

## Question 3 Subset Selection

### Forward Selection

Data Initialization and setting up the variables

```
ames      <- AmesHousing::make_ames()
numericVars <- ames %>% summarise_all(is.numeric) %>% unlist()
ames      <- ames[, numericVars]
NumCols   <- ncol(ames)

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

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found

## Reordering variables and trying again:

## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to replace is
## not a multiple of replacement length
```

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

```
## [1] 9.354907e+12 6.372705e+12 5.372622e+12 5.000405e+12 4.711132e+12
## [6] 4.509022e+12 4.366282e+12 4.216771e+12 4.101703e+12 4.014448e+12
## [11] 3.952959e+12 3.910112e+12 3.877808e+12 3.852701e+12 3.829707e+12
## [16] 3.808074e+12 3.788825e+12 3.772223e+12 3.759006e+12 3.747105e+12
## [21] 3.736053e+12 3.725905e+12 3.716953e+12 3.708821e+12 3.704526e+12
## [26] 3.703314e+12 3.702500e+12 3.701952e+12 3.701714e+12 3.701525e+12
## [31] 3.701381e+12 3.701365e+12 3.701352e+12
```

```
min_rss <- which.min(smm$rss)
min_bic <- which.min(smm$bic)
```

```
min_rss
```

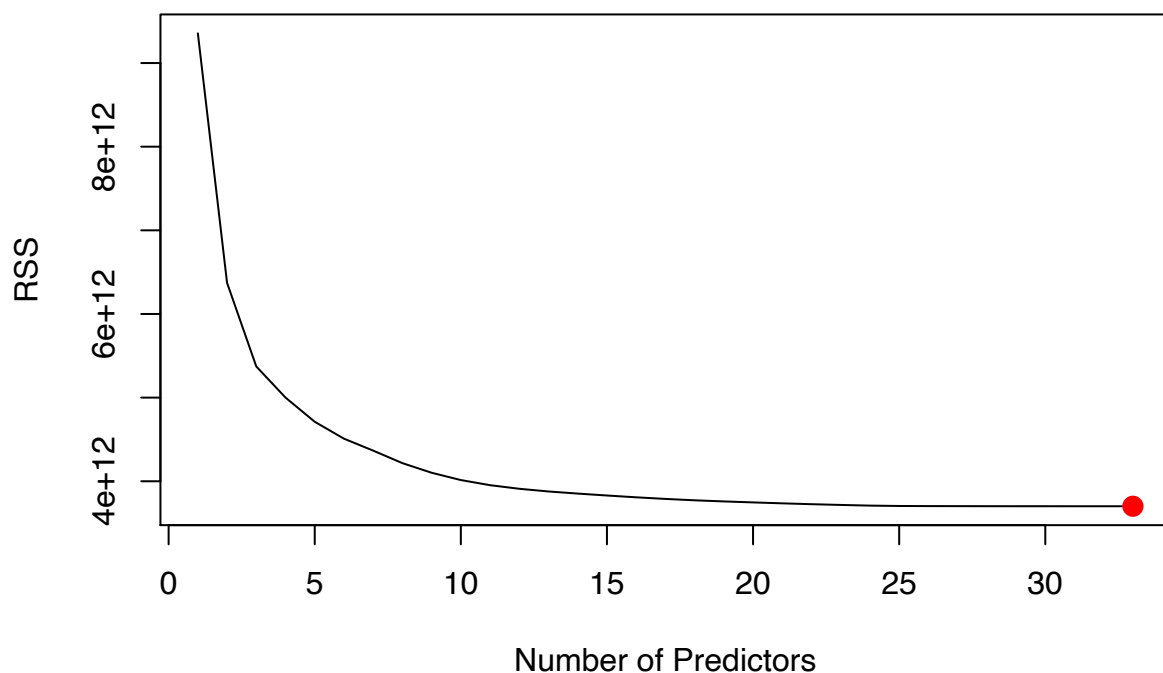
```
## [1] 33
```

```
min_bic
```

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

**RSS Plot (Forward Selection)** Plotting the RSS of each Model (Forward Selection)

```
plot(smm$rss,xlab="Number of Predictors", ylab="RSS", type='l')
points(min_rss, smm$rss[min_rss], col="red", cex=2, pch=20)
```



