Homework-2 Group: BUAN6356502 6

Nimish Bhandare-(NCB190000), Naga Deepa Alavalapati-(NXA180060), Shivdatta Tanaji Patil-(SXP190008), Bhushan Patil-(BBP190000),

01/10/2019

```r
air.df <- read.csv(file="Airfares.csv", header=TRUE, sep=",")
air.df<-air.df[ ,-c(1:4)]
str(air.df)
```

```
## 'data.frame':    638 obs. of  14 variables:
##  $ COUPON  : num  1 1.06 1.06 1.06 1.06 1.01 1.28 1.15 1.33 1.6 ...
##  $ NEW     : int  3 3 3 3 3 3 3 3 3 2 ...
##  $ VACATION: Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 2 1 1 ...
##  $ SW      : Factor w/ 2 levels "No","Yes": 2 1 1 2 2 2 1 2 2 2 ...
##  $ HI      : num  5292 5419 9185 2657 2657 ...
##  $ S_INCOME: num  28637 26993 30124 29260 29260 ...
##  $ E_INCOME: num  21112 29838 29838 29838 29838 ...
##  $ S_POP   : int  3036732 3532657 5787293 7830332 7830332 2230955 3036732 1440377 3770125 1694803 ..
##  $ E_POP   : int  205711 7145897 7145897 7145897 7145897 7145897 7145897 7145897 7145897 7145897 ...
##  $ SLOT    : Factor w/ 2 levels "Controlled","Free": 2 2 2 1 2 2 2 2 2 2 ...
##  $ GATE    : Factor w/ 2 levels "Constrained",..: 2 2 2 2 2 2 2 2 2 2 ...
##  $ DISTANCE: int  312 576 364 612 612 309 1220 921 1249 964 ...
##  $ PAX     : int  7864 8820 6452 25144 25144 13386 4625 5512 7811 4657 ...
##  $ FARE    : num  64.1 174.5 207.8 85.5 85.5 ...
```

