DATA 303/473 Assignment 3

Name:Nicholas Tran, ID:300296259

Due: 12 May 2022

Assignment Questions

Q1

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

```
library(ISLR2)
```

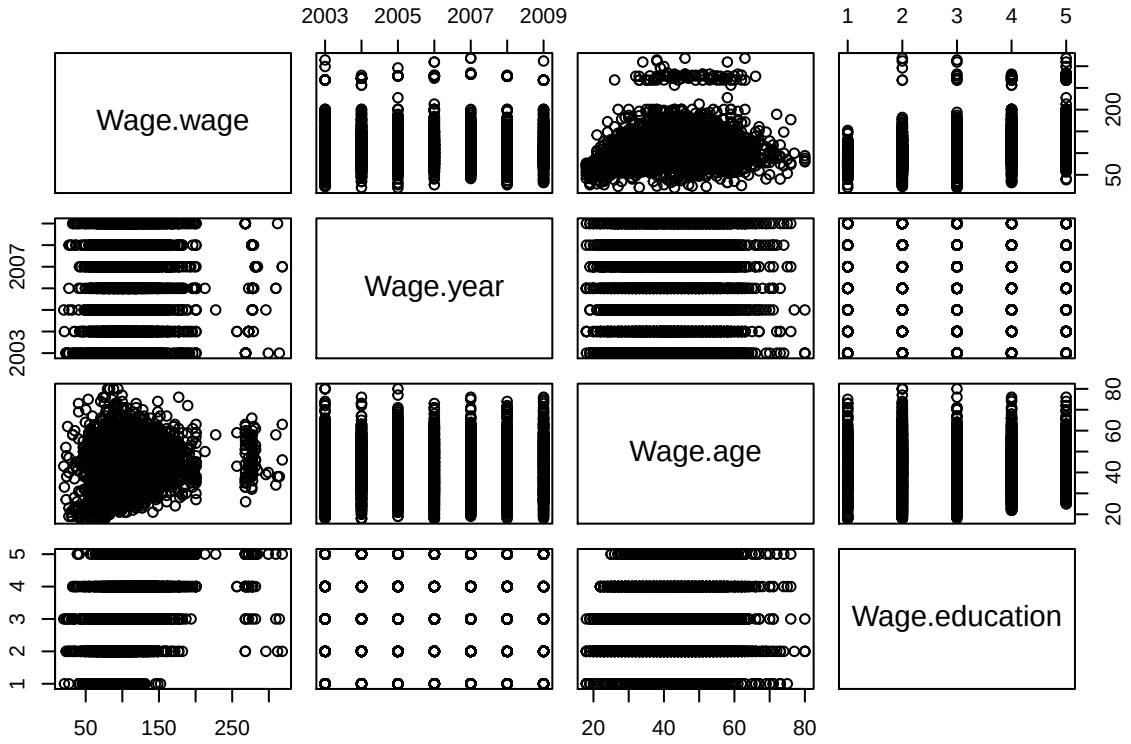
```
## Warning: package 'ISLR2' was built under R version 4.0.5
```

```
#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)

model1 <- lm(wage ~ year + ns(age, 1) + education, data = Wage)
model2 <- lm(wage ~ year + ns(age, 3) + education, data = Wage)
model3 <- lm(wage ~ year + ns(age, 5) + education, data = Wage)
model4 <- lm(wage ~ year + ns(age, 7) + education, data = Wage)
model5 <- lm(wage ~ year + ns(age, 9) + education, data = Wage)

anova(model1, model2, model3, model4, model5)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ year + ns(age, 1) + education
## Model 2: wage ~ year + ns(age, 3) + education
## Model 3: wage ~ year + ns(age, 5) + education
```

```
## Model 4: wage ~ year + ns(age, 7) + education
## Model 5: wage ~ year + ns(age, 9) + education
##   Res.Df     RSS Df Sum of Sq       F Pr(>F)
## 1   2993 3854286
## 2   2991 3699770  2    154516 62.5205 <2e-16 ***
## 3   2989 3694885  2      4885  1.9765 0.1387
## 4   2987 3692452  2      2433  0.9845 0.3737
## 5   2985 3688635  2      3817  1.5443 0.2136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

model 1 vs. model 2: Significant. p-value is incredibly low so we pick model 2 over model 1.

model 2 vs. model 3: Insignificant. p-value is high so there's insufficient evidence to choose model 3 over model 2.

model 3 vs. model 4: Insignificant. p-value is high so there's insufficient evidence to choose model 4 over model 3.

model 4 vs. model 5: Insignificant. p-value is high so there's insufficient evidence to choose model5 over model 4.

Based on the results of the deviance tests, we pick model 2 - the model with age as a 3rd degree natural spline - as the best model.

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

