SimulatedDataSet-HW2-P3

Jeetendra Gan | jgan2 | 50325023

p = Number of features(20)

n = Number of observations(1000)

I have generated 14 beta values from a normal distribution with mean equal to 2 and standard deviation of 3. The rest of the 4 beta values are set to 0 by default. One of the beta values is set to 6 on purpose(To have at least one highly correlated predictor with response)

Here are the true model coefficients.

```
##  [1]  4.9054679  4.6589676  7.8961034  0.0000000  2.8245505  0.0000000
##  [7]  0.2698100 -2.2529101  0.0000000  0.0000000 -3.3439323  0.1410946
## [13]  6.0000000  0.0000000  3.5256386  4.1645701  2.0916208  3.8572967
## [19]  4.8486723  2.0630175
```

I have randomly generated the range of each predictor. Each predictor will take on values from the following range.

```
##       [,1] [,2]
##  [1,]  886 2675
##  [2,] 1943 3300
##  [3,] 2721 3043
##  [4,]  564 1158
##  [5,]   71 1727
##  [6,]  951 3438
##  [7,] 2294 3405
##  [8,] 2831 3108
##  [9,] 1095 2901
## [10,] 2088 3124
## [11,]  982 3125
## [12,] 3230 3381
## [13,]  372 1662
## [14,]   95 2273
## [15,] 1992 3710
## [16,] 2599 3896
## [17,] 2689 3087
## [18,]  863 1908
## [19,] 1621 3772
## [20,] 1260 3842
```

Using the ranges of predictors, I have generated 1000 values for each predictor.

This is the head of predictors.

```
##           [,1]     [,2]     [,3]      [,4]      [,5]     [,6]     [,7]
## [1,] 1400.4762 2314.306 2772.415  686.2612  575.1927 1569.343 2783.130
## [2,] 2296.2779 2748.877 2767.534 1123.8682 1450.1479 3411.795 2735.503
## [3,] 1617.6597 2160.371 2769.036  789.3183 1054.0803 2734.482 2706.789
## [4,] 2465.7181 3101.105 2886.648  935.9866 1407.7176 2571.848 2881.506
## [5,] 2568.4960 3093.382 2879.690  673.0004  557.9481 1549.263 2375.976
## [6,]  967.5006 2591.492 2919.462  955.5693  304.6371 1166.854 3090.420
##          [,8]     [,9]    [,10]    [,11]    [,12]     [,13]     [,14]
## [1,] 3088.704 2138.706 2914.691 1647.598 3243.190  684.4441 1664.3372
## [2,] 2994.719 1977.855 2998.785 1677.446 3292.386  991.2411  496.2143
## [3,] 2853.089 2435.396 2339.191 2846.955 3244.807 1456.2251 1013.4784
## [4,] 3065.070 1502.450 2667.672 1686.348 3368.339 1651.7793 1489.1480
## [5,] 2839.250 2447.497 3025.345 1251.378 3239.756  687.0627 2194.3118
## [6,] 3056.661 2584.713 2096.494 1745.383 3288.882 1175.7200 1341.8869
##         [,15]    [,16]    [,17]     [,18]    [,19]    [,20]
## [1,] 2082.757 3424.971 2965.739 1497.7724 3714.556 2209.488
## [2,] 2152.835 2718.505 2839.575  974.6648 2120.978 2379.724
## [3,] 3104.789 3626.618 3072.949 1402.7762 2199.169 2399.989
## [4,] 2464.055 3821.221 2898.886 1566.5480 1861.816 3345.382
## [5,] 3660.918 2991.595 2706.639 1383.4648 2238.046 1906.201
## [6,] 3366.628 3329.221 2956.716 1525.2486 1844.132 1675.233
```

The error values are generated from a normal distribution with mean = 0, and standard deviation = 1600. The following cell shows the head of errors.

```
## [1]  342.3400  767.4530  140.5259  710.1736 -580.5407  196.2784
```

