MSCI 446 - Assignment 2 - Question 3

M. Harper, H. Gomaa, K. Morris

20/02/2021

Include Packages

```
library('tidyverse')
library('caret')
library('gridExtra')
library('plotly')
library('ISLR')
library('AmesHousing')
library('leaps')

theme_set(theme_classic())
```

Import Dataset

```
ames <- AmesHousing::make_ames()
numericVars <- ames %>% summarize_all(is.numeric) %>% unlist()
ames <- ames[, numericVars]
head(ames)
```

```
## # A tibble: 6 x 35
##   Lot_Frontage Lot_Area Year_Built Year_Remod_Add Mas_Vnr_Area BsmtFin_SF_1
##          <dbl>    <int>      <int>          <int>        <dbl>        <dbl>
## 1          141    31770       1960           1960          112            2
## 2           80    11622       1961           1961            0            6
## 3           81    14267       1958           1958          108            1
## 4           93    11160       1968           1968            0            1
## 5           74    13830       1997           1998            0            3
## 6           78     9978       1998           1998           20            3
## # ... with 29 more variables: BsmtFin_SF_2 <dbl>, Bsmt_Unf_SF <dbl>,
## #   Total_Bsmt_SF <dbl>, First_Flr_SF <int>, Second_Flr_SF <int>,
## #   Low_Qual_Fin_SF <int>, Gr_Liv_Area <int>, Bsmt_Full_Bath <dbl>,
## #   Bsmt_Half_Bath <dbl>, Full_Bath <int>, Half_Bath <int>,
## #   Bedroom_AbvGr <int>, Kitchen_AbvGr <int>, TotRms_AbvGrd <int>,
## #   Fireplaces <int>, Garage_Cars <dbl>, Garage_Area <dbl>, Wood_Deck_SF <int>,
## #   Open_Porch_SF <int>, Enclosed_Porch <int>, Three_season_porch <int>,
## #   Screen_Porch <int>, Pool_Area <int>, Misc_Val <int>, Mo_Sold <int>,
## #   Year_Sold <int>, Sale_Price <int>, Longitude <dbl>, Latitude <dbl>
```

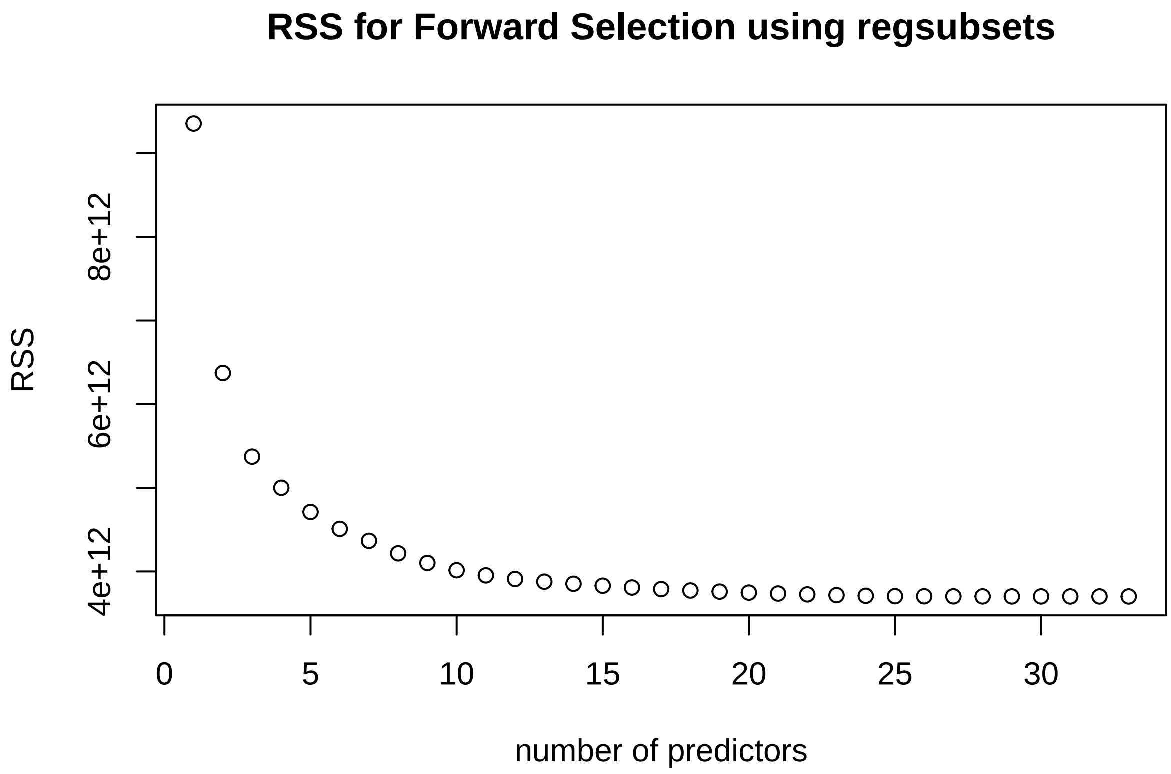
Forward Selection

```
NumCols <- ncol(ames)
res <- regsubsets(Sale_Price ~., data=ames, method='forward', nvmax=NumCols)
```

```
## Reordering variables and trying again:
```

```
smm <- summary(res)
```

```
plot(smm$rss, main="RSS for Forward Selection using regsubsets",
     xlab="number of predictors",ylab="RSS")
```