```
AIC(model1, model2, model3, model4, model5)
```

```
##        df      AIC
## model1  8 30004.62
## model2 10 29885.87
## model3 12 29885.91
## model4 14 29887.93
## model5 16 29888.83
```

The model with the lowest AIC is model 2 (29885.87 AIC), the model with age as a 3rd degree natural spline - the same model chosen by the deviance test.

(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)

train_index <- sample(nrow(Wage), nrow(Wage)*0.7)
train = Wage[train_index,]
test = Wage[-train_index,]

model1 <- lm(wage ~ year + ns(age, 1) + education, data = train)
model2 <- lm(wage ~ year + ns(age, 3) + education, data = train)
model3 <- lm(wage ~ year + ns(age, 5) + education, data = train)
model4 <- lm(wage ~ year + ns(age, 7) + education, data = train)
model5 <- lm(wage ~ year + ns(age, 9) + education, data = train)

y <- test$wage
y_hat1 <- predict(model1, newdata = test)
y_hat2 <- predict(model2, newdata = test)
y_hat3 <- predict(model3, newdata = test)
y_hat4 <- predict(model4, newdata = test)
```

```
y_hat5 <- predict(model5, newdata = test)

MSE1 <- mean((y-y_hat1)^2)
MSE2 <- mean((y-y_hat2)^2)
MSE3 <- mean((y-y_hat3)^2)
MSE4 <- mean((y-y_hat4)^2)
MSE5 <- mean((y-y_hat5)^2)

c(MSE1, MSE2, MSE3, MSE4, MSE5)
```

```
## [1] 1406.339 1335.942 1332.825 1334.092 1332.456
```

The model with the lowest MSE is model 5 (the model with age as a 9th degree natural spline), however model 2, model 3, model 4 and model 5 all have really similar MSEs. Potentially cross-validation may be needed to achieve more accurate results.

(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.

Even though the model with the lowest MSE is model 4, the AIC and deviance test recommends the second model and the MSE between model 2 and model 4 are close enough that model 2 seems like the better choice.

```
model2 <- lm(wage ~ year + ns(age, 3) + education, data = Wage)

summary(model2)
```

```
##
## Call:
## lm(formula = wage ~ year + ns(age, 3) + education, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -119.258  -19.694   -3.259   14.259  213.400
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -2322.8434   637.0840  -3.646 0.000271 ***
## year                            1.1815     0.3176   3.720 0.000203 ***
## ns(age, 3)1                    30.4808     2.9799  10.229  < 2e-16 ***
## ns(age, 3)2                    74.5353     8.0524   9.256  < 2e-16 ***
## ns(age, 3)3                     4.1361     6.3388   0.653 0.514124
## education2. HS Grad            10.9180     2.4282   4.496 7.18e-06 ***
## education3. Some College       23.4279     2.5550   9.170  < 2e-16 ***
## education4. College Grad       38.0297     2.5394  14.976  < 2e-16 ***
## education5. Advanced Degree    62.4889     2.7566  22.669  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.17 on 2991 degrees of freedom
## Multiple R-squared:  0.2915, Adjusted R-squared:  0.2896
## F-statistic: 153.8 on 8 and 2991 DF,  p-value: < 2.2e-16
```

p-values of regression coefficients: All predictors besides from the natural spline of degree 3 of age have incredibly low p-values which means that all the predictors besides from ns(age, 3)3 are significantly related to the response variable wage.

R-squared: The R^2 value of the model is 0.2915 which means only 29.15% of the variance in wage is

explained by the model. The model isn't a good predictor of wage.

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)

train_index <- sample(nrow(College), nrow(College)*0.7)
train = College[train_index,]
test = College[-train_index,]

dim(train)
```

```
## [1] 543  18
```

```
dim(test)
```

```
## [1] 234  18
```

(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)
```

```
## Warning: package 'glmnet' was built under R version 4.0.5
```

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

```
## Loaded glmnet 4.1-2
```

```
y <- train$Apps
x <- model.matrix(Apps ~., train)

lasso.mod <- glmnet(x, y, alpha=1)
grid <- 10 ^ seq(4, -2, length = 100)
cv.out <- cv.glmnet(x, y, alpha=1, lambda = grid)
lam1se <- cv.out$lambda.1se
lam1se
```

```
## [1] 403.7017
```

```
log(lam1se)
```

```
## [1] 6.000676
```

- Test MSE

```
y.test <- test$Apps
x.test <- model.matrix(Apps ~., test)

lasso.predict <- predict(lasso.mod, s=lam1se, newx = x.test)
MSE <- mean((lasso.predict - y.test)^2)
MSE
```

```
## [1] 542125.9
```