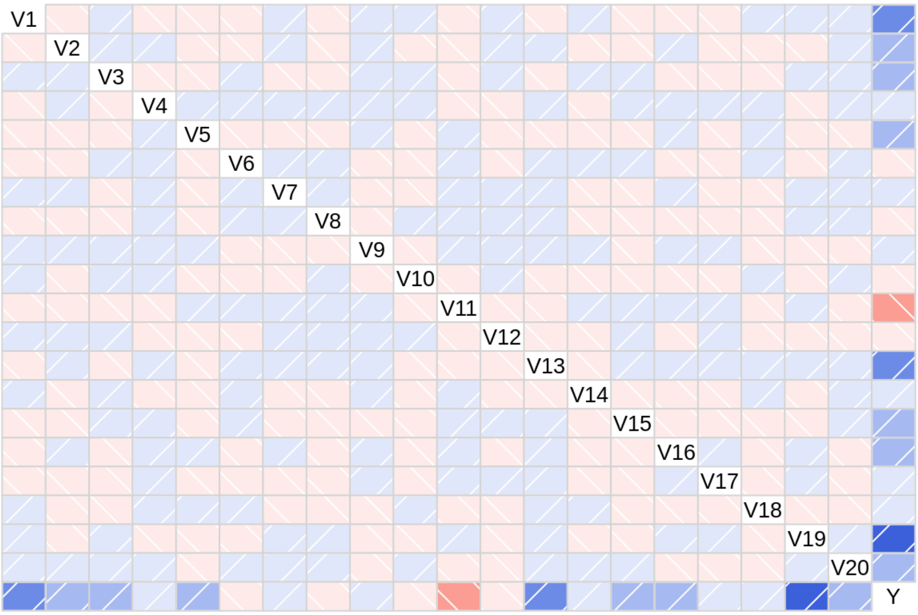
Why did I choose such a large value of standard deviation for error?

1. The range of values taken by predictors are large
2. To add variance to the data so that the coefficients with 0 values are not accidentally valued by the model.

This is the head of the response variable.

```
##           [,1]
## [1,]  90513.70
## [2,]  89385.15
## [3,]  90462.92
## [4,] 105278.13
## [5,]  96532.39
## [6,]  86757.79
```

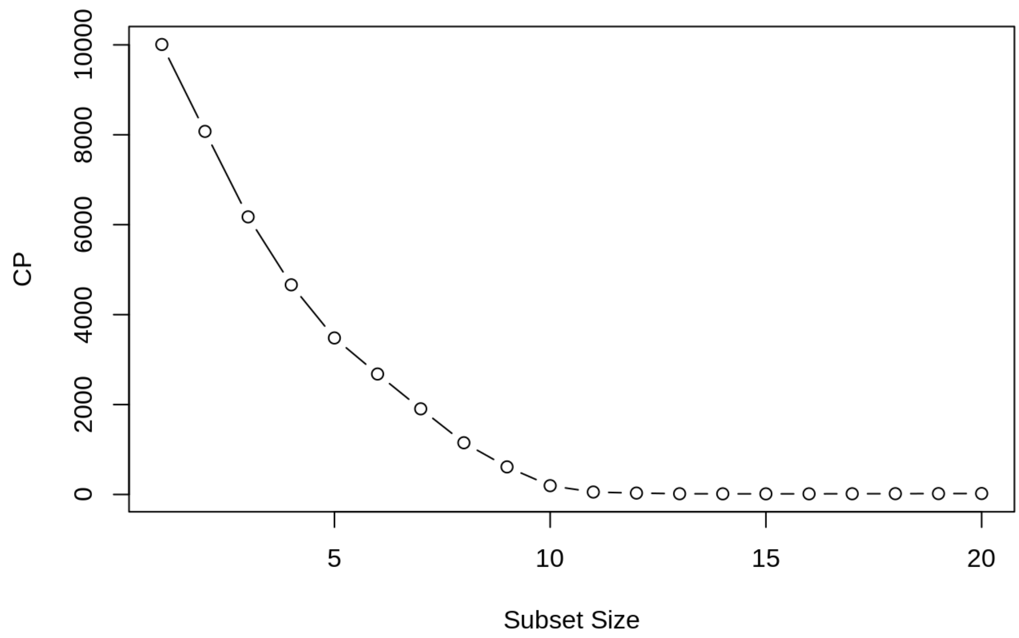
Following is the correlation matrix of the complete data.