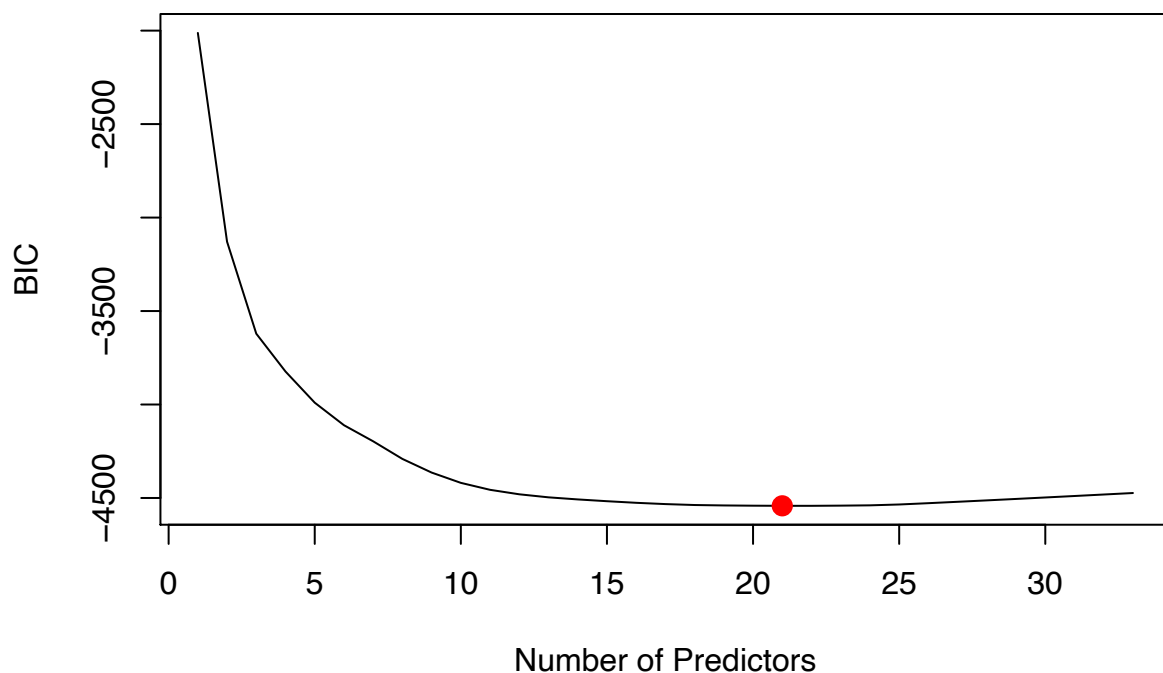
The best model, that is, a model that produces the least RSS is the model that uses 33 predictors. The coefficients of these are as follows:

```
coef(res, min_rss)
```

|    |                |                    |               |                 |
|----|----------------|--------------------|---------------|-----------------|
| ## | (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       |               |                 |

## BIC Plot (Forward Selection) Plotting the BIC of each Model (Forward Selection)

```
plot(smm$bic,xlab="Number of Predictors", ylab="BIC", type='l')
points(min_bic, smm$bic[min_bic], col="red", cex=2, pch=20)
```



The best model, that is, a model that produces the least BIC is the model that uses 21 predictors. The coefficients of these are as follows:

```
coef(res, min_bic)
```

```
##      (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

Data Initialization and setting up the variables