- Non-zero coefficient estimates

```
lasso.coeff <- predict(lasso.mod, type="coefficients", s=lam1se)
lasso.coeff
```

```
## 19 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept) -2.179777e+02
## (Intercept)  .
## PrivateYes   .
## Accept       1.318419e+00
## Enroll       .
## Top10perc    2.151183e+01
## Top25perc    .
## F.Undergrad  .
## P.Undergrad  .
## Outstate     .
## Room.Board   .
## Books        .
## Personal     .
## PhD          .
## Terminal     .
## S.F.Ratio    .
## perc.alumni  .
## Expend       6.356122e-04
## Grad.Rate    .
```

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

```
library(leaps)

regfit.full <- regsubsets(Apps ~ ., train)
reg.summary <- summary(regfit.full)
which.min(reg.summary$bic)
```

```
## [1] 8
```

```
coef(regfit.full, 8)
```

```
##   (Intercept)        Accept        Enroll     Top10perc     Top25perc
## -242.11867623    1.68394363   -1.46446495   62.13448396  -24.58660224
##   F.Undergrad     Outstate        Expend     Grad.Rate
##    0.15403138   -0.12246516    0.07623521    9.85210993
```

According to the results of BIC for subset selection, the best model chosen is the one with Accept, Enroll, Top10perc, Top25perc, F.undergrad, Outstate, Expend and Grad.Rate as predictors.

(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.

```
model.bic <- lm(Apps ~ Accept + Enroll + Top10perc + Top25perc + F.Undergrad + Outstate + Expend + Grad

summary(model.bic)
```

```
##
## Call:
## lm(formula = Apps ~ Accept + Enroll + Top10perc + Top25perc +
##     F.Undergrad + Outstate + Expend + Grad.Rate, data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5177.6  -440.2   -27.4   307.8  7559.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -406.23399  201.56571  -2.015 0.044211 *
## Accept         1.60792    0.04009  40.109  < 2e-16 ***
## Enroll        -0.98427    0.18567  -5.301 1.51e-07 ***
## Top10perc     47.83604    5.58532   8.565  < 2e-16 ***
## Top25perc    -16.27312    4.44104  -3.664 0.000265 ***
## F.Undergrad    0.09235    0.03105   2.974 0.003031 **
## Outstate      -0.10259    0.01557  -6.587 8.31e-11 ***
## Expend         0.07692    0.01145   6.719 3.56e-11 ***
## Grad.Rate      7.72318    2.85103   2.709 0.006901 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1057 on 768 degrees of freedom
## Multiple R-squared:  0.9261, Adjusted R-squared:  0.9254
## F-statistic:  1204 on 8 and 768 DF,  p-value: < 2.2e-16
```

```
model.lasso <- lm(Apps ~ Accept + Top10perc + Expend, data = College)

summary(model.lasso)
```

```
##
## Call:
## lm(formula = Apps ~ Accept + Top10perc + Expend, data = College)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5272.0  -475.8   -24.9   290.8  9816.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e+03  8.818e+01 -12.546  < 2e-16 ***
## Accept       1.440e+00  1.654e-02  87.089  < 2e-16 ***
## Top10perc    2.606e+01  3.038e+00   8.579  < 2e-16 ***
## Expend       4.989e-02  1.015e-02   4.915 1.09e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1108 on 773 degrees of freedom
## Multiple R-squared:  0.9183, Adjusted R-squared:  0.918
## F-statistic:  2898 on 3 and 773 DF,  p-value: < 2.2e-16
```

```
AIC(model.bic, model.lasso)
```

```
##             df      AIC
## model.bic   10 13037.24
## model.lasso  5 13105.23
```

```
BIC(model.bic, model.lasso)
```

```
##             df      BIC
## model.bic   10 13083.79
## model.lasso  5 13128.51
```

Regression coefficients: All the p-values for the coefficients in the model picked by BIC are all incredibly low which means that all predictors in the model have a significant relationship with the response variable. The same goes for the model chosen by LASSO.

Adjusted R-squared: The adjusted R^2 for the model chosen by BIC is 0.9254 which means that 92.54% of the variance in the response variable can be explained by the model. Compare to the model chosen by LASSO, the adjusted R^2 is 0.918, 91.8% which is slightly lower. From my understanding the adjusted R^2 already penalises for added predictors so if the model with more predictors has a better R^2 even with the penalty then that's the model that should be chosen which is the model chosen by BIC.

AIC and BIC: Both BIC and AIC chooses the model chosen by BIC (it has the lowest AIC/BIC),

```
13128.51 - 13083.79
```

```
## [1] 44.72
```

```
13105.23 - 13037.24
```

```
## [1] 67.99
```

The differences in AIC/BIC between the models isn't small enough for us to disregard the results and choose the simpler model.

Based on all the criteria above it appears the best model is the model chosen by BIC, the model with Accept, Enroll, Top10perc, Top25perc, F.undergrad, Outstate, Expend and Grad.Rate as predictors. The model chosen by LASSO omits too many important predictors that contributes to the variance of the response variable.

[**Total: 70 marks**]