70
##                                      3rd Qu.:4.857   3rd Qu.:128.68
##                                      Max.   :5.763   Max.   :318.34
##
```

In the first part of the assignment. We are interested in `wage` in relation to `year`, `age` and `education`. This is a paired plot.

```
pairs(data.frame(Wage$wage, Wage$year, Wage$age, Wage$education))
```



It is known that `year` has approximately linear trend and the variable `education` is a categorical variable. We use the natural spline curve fitting for the trend of `age`. For this we use function `ns()` in the `splines` package and `lm()` function. We fit the following models

$$\begin{aligned}
\text{model1:} &\quad \text{waga} \sim \text{year} + ns(\text{age}, df = 1) + \text{education},\\
\text{model2:} &\quad \text{waga} \sim \text{year} + ns(\text{age}, df = 3) + \text{education},\\
\text{model3:} &\quad \text{waga} \sim \text{year} + ns(\text{age}, df = 5) + \text{education},\\
\text{model4:} &\quad \text{waga} \sim \text{year} + ns(\text{age}, df = 7) + \text{education},\\
\text{model5:} &\quad \text{waga} \sim \text{year} + ns(\text{age}, df = 9) + \text{education}.
\end{aligned}$$

(a) **(10 marks)** Fit the model and use `anova()` function to do the deviance test to compare the models. Choose the best model.

```
library(splines)
```

(b) **(5 marks)** Calculate `AIC` for each model fitted in (a). Choose the best model using the value of `AIC`.

(c) **(10 marks)** Split the data set (100%) into a training set (70%) and a test set (30%). Then fit model1–model5 on the training set, and calculate the test MSE for each model. Choose the best model.

```
set.seed(11)
```

(d) **(10 marks)** By combining the result from (a), (b) and (c), decide the best model. Refit the chosen model using all of the `Wage` data set. Interpret the out of the `summary()` function.

Q2

Here we will predict the number of applications received `Apps` using the other variables in the "College" data set.

The data set contains 777 observations on the following 18 variables.

```
# Private: A factor with levels No and Yes indicating private or public university
# Apps: Number of applications received
```

```
# Accept: Number of applications accepted
# Enroll: Number of new students enrolled
# Top10perc: Pct. new students from top 10% of H.S. class
# Top25perc: Pct. new students from top 25% of H.S. class
# F.Undergrad: Number of fulltime undergraduates
# P.Undergrad: Number of parttime undergraduates
# Outstate: Out-of-state tuition
# Room.Board: Room and board costs
# Books: Estimated book costs
# Personal: Estimated personal spending
# PhD: Pct. of faculty with Ph.D.'s
# Terminal: Pct. of faculty with terminal degree
# S.F.Ratio: Student/faculty ratio
# perc.alumni: Pct. alumni who donate
# Expend: Instructional expenditure per student
# Grad.Rate: Graduation rate
```

```
library(ISLR)
```

```
##
## Attaching package: 'ISLR'
```

```
## The following objects are masked from 'package:ISLR2':
##
##     Auto, Credit
```

```
data(College)
summary(College)
```

```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56
##            3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00
##            Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00
##    Top25perc      F.Undergrad     P.Undergrad         Outstate
##  Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340
##  1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal         PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   Median :1200   Median : 75.00
##  Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66
##  3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00
##  Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :103.00
##     Terminal       S.F.Ratio      perc.alumni        Expend
##  Min.   : 24.0   Min.   : 2.50   Min.   : 0.00   Min.   : 3186
##  1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751
##  Median : 82.0   Median :13.60   Median :21.00   Median : 8377
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
```

```
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

(a) **(5 marks)** (Create trainig set and test set) Split the data set (100%) into a training set (70%) and a test set (30%).

```
set.seed(11)
```

(b) **(10 marks)** (LASSO) Fit a lasso model on the training set, with λ chosen by cross-validation with the `1 se rule` . Report the test error obtained, along with the of non-zero coefficient estimates.

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-3
```

```
grid <- 10 ^ seq(4, -2, length = 100)
```

- Test MSE
- Non-zero coefficient estimates

(c) **(10 marks)** Do the best subset selection with `BIC` and choose the best model.

```
library(leaps)
```

(d) **(10 marks)** Use all of the `College` data set, refit the models chosen by LASSO in (b) and best subset selection in (c). Print output of the function `summary()` for these models. Then compute 'AIC' and 'BIC'. Between these 2 models, which model is the better model. Give reasons why.

[**Total: 70 marks**]