# Find number of predictors for smallest RSS value:

```
which.min(smm$rss)
```

```
## [1] 33
```

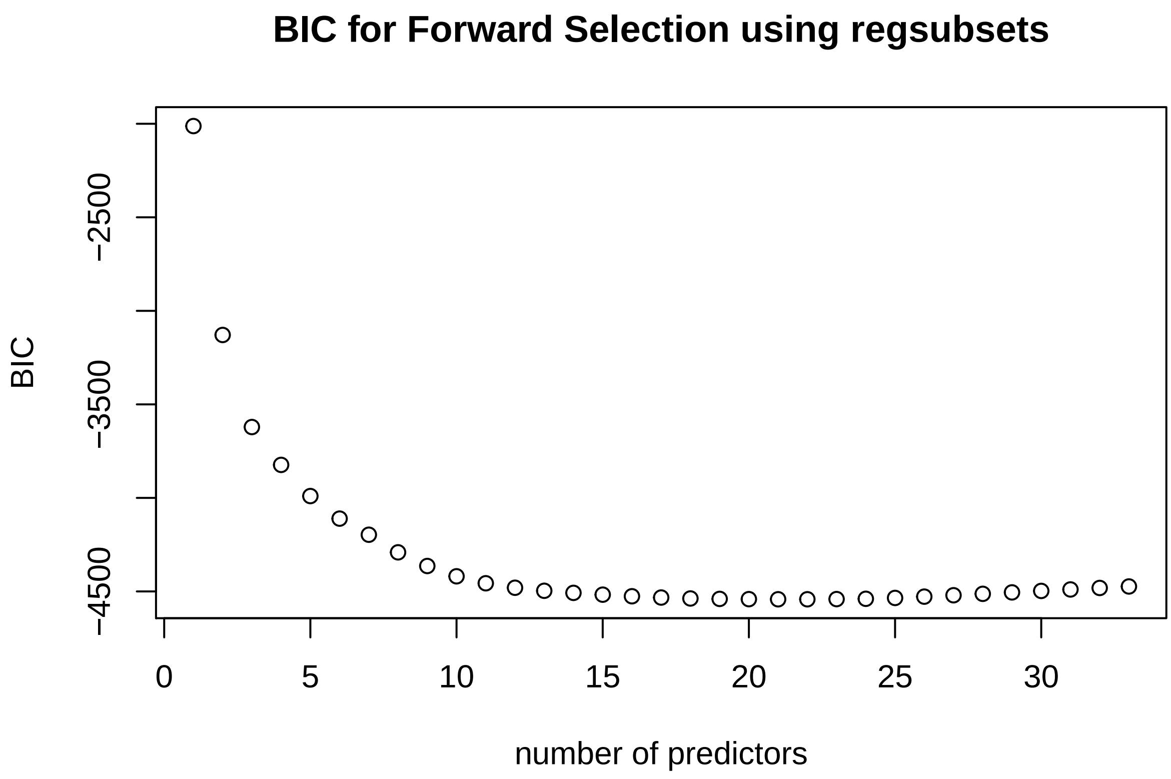
As seen above, using 33 predictors gives the smallest RSS value for Forward Selection. The values of these 33 parameters are shown below:

```
coef(res,33)
```

```
##        (Intercept)       Lot_Frontage           Lot_Area         Year_Built 
##      -1.142977e+07       8.737532e+01       3.141331e-01       3.845931e+02 
##     Year_Remod_Add       Mas_Vnr_Area       BsmtFin_SF_1       BsmtFin_SF_2 
##       5.129858e+02       3.794721e+01       3.002994e+02      -1.338433e+01 
##        Bsmt_Unf_SF      Total_Bsmt_SF       First_Flr_SF    Low_Qual_Fin_SF 
##      -1.337146e+01       3.759189e+01       3.554565e-01      -4.417005e+01 
##     Bsmt_Full_Bath     Bsmt_Half_Bath          Full_Bath          Half_Bath 
##       6.504458e+03      -1.883312e+03       1.949198e+03      -3.471763e+03 
##      Bedroom_AbvGr      Kitchen_AbvGr      TotRms_AbvGrd         Fireplaces 
##      -1.034286e+04      -3.360632e+04       4.068734e+03       7.084818e+03 
##        Garage_Cars        Garage_Area       Wood_Deck_SF      Open_Porch_SF 
##       7.737977e+03       2.082670e+01       2.430170e+01      -4.100172e+00 
##     Enclosed_Porch Three_season_porch       Screen_Porch          Pool_Area 
##       2.974408e+01       8.723251e+00       6.200042e+01      -6.447100e+01 
##           Misc_Val            Mo_Sold          Year_Sold          Longitude 
##      -9.497111e+00       2.762025e+01      -9.346976e+02      -1.299076e+04 
##           Latitude        Gr_Liv_Area 
##       2.464128e+05       6.324190e+01
```

