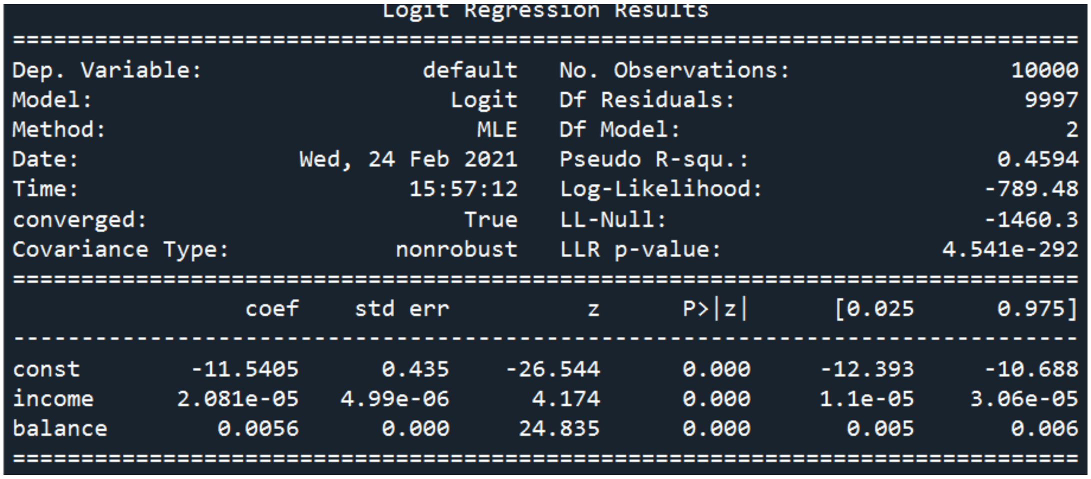
Jayoung Kang

Problem Set 3

Chapter 5: Question 6

a. *Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with income and balance in a multiple logistic regression model that uses both predictors.*
   - Income std error: 4.99e-06
   - Balance std error: 0.000
   - Intercept std error: 0.435



```
                           Logit Regression Results
==============================================================================
Dep. Variable:                default   No. Observations:                10000
Model:                          Logit   Df Residuals:                     9997
Method:                           MLE   Df Model:                            2
Date:                Wed, 24 Feb 2021   Pseudo R-squ.:                  0.4594
Time:                        15:57:12   Log-Likelihood:                -789.48
converged:                       True   LL-Null:                       -1460.3
Covariance Type:            nonrobust   LLR p-value:                 4.541e-292
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -11.5405      0.435    -26.544      0.000     -12.393     -10.688
income      2.081e-05   4.99e-06      4.174      0.000      1.1e-05    3.06e-05
balance        0.0056      0.000     24.835      0.000       0.005       0.006
==============================================================================
```

b. *Write a function, boot.fn(), that takes as input the Default data set as well as an index of the observations, and that outputs the coefficient estimates for income and balance in the multiple logistic regression model.*
   - Income coefficient estimate: -0.00013
   - Balance coefficient estimate: 0.00047

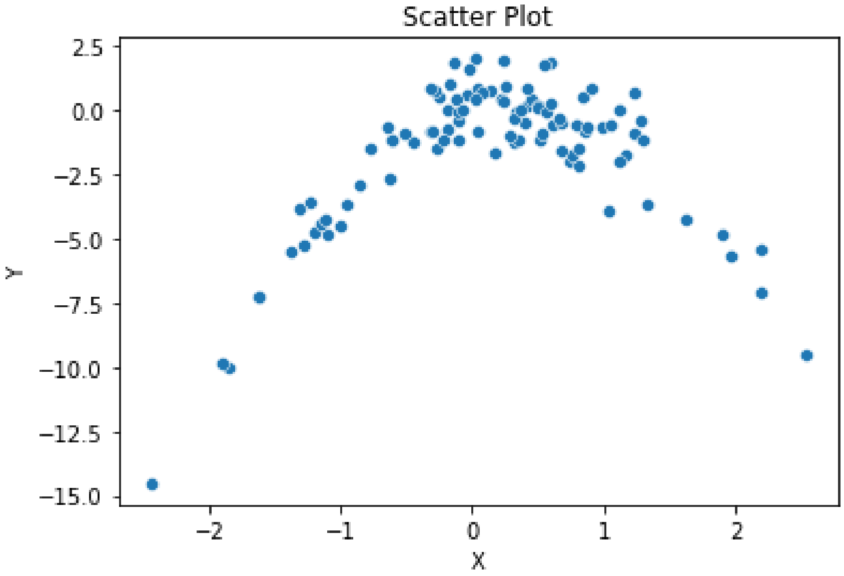
c. *Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for income and balance.*
   - Income coefficient std error: 7.82263e-05
   - Balance coefficient std error: 0.00265