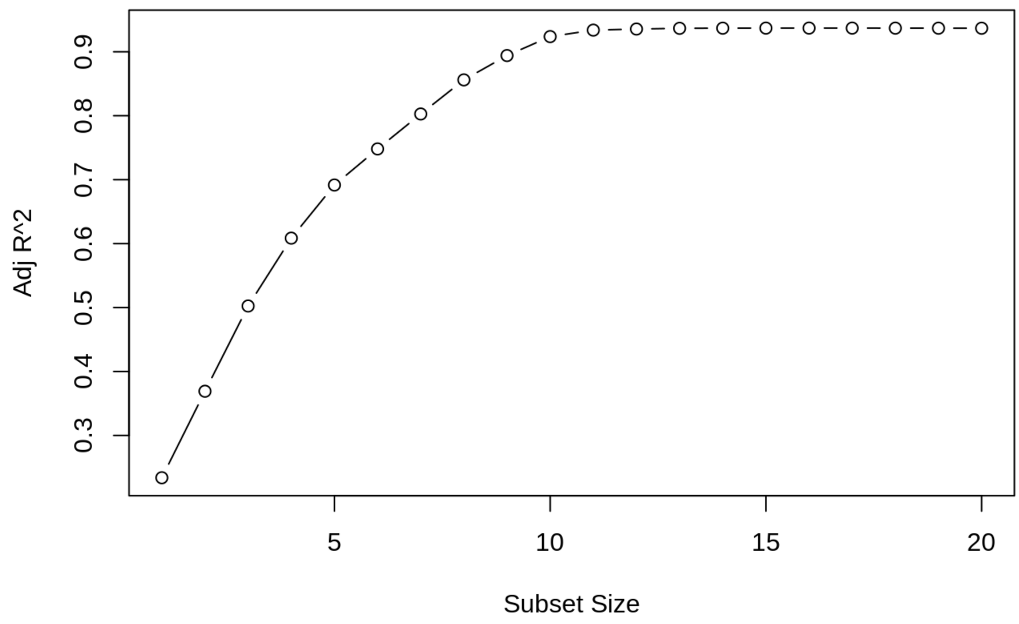


A few variables are strongly-positively correlated with the response, and only a few are strongly-weakly correlated. There is a little to no correlation between some variables and response. Also, the predictors themselves are not correlated to each other.