Repeat Using BIC Metric

```
plot(smm$bic, main="BIC for Forward Selection using regsubsets",
     xlab="number of predictors",ylab="BIC")
```



```
which.min(smm$bic)
```

```
## [1] 21
```

21 predictors give the smallest BIC value when using Forward Selection. The reason fewer predictors optimize BIC as apposed to the 33 predictors needed to optimize RSS is due to the BIC calculation penalizing the number of predictors used in a model. The 21 parameter model is shown below:

```
coef(res, 21)
```

```
##    (Intercept)   Lot_Frontage       Lot_Area     Year_Built Year_Remod_Add
##  -1.804094e+06   9.403297e+01   2.439368e-01   3.616190e+02   5.689112e+02
##   Mas_Vnr_Area   BsmtFin_SF_2    Bsmt_Unf_SF  Total_Bsmt_SF Bsmt_Full_Bath
##   4.363806e+01  -1.280552e+01  -1.309842e+01   4.126980e+01   6.192556e+03
## Bsmt_Half_Bath  Kitchen_AbvGr  TotRms_AbvGrd     Fireplaces    Garage_Cars
##  -4.186852e+03  -3.385257e+04   5.606576e+02   9.867642e+03   1.004416e+04
##    Garage_Area   Wood_Deck_SF  Open_Porch_SF      Pool_Area       Misc_Val
##   2.165199e+01   1.963979e+01   1.895785e+00  -5.499532e+01  -9.029755e+00
##        Mo_Sold    Gr_Liv_Area
##   9.536313e+01   5.928065e+01
```