d. *Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.*
   - The standard errors obtained from the bootstrap are similar but still slightly larger than the estimated std errors from part a). This is somewhat expected because we have bootstrapped for only 1000 times. If we increase the number of times we bootstrap we will likely get a more precise estimate of the std errors.

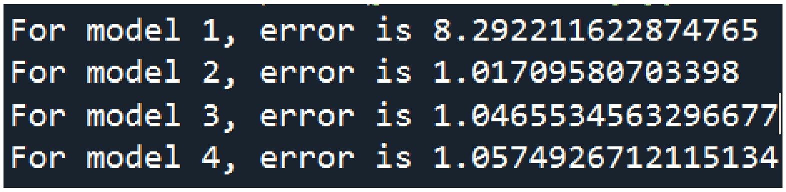
Chapter 5: Question 8

a. *In this data set, what is n and what is p? Write out the model used to generate the data in equation form*
   - n: n is 100, the number of examples
   - p: p is 2, the number of features x and x^2
   - model: $y = x -2x^2 + \varepsilon$

b. *Create a scatterplot of X against Y . Comment on what you find*
   - The scatter plot takes on a concave parabola shape and although it is not a perfect parabola due to the randomness in the generated data, the parabola indicates the quadratic relationship between x and y.
   - We can see that there is a positive correlation between x and y until x=0.25 and a negative correlation for values of x above 0.25. The precise value for where x peaks can be found through finding the FOC of the model.



c. *Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:*
   - The LOOCV error is really high when the model is just a regular linear regression (model 1). It goes down as we add the squared term (model 2) but then increases slightly as we add higher powers to the model. In this case it is likely because the true relationship we assigned only has a squared term. But the LOOCV error could go up if the model is overfitting as well.

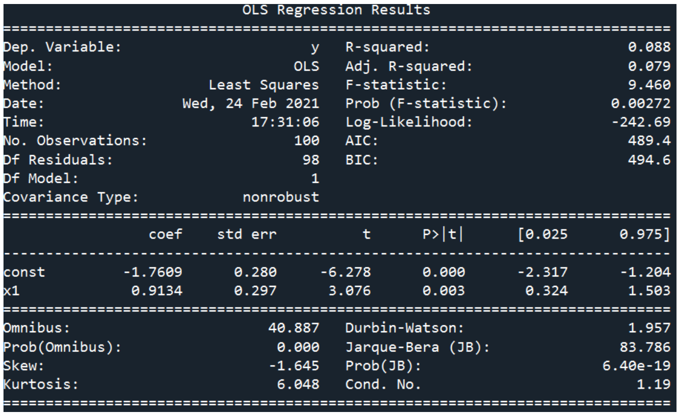


```
For model 1, error is 8.292211622874765
For model 2, error is 1.01709580703398
For model 3, error is 1.0465534563296677
For model 4, error is 1.0574926712115134
```