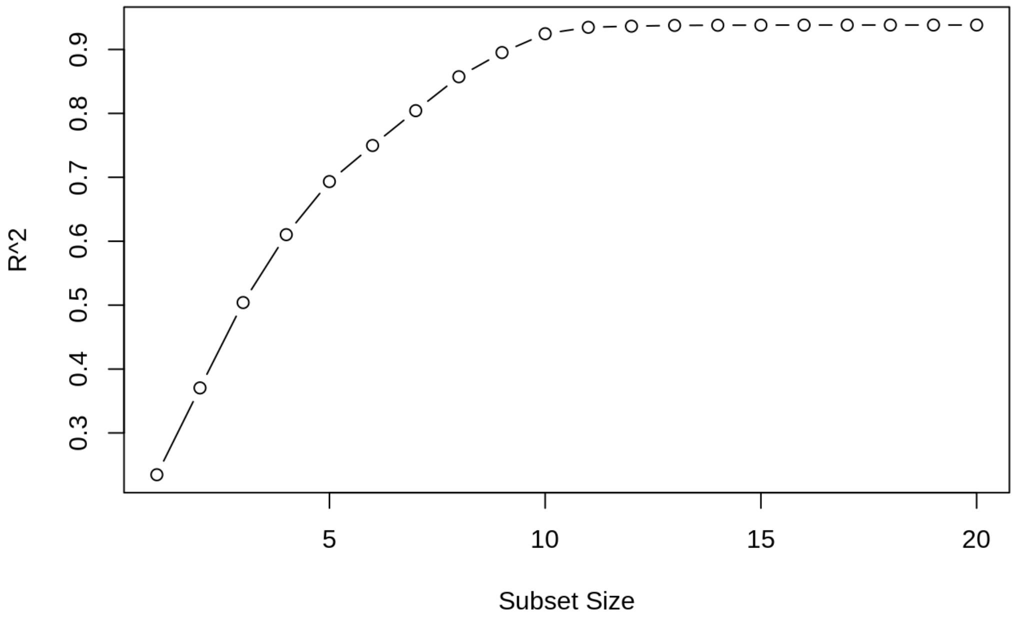
Results of subset selection

Following are a few graps for each of the subsets

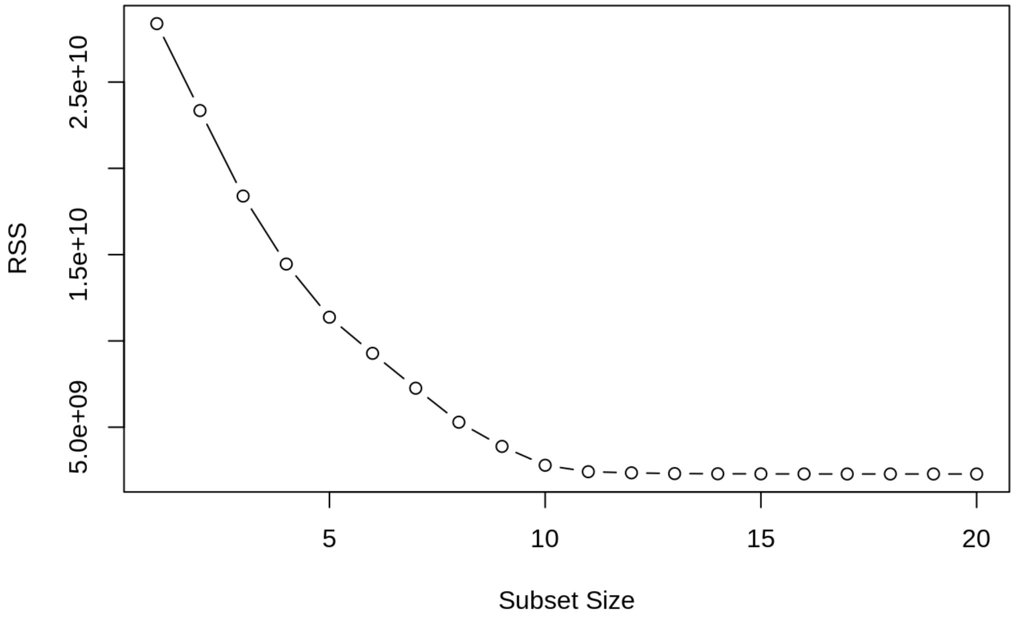




The graph below shows that as the subset size increases, the CP and adjusted R-square, which estimate the test error also decreases.

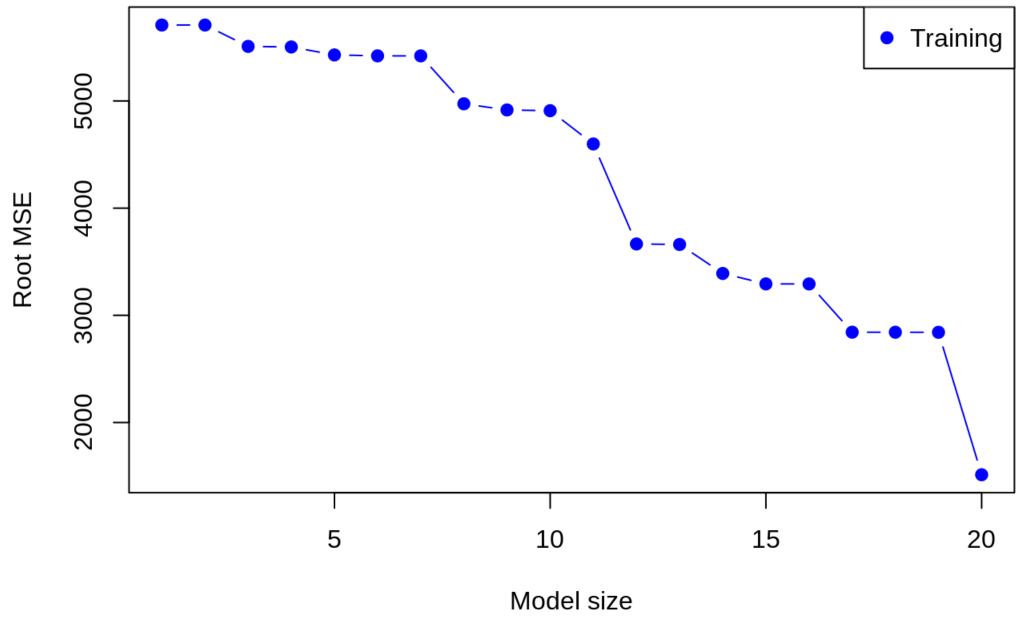


R-square increases as the number of variables are added to the model. This is explained because addition of variables helps us explain variance in the **training set**.