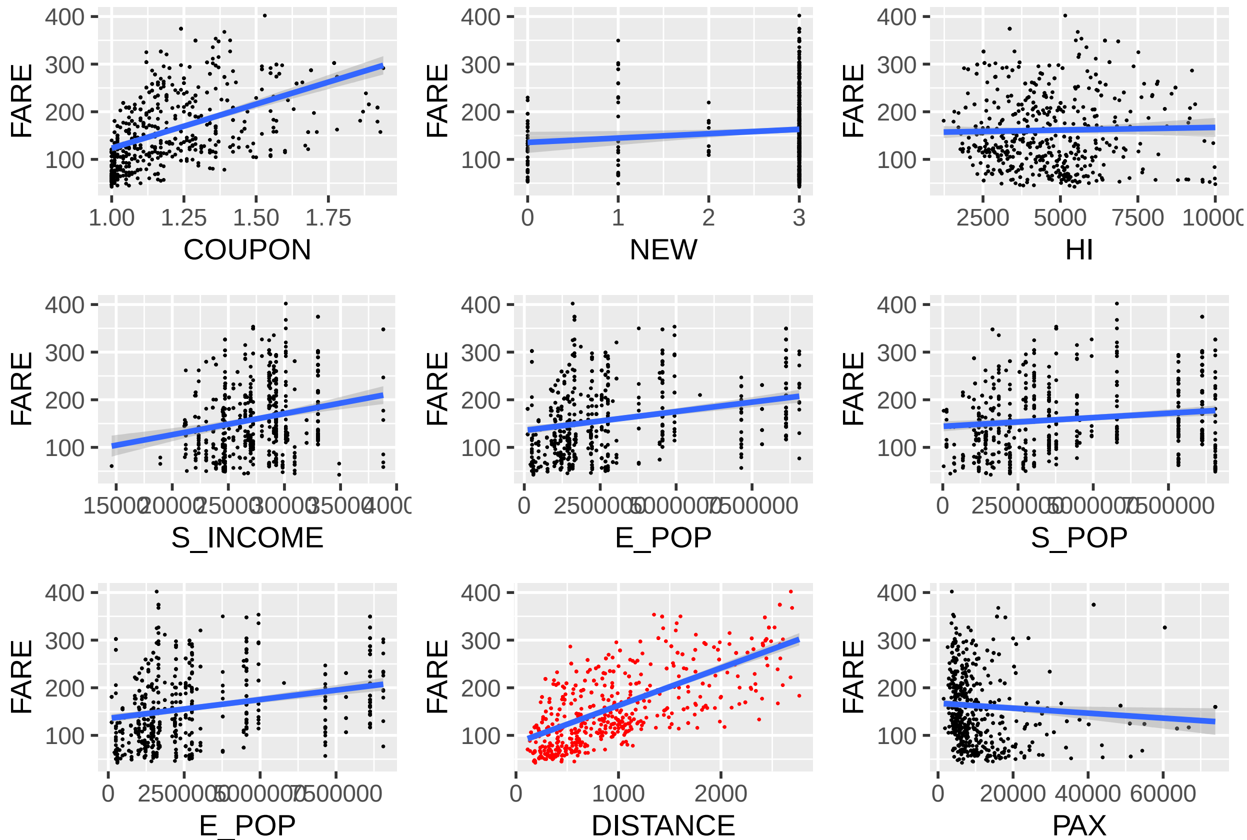
Q1. Create a correlation table and scatterplots between FARE and the predictors. What seems to be the best single predictor of FARE? Explain your answer

```r
corr <- cor(air.df[,-c(3,4,10,11)])
corr
```

```
##               COUPON         NEW          HI    S_INCOME   E_INCOME
## COUPON    1.00000000  0.02022307 -0.34725207 -0.08840265  0.0468892
## NEW       0.02022307  1.00000000  0.05414685  0.02659673  0.1133766
## HI       -0.34725207  0.05414685  1.00000000 -0.02738221  0.0823926
## S_INCOME -0.08840265  0.02659673 -0.02738221  1.00000000 -0.1388642
## E_INCOME  0.04688920  0.11337664  0.08239260 -0.13886420  1.0000000
## S_POP    -0.10776336 -0.01667212 -0.17249541  0.51718718 -0.1440586
## E_POP     0.09496994  0.05856818 -0.06245600 -0.27228027  0.4584181
## DISTANCE  0.74680521  0.08096520 -0.31237457  0.02815334  0.1765307
## PAX      -0.33697358  0.01049527 -0.16896078  0.13819710  0.2599611
## FARE      0.49653696  0.09172969  0.02519492  0.20913485  0.3260923
##                S_POP       E_POP    DISTANCE         PAX        FARE
## COUPON   -0.10776336  0.09496994  0.74680521 -0.33697358  0.49653696
## NEW      -0.01667212  0.05856818  0.08096520  0.01049527  0.09172969
## HI       -0.17249541 -0.06245600 -0.31237457 -0.16896078  0.02519492
## S_INCOME  0.51718718 -0.27228027  0.02815334  0.13819710  0.20913485
## E_INCOME -0.14405857  0.45841806  0.17653074  0.25996105  0.32609229
## S_POP     1.00000000 -0.28014283  0.01843667  0.28461056  0.14509708
## E_POP    -0.28014283  1.00000000  0.11563970  0.31469750  0.28504299
```

```
## DISTANCE  0.01843667  0.11563970  1.00000000 -0.10248160  0.67001599
## PAX       0.28461056  0.31469750 -0.10248160  1.00000000 -0.09070541
## FARE      0.14509708  0.28504299  0.67001599 -0.09070541  1.00000000
```

```
cplot<- ggplot(data = air.df, aes(x = COUPON, y= FARE ))+geom_point(color = 'black', size=0.001)+stat_sm

nplot<-ggplot(data = air.df, aes(x = NEW, y= FARE ))+geom_point(color = 'black',size=0.001)+stat_smooth


hplot<-ggplot(data = air.df, aes(x = HI, y= FARE ))+geom_point(color = 'black',size=0.001)+stat_smooth(

splot<-ggplot(data = air.df, aes(x = S_INCOME, y= FARE ))+geom_point(color = 'black', size=0.001)+stat_

eplot<-ggplot(data = air.df, aes(x = E_INCOME, y= FARE ))+geom_point(color = 'black', size=0.001)+stat_

pplot<-ggplot(data = air.df, aes(x = S_POP, y= FARE ))+geom_point(color = 'black',size=0.001)+stat_smoo

eplot<-ggplot(data = air.df, aes(x = E_POP, y= FARE ))+geom_point(color = 'black',size=0.001)+stat_smoo


dplot<-ggplot(data = air.df, aes(x = DISTANCE, y= FARE ))+geom_point(color = 'red', size=0.001)+stat_sm

aplot<-ggplot(data = air.df, aes(x = PAX, y= FARE ))+geom_point(color = 'black',size=0.001)+stat_smooth
ggarrange(cplot,nplot,hplot,splot,eplot,pplot,eplot,dplot,aplot)
```