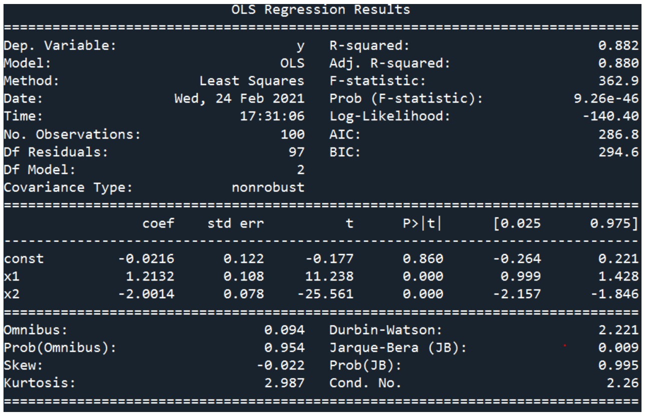
d. *Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?*
   - The results are exactly the same as part c).
   - This is because there is no random sampling in LOOCV. Every time we will be fitting n models such that each time the model will be trained on n - 1 observations and then tested on a validation set with that left out observation.

e. *Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.*
   - The best error value is for degree = 2, as expected. Since this is from a simulated data, we know that the real relationship is also quadratic so we would expect the smallest LOOCV error from model 2.

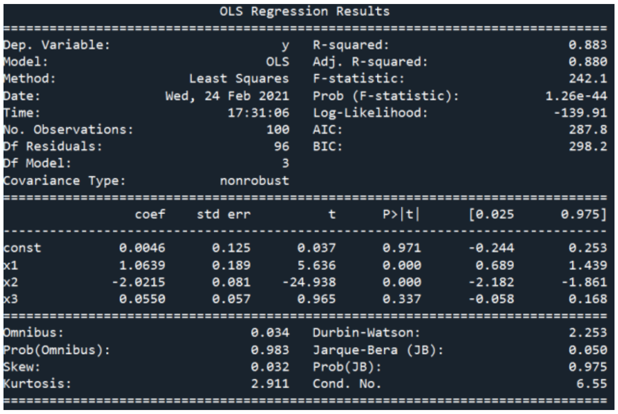
f. *Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (c) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?*
    - We can see that from all the regression results only x1 and x2 are statistically significant. This agrees with the conclusions drawn from the CV results because it indicates that the best model is the quadratic model.
    - The adjusted R-squared values also agree with our previous finding because it does not increase after model 2, indicating that adding the terms with the additional powers does not meaningfully increase the predictive power of the model.