```

ames      <- AmesHousing::make_ames()
numericVars <- ames %>% summarise_all(is.numeric) %>% unlist()
ames      <- ames[, numericVars]
NumCols   <- ncol(ames)

res <- regsubsets(Sale_Price ~ ., data=ames, method = "backward", nvmax=NumCols)

## Warning in leaps.setup(x, y, wt = wt, nbest = nbest, nvmax = nvmax, force.in =
## force.in, : 1 linear dependencies found

## Reordering variables and trying again:

## Warning in rval$lopt[] <- rval$vorder[rval$lopt]: number of items to replace is
## not a multiple of replacement length

```

```

smm <- summary(res)
smm$rss

```

```

## [1] 1.146822e+13 7.869601e+12 5.693659e+12 5.277521e+12 4.915896e+12
## [6] 4.671204e+12 4.481447e+12 4.314304e+12 4.179940e+12 4.092241e+12
## [11] 4.002680e+12 3.946271e+12 3.902998e+12 3.867500e+12 3.842522e+12
## [16] 3.819776e+12 3.796777e+12 3.777550e+12 3.763157e+12 3.750030e+12
## [21] 3.738591e+12 3.727819e+12 3.718299e+12 3.711271e+12 3.707002e+12
## [26] 3.704526e+12 3.703298e+12 3.702482e+12 3.701936e+12 3.701699e+12
## [31] 3.701509e+12 3.701367e+12 3.701352e+12

```

```

min_rss <- which.min(smm$rss)
min_bic <- which.min(smm$bic)

min_rss

```

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

```
min_bic
```

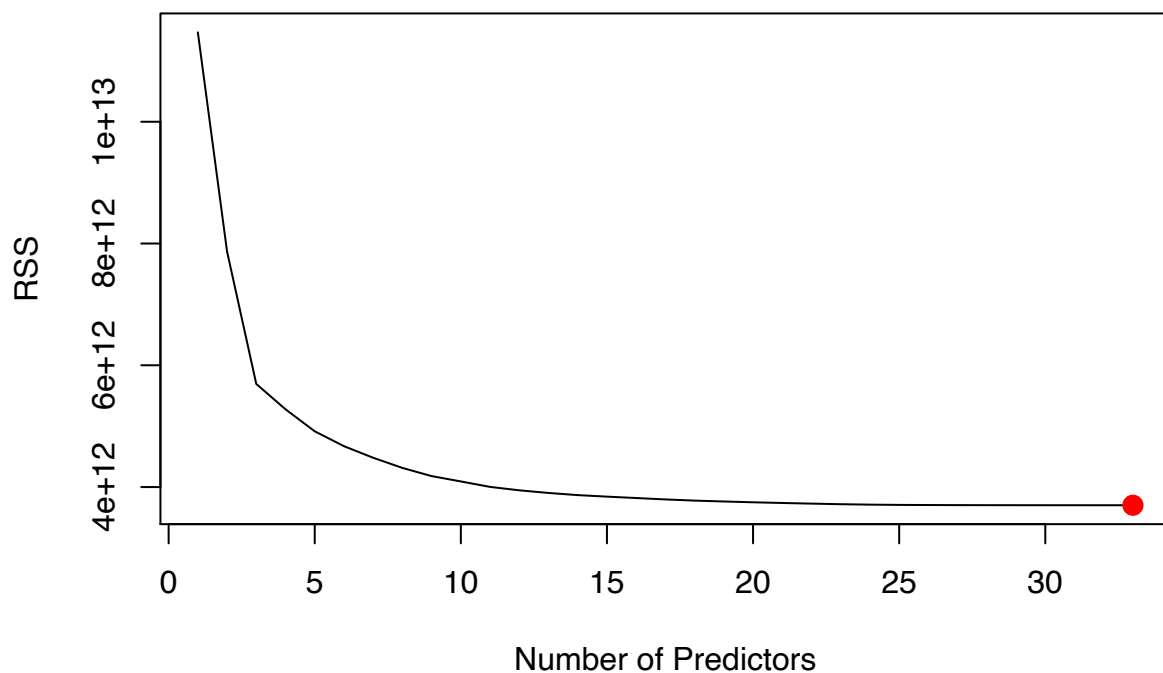
```
## [1] 22
```

**RSS Plot (Backward Selection)** Plotting the RSS of each Model (Backward Selection)