It can be seen from both correlation table and the scatter plots, DISTANCE has the highest correlation with FARE. FARE has correlation of 0.67 with DISTANCE in coerrelation table and amongst all scatter plots it has the most positive trend. Hence DISTANCE is the best single predictor of FARE, as more the DISTANCE, more is the FARE of air ticket on that route.

Q2. Explore the categorical predictors by computing the percentage of flights in each category. Create a pivot table with the average fare in each category. Which categorical predictor seems best for predicting FARE? Explain your answer

```
Vacation <- air.df %>%
        dplyr::select(VACATION,FARE) %>%
        group_by(VACATION) %>%
        summarise(Count = length(VACATION),Total = nrow(air.df), Percent = percent(length(VACATION)/nrow

Southwest <- air.df %>%
        dplyr::select(SW,FARE) %>%
        group_by(SW) %>%
        summarise(Count = length(SW),Total = nrow(air.df), Percent = percent(length(SW)/nrow(air.df)), 

Gate <- air.df %>%
        dplyr::select(GATE,FARE) %>%
        group_by(GATE) %>%
        summarise(Count = length(GATE),Total = nrow(air.df), Percent = percent(length(GATE)/nrow(air.df

Slot <- air.df %>%
        dplyr::select(SLOT,FARE) %>%
        group_by(SLOT) %>%
        summarise(Count = length(SLOT),Total = nrow(air.df), Percent = percent(length(SLOT)/nrow(air.df

Southwest
```

```
## # A tibble: 2 x 5
##   SW    Count Total Percent AvgFare
##   <fct> <int> <int> <chr>     <dbl>
## 1 No      444   638 69.6%     188.
## 2 Yes     194   638 30.4%      98.4
```

```
Vacation
```

```
## # A tibble: 2 x 5
##   VACATION Count Total Percent AvgFare
##   <fct>    <int> <int> <chr>     <dbl>
## 1 No         468   638 73.4%      174.
## 2 Yes        170   638 26.6%      126.
```

```
Gate
```

```
## # A tibble: 2 x 5
##   GATE        Count Total Percent AvgFare
##   <fct>       <int> <int> <chr>     <dbl>
## 1 Constrained   124   638 19.4%      193.
## 2 Free          514   638 80.6%      153.
```

```
Slot
```

```
## # A tibble: 2 x 5
##   SLOT       Count Total Percent AvgFare
##   <fct>      <int> <int> <chr>     <dbl>
## 1 Controlled   182   638 28.5%      186.
## 2 Free         456   638 71.5%      151.
```

As seen from the pivot tables, compared to other categorical variables. SW depicts a significant change in average fare($188.18 to $98.38),i.ethe FARE gets affected largely if the route is covered by SouthWest Airlines.