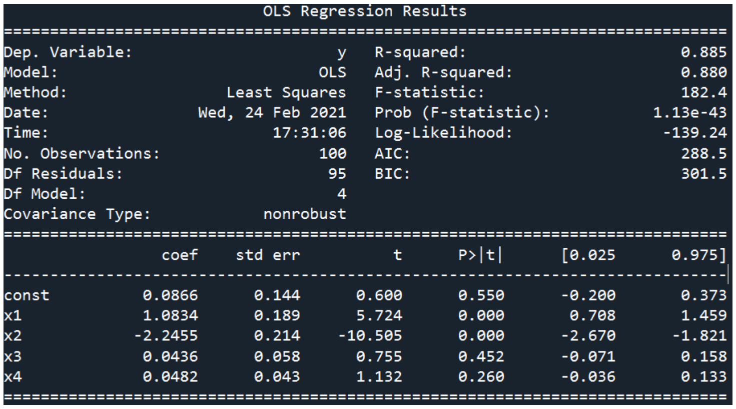
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.088
Model:                            OLS   Adj. R-squared:                  0.079
Method:                 Least Squares   F-statistic:                     9.460
Date:                Wed, 24 Feb 2021   Prob (F-statistic):            0.00272
Time:                        17:31:06   Log-Likelihood:                -242.69
No. Observations:                 100   AIC:                             489.4
Df Residuals:                      98   BIC:                             494.6
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.7609      0.280     -6.278      0.000      -2.317      -1.204
x1             0.9134      0.297      3.076      0.003       0.324       1.503
==============================================================================
Omnibus:                       40.887   Durbin-Watson:                   1.957
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               83.786
Skew:                          -1.645   Prob(JB):                     6.40e-19
Kurtosis:                       6.048   Cond. No.                         1.19
==============================================================================
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.882
Model:                            OLS   Adj. R-squared:                  0.880
Method:                 Least Squares   F-statistic:                     362.9
Date:                Wed, 24 Feb 2021   Prob (F-statistic):           9.26e-46
Time:                        17:31:06   Log-Likelihood:                -140.40
No. Observations:                 100   AIC:                             286.8
Df Residuals:                      97   BIC:                             294.6
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.0216      0.122     -0.177      0.860      -0.264       0.221
x1             1.2132      0.108     11.238      0.000       0.999       1.428
x2            -2.0014      0.078    -25.561      0.000      -2.157      -1.846
==============================================================================
Omnibus:                        0.094   Durbin-Watson:                   2.221
Prob(Omnibus):                  0.954   Jarque-Bera (JB):                0.009
Skew:                          -0.022   Prob(JB):                        0.995
Kurtosis:                       2.987   Cond. No.                         2.26
==============================================================================
```



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.883
Model:                            OLS   Adj. R-squared:                  0.880
Method:                 Least Squares   F-statistic:                     242.1
Date:                Wed, 24 Feb 2021   Prob (F-statistic):           1.26e-44
Time:                        17:31:06   Log-Likelihood:                -139.91
No. Observations:                 100   AIC:                             287.8
Df Residuals:                      96   BIC:                             298.2
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0046      0.125      0.037      0.971      -0.244       0.253
x1             1.0639      0.189      5.636      0.000       0.689       1.439
x2            -2.0215      0.081    -24.938      0.000      -2.182      -1.861
x3             0.0550      0.057      0.965      0.337      -0.058       0.168
==============================================================================
Omnibus:                        0.034   Durbin-Watson:                   2.253
Prob(Omnibus):                  0.983   Jarque-Bera (JB):                0.050
Skew:                           0.032   Prob(JB):                        0.975
Kurtosis:                       2.911   Cond. No.                         6.55
==============================================================================
```

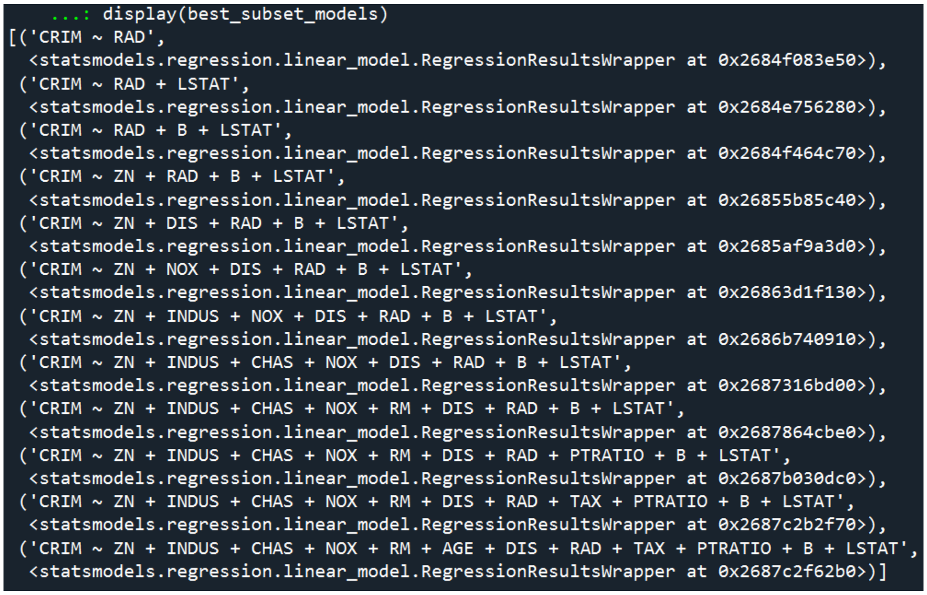


```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.885
Model:                            OLS   Adj. R-squared:                  0.880
Method:                 Least Squares   F-statistic:                     182.4
Date:                Wed, 24 Feb 2021   Prob (F-statistic):           1.13e-43
Time:                        17:31:06   Log-Likelihood:                -139.24
No. Observations:                 100   AIC:                             288.5
Df Residuals:                      95   BIC:                             301.5
Df Model:                           4
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          0.0866      0.144      0.600      0.550      -0.200       0.373
x1             1.0834      0.189      5.724      0.000       0.708       1.459
x2            -2.2455      0.214    -10.505      0.000      -2.670      -1.821
x3             0.0436      0.058      0.755      0.452      -0.071       0.158
x4             0.0482      0.043      1.132      0.260      -0.036       0.133
==============================================================================
```