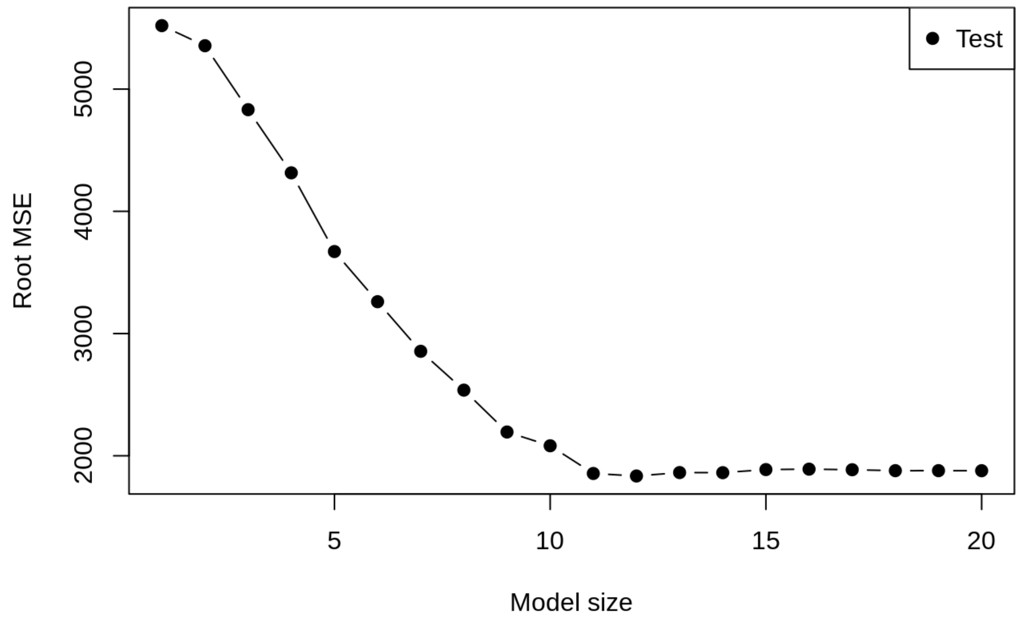


RSS also decreases with an increase in the number of predictors. We essentially overfit the data as we add more variables, hence the RSS gets reduced.

Following graph shows the root mean squared error for each model on the training data. The decrease in RMSE is expected as the RSS also decreases, which indicates an increase in variance which can further lead to overfitting.



The graph below shows the root mean squared error on the test data.



For which model size does the test set MSE take on its minimum value?

```
## [1] 12
```

There is not much decrease in the RMSE after 12, actually it increases slightly and does not drop back. The model with 11 predictors also has the same RMSE. It might perform similar to the one with 12.

The best model has the following coefficient values.

```
##   (Intercept)           V1          V2          V3          V5
## 14411.661311     4.990391    4.834792    6.727235    2.722149
##           V8          V11         V13         V15         V16
##    -3.371737    -3.278417    5.979938    3.324505    3.965697
##          V18          V19         V20
##     3.644985     4.916581    2.147284
```

For a model with 11 coefficients has the following values.

```
## (Intercept)          V1          V2          V3          V5         V11
## 3801.812829    4.984949    4.854083    6.899378    2.734663   -3.298878
##         V13         V15         V16         V18         V19         V20
##    5.988277    3.329127    3.993191    3.665485    4.907054    2.141248
```

It drops the predictor V8. The true value of the the coefficient is -2.252910. Model with 12 features does a better job of estimating the true model.

The true model is:

```
## V1: 4.905468
## V2: 4.658968
## V3: 7.896103
## V4: 0.000000
## V5: 2.824551
## V6: 0.000000
## V7: 0.269810
## V8: -2.252910
## V9: 0.000000
## V10: 0.000000
## V11: -3.343932
## V12: 0.141095
## V13: 6.000000
## V14: 0.000000
## V15: 3.525639
## V16: 4.164570
## V17: 2.091621
## V18: 3.857297
## V19: 4.848672
## V20: 2.063017
```

There is a striking resemblance between the true coefficients and the coefficients for the model with least test error. Best subset selection has removed each of the 4 zero coefficients that I added before and it has also removed those coefficients that have a very small value. Also, the coefficient values are very close to the true model.

V17 which does not have a very small coefficient, has been removed in the model with 12 coefficients. It has been added to the model with 13 coefficients as shown below.

```
## (Intercept)          V1          V2          V3          V5          V8
## 8122.339020    4.995348    4.843370    6.821939    2.741881   -3.234408
##         V11         V13         V15         V16         V17         V18
##   -3.304415    5.942521    3.349906    3.963433    1.926285    3.701923
##         V19         V20
##    4.889109    2.150658
```

We could also choose this model for deployment. But there will be an obvious bias towards the choice as we already know the true values of the coefficients. In practice we would choose a simpler model.