Q3. Create data partition by assigning 80% of the records to the training dataset. Use rounding if 80% of the index generates a fraction. Also, set the seed at 42

```
set.seed(42)
train.index <- sample(1:nrow(air.df), 0.8 *round(nrow(air.df)))
train.df <- air.df[train.index, ]
valid.df <- air.df[-train.index, ]
```

Q4. Using leaps package, run stepwise regression to reduce the number of predictors. Discuss the results from this model.

```
search <- regsubsets(FARE ~ ., data = train.df, nbest = 1, nvmax = dim(air.df)[2],method = "seqrep")
stepwise <- summary(search)
stepwise$which
```

```
##    (Intercept) COUPON   NEW VACATIONYes SWYes    HI S_INCOME E_INCOME
## 1         TRUE  FALSE FALSE       FALSE FALSE FALSE    FALSE    FALSE
## 2         TRUE  FALSE FALSE       FALSE  TRUE FALSE    FALSE    FALSE
## 3         TRUE  FALSE FALSE        TRUE  TRUE FALSE    FALSE    FALSE
## 4         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 5         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 6         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 7         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE
## 8         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE
## 9         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 10        TRUE   TRUE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
## 11        TRUE  FALSE  TRUE        TRUE  TRUE  TRUE    FALSE     TRUE
## 12        TRUE  FALSE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
## 13        TRUE   TRUE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
##    S_POP E_POP SLOTFree GATEFree DISTANCE   PAX
## 1  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 2  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 3  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 4  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 5  FALSE FALSE     TRUE    FALSE     TRUE FALSE
## 6  FALSE FALSE     TRUE     TRUE     TRUE FALSE
## 7  FALSE FALSE     TRUE     TRUE     TRUE FALSE
## 8   TRUE  TRUE    FALSE    FALSE     TRUE  TRUE
## 9   TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 10  TRUE  TRUE     TRUE    FALSE    FALSE FALSE
## 11  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
```

```
## 12  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 13  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
```

```r
print("R-square")
```

```
## [1] "R-square"
```

```r
stepwise$rsq
```

```
##  [1] 0.4168069 0.5793894 0.6966218 0.7232479 0.7366555 0.7565835 0.7604199
##  [8] 0.7674947 0.7748171 0.6303171 0.7809073 0.7813501 0.7816700
```

```r
print("Adjusted R-square")
```

```
## [1] "Adjusted R-square"
```

```r
stepwise$adjr2
```

```
##  [1] 0.4156589 0.5777302 0.6948231 0.7210558 0.7340429 0.7536799 0.7570792
##  [8] 0.7637820 0.7707638 0.6229086 0.7760679 0.7760708 0.7759476
```

```r
print("Mallow’s Cp")
```

```
## [1] "Mallow’s Cp"
```

```r
stepwise$cp
```

```
##  [1] 818.89220 451.53899 187.21153 128.72255 100.26346  56.99127  50.27558
##  [8]  36.20326  21.56831 351.84190  11.73270  12.72670  14.00000
```

We can interpret this model by taking into consideration the Adjusted R-square and Mallow’s Cp values. As seen from above Adjusted R-square values there is no significant increase in adjusted r-square after considering 11 variables (0.7760).The Mallow’s Cp value for 11 variables in our model is 11.7320 which is closest to the ideal value of 12 according to the formula (p+1). Therefore according to stepwise search the best variables for predicting FARE are NEW, VACATION, SW, HI, E_INCOME, S_POP, E_POP, SLOT, GATE, DISTANCE, PAX.