Chapter 6: Question 11

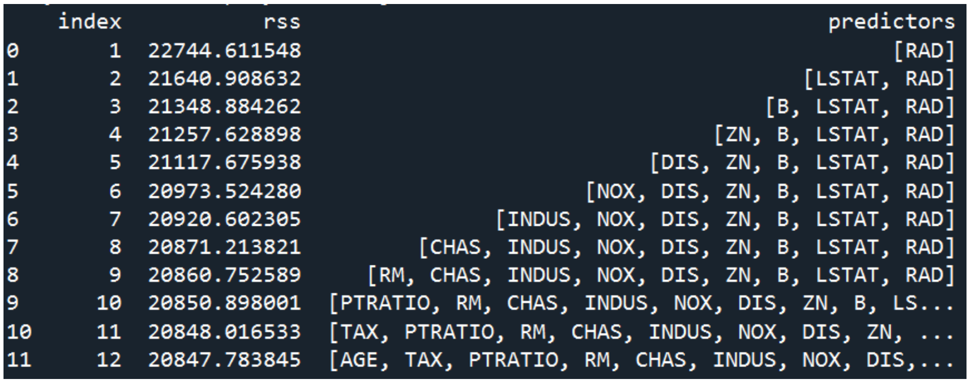
a. *Present and discuss results for the approaches that you consider: best subset, forward stepwise, & backwards stepwise selection*
    - For best subset, forward stepwise and backward stepwise we can see that when they are choosing the best among the models with the given number of predictors, they are choosing the same variables. By best among the models with the given number of predictors, it means that they are choosing the model with the lowest RSS. For example, when forward stepwise chooses the best model with 5 predictors, the chose predictors are DIS, ZN, B, LSTAT, RAD, which is identical to the predictors from the backward stepwise model chosen with 5 predictors and also the same for the best subset model. This is the case for all models with the same number of predictors.