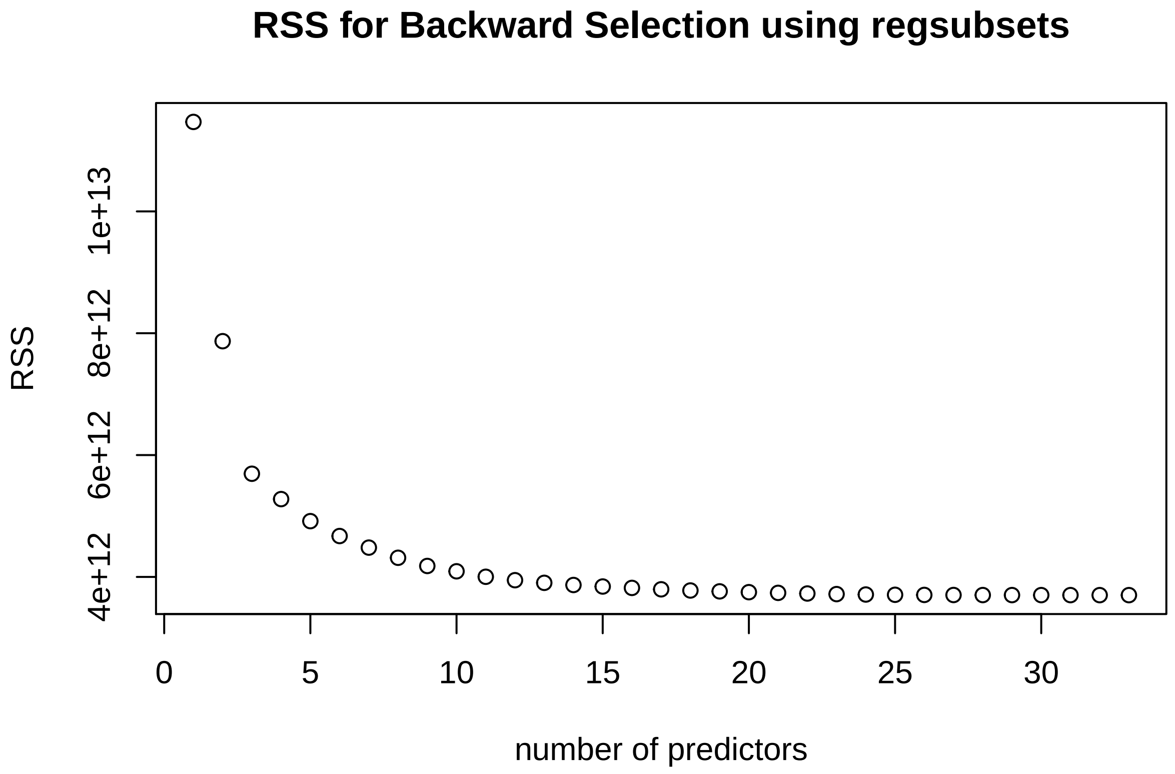
Backward Selection

```
resbkw <- regsubsets(Sale_Price ~., data=ames, method='backward', nvmax=NumCols)
```

```
## Reordering variables and trying again:
```

```
smmbkw <- summary(resbkw)
```

```
plot(smmbkw$rss, main="RSS for Backward Selection using regsubsets",
     xlab="number of predictors",ylab="RSS")
```



```
which.min(smmbkw$rss)
```

```
## [1] 33
```

As seen in Forward Selection, the best RSS value for Backward Selection is at 33 parameters. This model is shown below:

```
coef(resbkw,33)
```

```
##       (Intercept)       Lot_Frontage           Lot_Area         Year_Built
##     -1.170805e+07       8.688692e+01       3.250816e-01       3.915167e+02
##    Year_Remod_Add       Mas_Vnr_Area       BsmtFin_SF_1       BsmtFin_SF_2
##      5.250215e+02       3.754647e+01       1.414811e+02      -1.391134e+01
##      Bsmt_Unf_SF      Total_Bsmt_SF       First_Flr_SF      Second_Flr_SF
##     -1.797736e+01       4.219896e+01       6.308277e+01       6.342274e+01
##   Low_Qual_Fin_SF     Bsmt_Half_Bath          Full_Bath          Half_Bath
##      1.994256e+01      -4.985513e+03       1.170822e+03      -3.889125e+03
##     Bedroom_AbvGr       Kitchen_AbvGr      TotRms_AbvGrd         Fireplaces
##     -1.045933e+04      -3.204082e+04       4.031002e+03       7.123055e+03
##       Garage_Cars        Garage_Area       Wood_Deck_SF      Open_Porch_SF
##      8.075298e+03       1.987748e+01       2.550571e+01      -2.347879e+00
##    Enclosed_Porch Three_season_porch       Screen_Porch          Pool_Area
##      3.067302e+01       9.134332e+00       6.239160e+01      -6.435958e+01
##          Misc_Val            Mo_Sold          Year_Sold          Longitude
##     -9.835393e+00       4.225967e+01      -8.848423e+02      -1.570146e+04
##           Latitude        Gr_Liv_Area
##       2.437618e+05       0.000000e+00
```