Q5. Repeat the process in (4) using exhaustive search instead of stepwise regression. Compare the resulting best model to the one you obtained in (4) in terms of the predictors included in the final model.

```r
search <- regsubsets(FARE ~ ., data = train.df, nbest = 1, nvmax = dim(air.df)[2],method = "exhaustive")
exhaustive <- summary(search)
exhaustive$which
```

```
##    (Intercept) COUPON   NEW VACATIONYes SWYes    HI S_INCOME E_INCOME
## 1         TRUE  FALSE FALSE       FALSE FALSE FALSE    FALSE    FALSE
## 2         TRUE  FALSE FALSE       FALSE  TRUE FALSE    FALSE    FALSE
## 3         TRUE  FALSE FALSE        TRUE  TRUE FALSE    FALSE    FALSE
## 4         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
```

```
## 5          TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 6          TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 7          TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 8          TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE
## 9          TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 10         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE
## 11         TRUE  FALSE  TRUE        TRUE  TRUE  TRUE    FALSE     TRUE
## 12         TRUE  FALSE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
## 13         TRUE   TRUE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
##    S_POP E_POP SLOTFree GATEFree DISTANCE   PAX
## 1  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 2  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 3  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 4  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 5  FALSE FALSE     TRUE    FALSE     TRUE FALSE
## 6  FALSE FALSE     TRUE     TRUE     TRUE FALSE
## 7   TRUE  TRUE    FALSE    FALSE     TRUE  TRUE
## 8   TRUE  TRUE    FALSE    FALSE     TRUE  TRUE
## 9   TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 10  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 11  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 12  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 13  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
```

```r
print("R-square")
```

```
## [1] "R-square"
```

```r
exhaustive$rsq
```

```
##  [1] 0.4168069 0.5793894 0.6966218 0.7232479 0.7366555 0.7565835 0.7607777
##  [8] 0.7674947 0.7748171 0.7803115 0.7809073 0.7813501 0.7816700
```

```r
print("Adjusted R-square")
```

```
## [1] "Adjusted R-square"
```

```r
exhaustive$adjr2
```

```
##  [1] 0.4156589 0.5777302 0.6948231 0.7210558 0.7340429 0.7536799 0.7574419
##  [8] 0.7637820 0.7707638 0.7759090 0.7760679 0.7760708 0.7759476
```

```r
print("Mallow's Cp")
```

```
## [1] "Mallow's Cp"
```

```r
exhaustive$cp
```

```
##  [1] 818.89220 451.53899 187.21153 128.72255 100.26346  56.99127  49.46286
##  [8]  36.20326  21.56831  11.08605  11.73270  12.72670  14.00000
```

We can interpret this model by taking into consideration the Adjusted R-square and Mallow's Cp values. As seen from above Adjusted R-square values there is no significant increase in adjusted r-square after considering 10 variables (0.7759) . The Mallow's Cp value for 10 variables in our model is 11.08605 which is closest to the ideal value of 11 according to the formula (p+1). Therefore according to stepwise search the best variables for predicting FARE are VACATION, SW, HI, E_INCOME, S_POP, E_POP, SLOT, GATE, DISTANCE, PAX.

Q6.Compare the predictive accuracy of both models—stepwise regression and exhaustive search—using measures such as RMSE.

```
print("Stepwise Search")
```

```
## [1] "Stepwise Search"
```

```
stepwise.lm<-lm(formula = FARE ~ NEW+ VACATION + SW + HI + E_INCOME + S_POP + E_POP +SLOT + GATE + DIST
stepwise.lm.pred <- predict(stepwise.lm,valid.df)
accuracy(stepwise.lm.pred,valid.df$FARE)
```

```
##                ME     RMSE      MAE       MPE     MAPE
## Test set 3.166677 36.82363 27.57897 -5.812025 21.44043
```

```
print("Exhaustive Search")
```

```
## [1] "Exhaustive Search"
```

```
exhaustive.lm<-lm(formula = FARE ~ VACATION + SW + HI + E_INCOME + S_POP + E_POP +
    SLOT + GATE + DISTANCE + PAX, data = train.df )
exhaustive.lm.pred <- predict(exhaustive.lm,valid.df)
accuracy(exhaustive.lm.pred,valid.df$FARE)
```

```
##               ME    RMSE      MAE       MPE     MAPE
## Test set 3.06081 36.8617 27.70568 -5.938062 21.62142
```

RMSE is a measure of how spread out the residuals are, therfore lower the RMSE value signifies a better fit. As seen from above comparison it is evident that stepwise search has slightly low RMSE (36.823) than RMSE value of exhaustive search (36.861). Hence stepwise model is a better fit.