<Best subset>



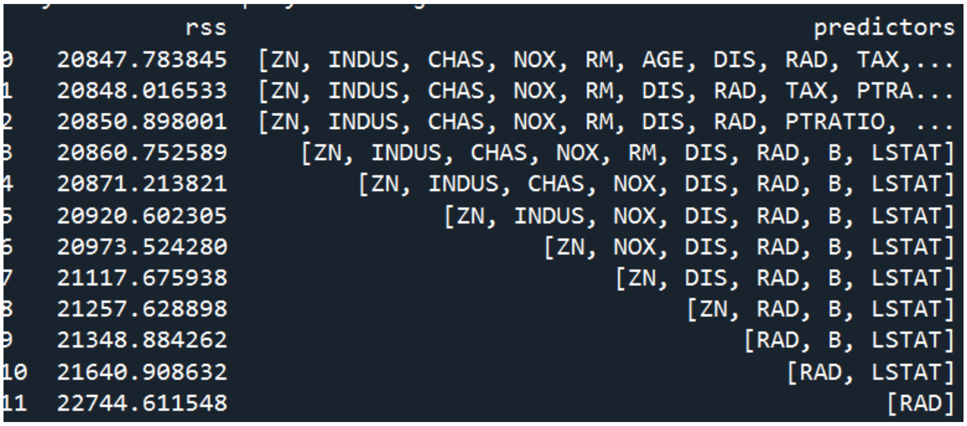
```
    ...: display(best_subset_models)
[('CRIM ~ RAD',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2684f083e50>),
 ('CRIM ~ RAD + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2684e756280>),
 ('CRIM ~ RAD + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2684f464c70>),
 ('CRIM ~ ZN + RAD + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x26855b85c40>),
 ('CRIM ~ ZN + DIS + RAD + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2685af9a3d0>),
 ('CRIM ~ ZN + NOX + DIS + RAD + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x26863d1f130>),
 ('CRIM ~ ZN + INDUS + NOX + DIS + RAD + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2686b740910>),
 ('CRIM ~ ZN + INDUS + CHAS + NOX + DIS + RAD + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2687316bd00>),
 ('CRIM ~ ZN + INDUS + CHAS + NOX + RM + DIS + RAD + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2687864cbe0>),
 ('CRIM ~ ZN + INDUS + CHAS + NOX + RM + DIS + RAD + PTRATIO + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2687b030dc0>),
 ('CRIM ~ ZN + INDUS + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2687c2b2f70>),
 ('CRIM ~ ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD + TAX + PTRATIO + B + LSTAT',
  <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x2687c2f62b0>)]
```

<Forward Stepwise>



```
     index           rss                                         predictors
0        1  22744.611548                                              [RAD]
1        2  21640.908632                                       [LSTAT, RAD]
2        3  21348.884262                                    [B, LSTAT, RAD]
3        4  21257.628898                                [ZN, B, LSTAT, RAD]
4        5  21117.675938                           [DIS, ZN, B, LSTAT, RAD]
5        6  20973.524280                      [NOX, DIS, ZN, B, LSTAT, RAD]
6        7  20920.602305               [INDUS, NOX, DIS, ZN, B, LSTAT, RAD]
7        8  20871.213821         [CHAS, INDUS, NOX, DIS, ZN, B, LSTAT, RAD]
8        9  20860.752589     [RM, CHAS, INDUS, NOX, DIS, ZN, B, LSTAT, RAD]
9       10  20850.898001  [PTRATIO, RM, CHAS, INDUS, NOX, DIS, ZN, B, LS...
10      11  20848.016533  [TAX, PTRATIO, RM, CHAS, INDUS, NOX, DIS, ZN, ...
11      12  20847.783845  [AGE, TAX, PTRATIO, RM, CHAS, INDUS, NOX, DIS,...
```