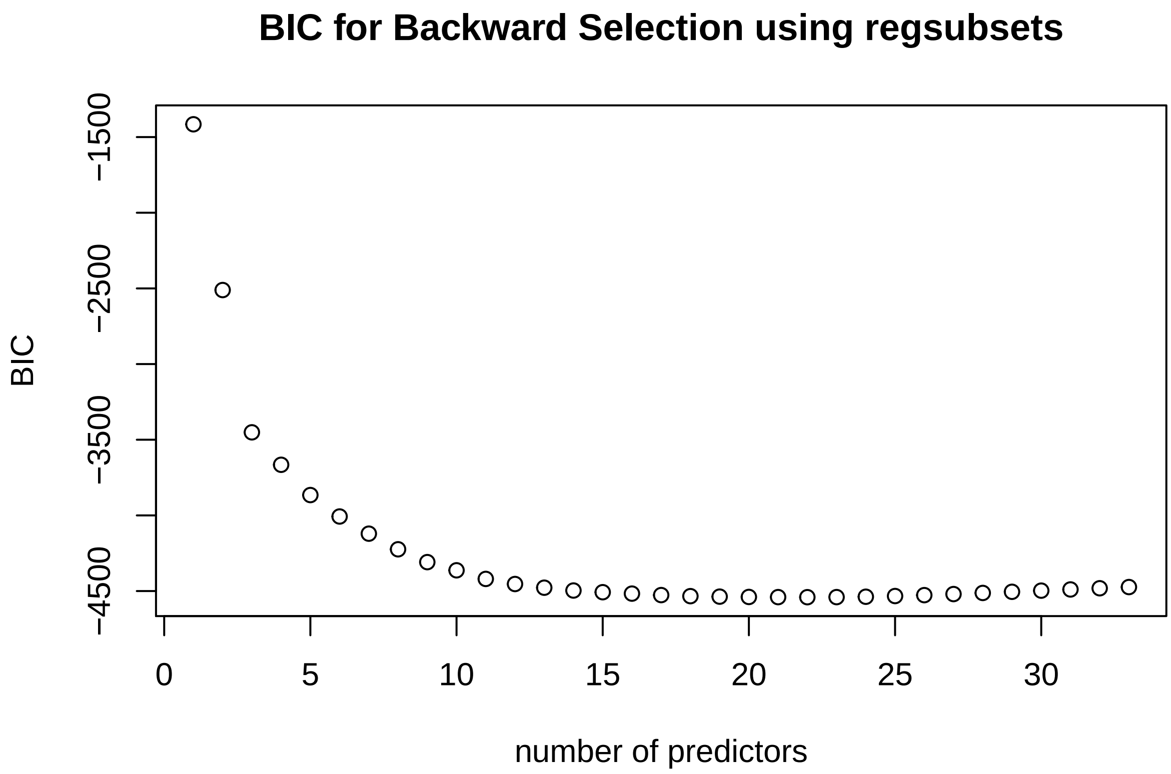
```
names(coef(resbkw,33)) == names(coef(res,33))
```

```
##  [1]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE
## [13] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## [25]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

It appears as though many of the same predictor variables are being used as well, with a few exceptions.

Repeat Using BIC Metric

```
plot(smmbkw$bic, main="BIC for Backward Selection using regsubsets",
     xlab="number of predictors",ylab="BIC")
```



```
which.min(smmbkw$bic)
```

```
## [1] 22
```

The model with minimum BIC value was found to have 22 parameters when using Backward Selection. In comparison, the model with minimum BIC value for Backward Selection had 22 parameters. The predictors and their respective coefficients of the 22 variable model is summarized below.

```
coef(resbkw, 22)
```

```
##    (Intercept)   Lot_Frontage       Lot_Area     Year_Built Year_Remod_Add 
##  -1.816554e+06   8.988071e+01   2.252629e-01   3.568215e+02   5.800933e+02 
##   Mas_Vnr_Area   BsmtFin_SF_2    Bsmt_Unf_SF  Total_Bsmt_SF   First_Flr_SF 
##   4.228330e+01  -1.357431e+01  -1.785513e+01   4.275903e+01   4.067518e+01 
##  Second_Flr_SF Bsmt_Half_Bath   Kitchen_AbvGr  TotRms_AbvGrd     Fireplaces 
##   3.517944e+01  -7.244059e+03  -3.430550e+04   6.485669e+02   9.556181e+03 
##    Garage_Cars    Garage_Area   Wood_Deck_SF  Open_Porch_SF      Pool_Area 
##   1.024918e+04   2.015845e+01   2.092499e+01   3.072430e+00  -5.526520e+01 
##       Misc_Val        Mo_Sold    Gr_Liv_Area 
##  -9.437835e+00   9.540143e+01   2.300514e+01
```