Q7. Using the exhaustive search model, predict the average fare on a route with the following characteristics: COUPON = 1.202, NEW = 3, VACATION = No, SW = No, HI = 4442.141, S_INCOME = $28,760, E_INCOME = $27,664, S_POP = 4,557,004, E_POP = 3,195,503, SLOT = Free, GATE = Free, PAX = 12,782, DISTANCE = 1976 miles.

```
valid1.df <- data.frame('COUPON' = 1.202, 'NEW' = 3, 'VACATION' = 'No', 'SW' =
'No', 'HI' = 4442.141, 'S_INCOME' = 28760, 'E_INCOME' = 27664, 'S_POP' =
4557004, 'E_POP' = 3195503, 'SLOT' = 'Free', 'GATE' = 'Free', 'PAX' = 12782,
'DISTANCE' = 1976)

exhaustive.lm.pred <- predict(exhaustive.lm,valid1.df)
exhaustive.lm.pred
```

```
##       1
## 247.684
```

According to given variable values the exhaustive search model predicts a average fare of $247.684.

Q8. Predict the reduction in average fare on the route in question (7.), if Southwest decides to cover this route [using the exhaustive search model above].

```
valid2.df <- data.frame('COUPON' = 1.202, 'NEW' = 3, 'VACATION' = 'No', 'SW' =
'Yes', 'HI' = 4442.141, 'S_INCOME' = 28760, 'E_INCOME' = 27664, 'S_POP' =
4557004, 'E_POP' = 3195503, 'SLOT' = 'Free', 'GATE' = 'Free', 'PAX' = 12782,
'DISTANCE' = 1976)
exhaustive.lm.pred <- predict(exhaustive.lm,valid2.df)
exhaustive.lm.pred
```

```
##        1
## 207.1558
```

According to given variable values the exhaustive search model predicts a average fare of $207.1558. We can conclude that there is a reduction in average fare when Southwest airlines covers the route as compared to average fare on the route which Southwest airlines doesn't operate on.