```

plot(smm$rss, xlab="Number of Predictors", ylab="RSS", type='l')
points(min_rss, smm$rss[min_rss], col="red", cex=2, pch=20)

```



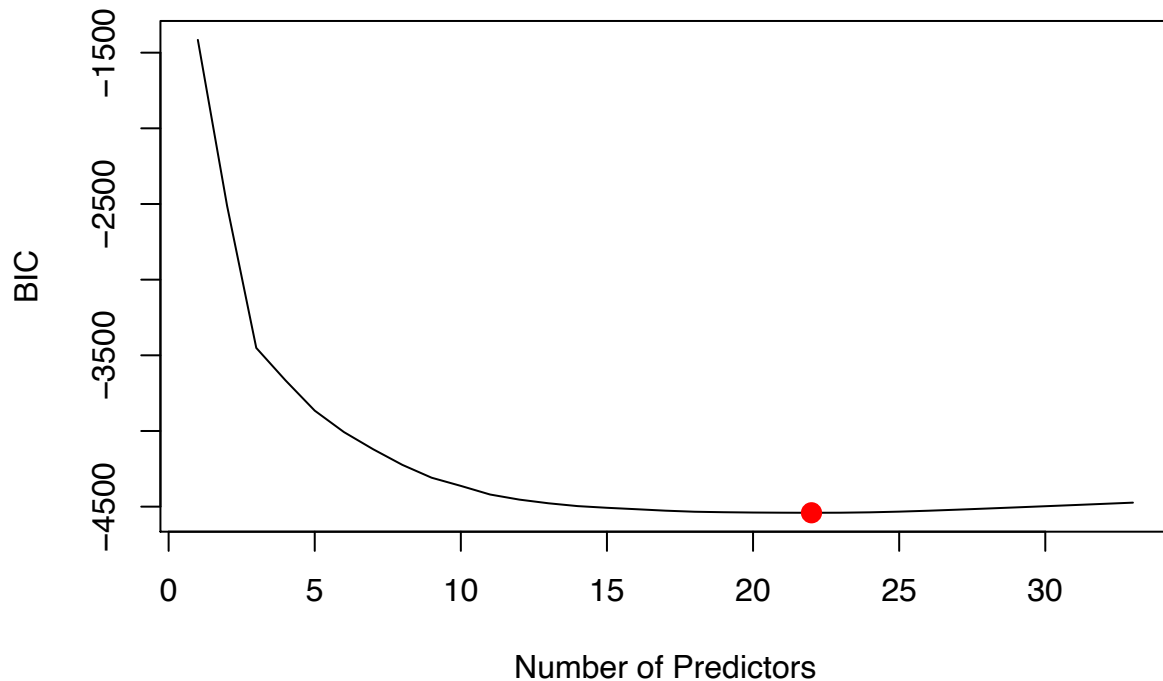
The best model, that is, a model that produces the least RSS is the model that uses 33 predictors. The coefficients of these are as follows:

```
coef(res, min_rss)
```

|    |                 |                    |               |               |
|----|-----------------|--------------------|---------------|---------------|
| ## | (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       |               |               |

**BIC Plot (Backward Selection)** Plotting the BIC of each Model (Backward Selection)

```
plot(smm$bic,xlab="Number of Predictors", ylab="BIC", type='l')
points(min_bic, smm$bic[min_bic], col="red", cex=2, pch=20)
```



The best model, that is, a model that produces the least BIC is the model that uses 22 predictors. The coefficients of these are as follows:

```
coef(res, min_bic)
```

```
##      (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
```