Q9. Using leaps package, run backward selection regression to reduce the number of predictors. Discuss the results from this model.

```
search <- regsubsets(FARE ~ ., data = train.df, nbest = 1, nvmax = dim(air.df)[2],method = "backward")
backward <- summary(search)
backward$which
```

```
##    (Intercept) COUPON   NEW VACATIONYes SWYes    HI S_INCOME E_INCOME
## 1         TRUE  FALSE FALSE       FALSE FALSE FALSE    FALSE    FALSE
## 2         TRUE  FALSE FALSE       FALSE  TRUE FALSE    FALSE    FALSE
## 3         TRUE  FALSE FALSE        TRUE  TRUE FALSE    FALSE    FALSE
## 4         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 5         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 6         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 7         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 8         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 9         TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE    FALSE
## 10        TRUE  FALSE FALSE        TRUE  TRUE  TRUE    FALSE     TRUE
## 11        TRUE  FALSE  TRUE        TRUE  TRUE  TRUE    FALSE     TRUE
## 12        TRUE  FALSE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
## 13        TRUE   TRUE  TRUE        TRUE  TRUE  TRUE     TRUE     TRUE
##    S_POP E_POP SLOTFree GATEFree DISTANCE   PAX
## 1  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 2  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 3  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 4  FALSE FALSE    FALSE    FALSE     TRUE FALSE
## 5  FALSE  TRUE    FALSE    FALSE     TRUE FALSE
## 6   TRUE  TRUE    FALSE    FALSE     TRUE FALSE
## 7   TRUE  TRUE    FALSE    FALSE     TRUE  TRUE
## 8   TRUE  TRUE    FALSE     TRUE     TRUE  TRUE
## 9   TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 10  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 11  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 12  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
## 13  TRUE  TRUE     TRUE     TRUE     TRUE  TRUE
```

```r
print("R-square")
```

```
## [1] "R-square"
```

```r
backward$rsq
```

```
##  [1] 0.4168069 0.5793894 0.6966218 0.7232479 0.7322282 0.7509946 0.7607777
##  [8] 0.7663728 0.7748171 0.7803115 0.7809073 0.7813501 0.7816700
```

```r
print("Adjusted R-square")
```

```
## [1] "Adjusted R-square"
```

```r
backward$adjr2
```

```
##  [1] 0.4156589 0.5777302 0.6948231 0.7210558 0.7295718 0.7480243 0.7574419
##  [8] 0.7626422 0.7707638 0.7759090 0.7760679 0.7760708 0.7759476
```

```r
print("Mallow's Cp")
```

```
## [1] "Mallow's Cp"
```

```r
backward$cp
```

```
##  [1] 818.89220 451.53899 187.21153 128.72255 110.32120  69.68802  49.46286
##  [8]  38.75199  21.56831  11.08605  11.73270  12.72670  14.00000
```

We can interpret this backward search model by taking into consideration the Adjusted R-square and Mallow's Cp values. As seen from above Adjusted R-square values there is no significant increase in adjusted r-square after considering 10 variables (0.7759) . The Mallow's Cp value for 10 variables in our model is 11.08605 which is closest to the ideal value of 11 according to the formula (p+1). Therefore according to stepwise search the best variables for predicting FARE are VACATION, SW, HI, E_INCOME, S_POP, E_POP, SLOT, GATE, DISTANCE, PAX. However backward search model in not reccomended when the number of predictor variables is high, as its computation is expensive.

Q10. Now run a backward selection model using stepAIC() function. Discuss the results from this model, including the role of AIC in this model.

```r
air.lm<-lm(FARE ~ .,data = train.df)
air.lm<- stepAIC(air.lm,direction = "backward")
```

```
## Start:  AIC=3652.06
## FARE ~ COUPON + NEW + VACATION + SW + HI + S_INCOME + E_INCOME +
##     S_POP + E_POP + SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq    RSS    AIC
## - COUPON    1       911 622732 3650.8
## - NEW       1      1459 623280 3651.3
## - S_INCOME  1      1460 623281 3651.3
```

```
## <none>                   621821 3652.1
## - E_INCOME  1     17499  639320 3664.2
## - SLOT      1     17769  639590 3664.4
## - PAX       1     24441  646263 3669.7
## - E_POP     1     28296  650118 3672.8
## - GATE      1     28881  650702 3673.2
## - S_POP     1     36680  658501 3679.3
## - HI        1     76469  698290 3709.2
## - SW        1    105205  727026 3729.8
## - VACATION  1    113382  735204 3735.5
## - DISTANCE  1    417379 1039200 3912.0
##
## Step:  AIC=3650.81
## FARE ~ NEW + VACATION + SW + HI + S_INCOME + E_INCOME + S_POP +
##     E_POP + SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## - S_INCOME  1      1261  623994 3649.8
## - NEW       1      1678  624410 3650.2
## <none>                   622732 3650.8
## - E_INCOME  1     17126  639859 3662.6
## - SLOT      1     18407  641139 3663.7
## - GATE      1     29285  652018 3672.2
## - E_POP     1     29484  652217 3672.4
## - PAX       1     34128  656860 3676.0
## - S_POP     1     36089  658821 3677.5
## - HI        1     78594  701326 3709.4
## - SW        1    107735  730468 3730.2
## - VACATION  1    114276  737009 3734.7
## - DISTANCE  1    824468 1447200 4078.9
##
## Step:  AIC=3649.84
## FARE ~ NEW + VACATION + SW + HI + E_INCOME + S_POP + E_POP +
##     SLOT + GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
## - NEW       1      1697  625690 3649.2
## <none>                   623994 3649.8
## - E_INCOME  1     16167  640161 3660.9
## - SLOT      1     20012  644006 3663.9
## - E_POP     1     28559  652552 3670.7
## - GATE      1     29766  653759 3671.6
## - PAX       1     32869  656863 3674.0
## - S_POP     1     41722  665715 3680.8
## - HI        1     79501  703495 3709.0
## - SW        1    126837  750831 3742.2
## - VACATION  1    128080  752073 3743.1
## - DISTANCE  1    826967 1450960 4078.2
##
## Step:  AIC=3649.22
## FARE ~ VACATION + SW + HI + E_INCOME + S_POP + E_POP + SLOT +
##     GATE + DISTANCE + PAX
##
##            Df Sum of Sq     RSS    AIC
```

```
## <none>                    625690 3649.2
## - E_INCOME  1     15649  641339 3659.8
## - SLOT      1     19217  644907 3662.6
## - E_POP     1     28766  654456 3670.1
## - GATE      1     29165  654856 3670.5
## - PAX       1     32706  658396 3673.2
## - S_POP     1     42648  668338 3680.9
## - HI        1     78891  704581 3707.8
## - SW        1    126577  752267 3741.2
## - VACATION  1    127066  752756 3741.5
## - DISTANCE  1    825966 1451656 4076.4
```

```r
summary(air.lm)
```

```
##
## Call:
## lm(formula = FARE ~ VACATION + SW + HI + E_INCOME + S_POP + E_POP +
##     SLOT + GATE + DISTANCE + PAX, data = train.df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -99.148 -22.077  -2.028  21.491 107.744
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.208e+01  1.476e+01   2.851 0.004534 **
## VACATIONYes -3.876e+01  3.850e+00 -10.067  < 2e-16 ***
## SWYes       -4.053e+01  4.034e+00 -10.047  < 2e-16 ***
## HI           8.268e-03  1.042e-03   7.932 1.43e-14 ***
## E_INCOME     1.445e-03  4.089e-04   3.533 0.000450 ***
## S_POP        4.185e-06  7.176e-07   5.832 9.85e-09 ***
## E_POP        3.779e-06  7.890e-07   4.790 2.21e-06 ***
## SLOTFree    -1.685e+01  4.305e+00  -3.915 0.000103 ***
## GATEFree    -2.122e+01  4.399e+00  -4.823 1.88e-06 ***
## DISTANCE     7.367e-02  2.870e-03  25.666  < 2e-16 ***
## PAX         -7.619e-04  1.492e-04  -5.107 4.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35.41 on 499 degrees of freedom
## Multiple R-squared:  0.7803, Adjusted R-squared:  0.7759
## F-statistic: 177.2 on 10 and 499 DF,  p-value: < 2.2e-16
```

```r
air.lm.pred <- predict(air.lm, valid.df)
accuracy(air.lm.pred, valid.df$FARE)
```

```
##                ME    RMSE      MAE       MPE     MAPE
## Test set 3.06081 36.8617 27.70568 -5.938062 21.62142
```

By running backward seection using step AIC function, we get the best model with 10 predictors which are VACATION, SW, HI, E_INCOME, S_POP, E_POP, SLOT,GATE, DISTANCE and PAX. The role of AIC in this model is that at every step we drop a variable which decreases the considered AIC value the most. Therefore at every step we lose a less significant predictor. In first step we eliminated COUPON, in the second we eliminated S_INCOME and in the third step we eliminated NEW predictor.