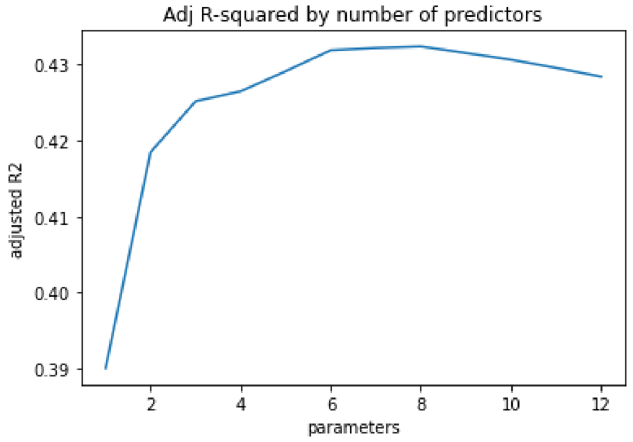
<Backward Stepwise>

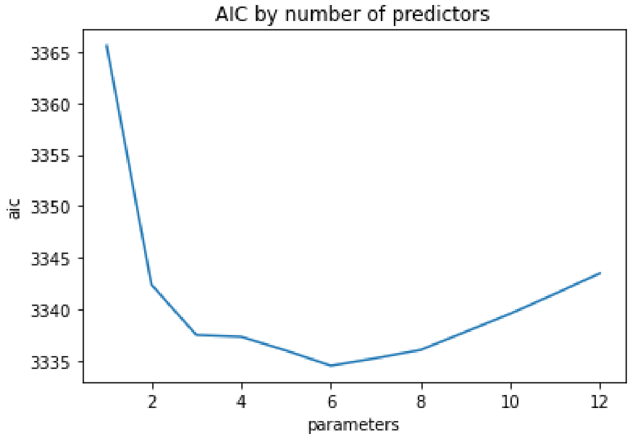


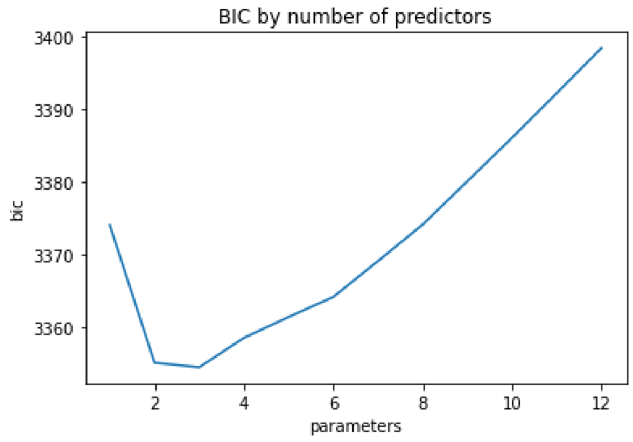
```
             rss                                         predictors
0   20847.783845  [ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX,...
1   20848.016533  [ZN, INDUS, CHAS, NOX, RM, DIS, RAD, TAX, PTRA...
2   20850.898001  [ZN, INDUS, CHAS, NOX, RM, DIS, RAD, PTRATIO, ...
3   20860.752589     [ZN, INDUS, CHAS, NOX, RM, DIS, RAD, B, LSTAT]
4   20871.213821         [ZN, INDUS, CHAS, NOX, DIS, RAD, B, LSTAT]
5   20920.602305               [ZN, INDUS, NOX, DIS, RAD, B, LSTAT]
6   20973.524280                      [ZN, NOX, DIS, RAD, B, LSTAT]
7   21117.675938                           [ZN, DIS, RAD, B, LSTAT]
8   21257.628898                                [ZN, RAD, B, LSTAT]
9   21348.884262                                    [RAD, B, LSTAT]
10  21640.908632                                       [RAD, LSTAT]
11  22744.611548                                              [RAD]
```

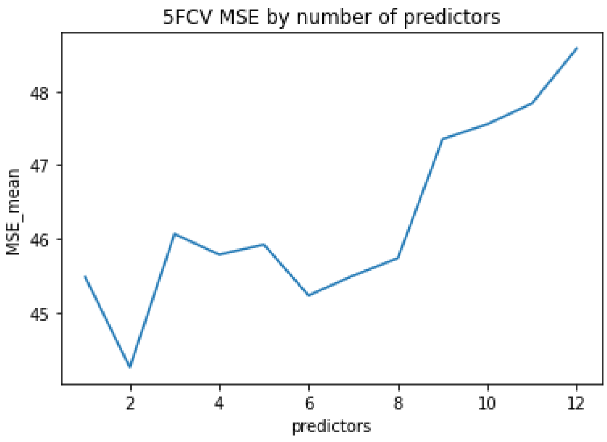
b. *Compare the results of using the mathematical-adjustment approaches (AIC, BIC, & adjusted R2) to using 5-Fold Cross-Validation (5FCV). Propose a model (or set of models) that seem to perform well on this data set, and justify your answer.*

- Adjusted R-squared:
    - The model with the highest adjusted R-squared has the 8 parameters as shown below
    - CRIM ~ ZN + INDUS + CHAS + NOX + DIS + RAD + B + LSTAT
- AIC
    - The model with the lowest AIC has the 6 parameters as shown below
    - CRIM ~ ZN + INDUS + NOX + DIS + RAD + B + LSTAT
- BIC:
    - The model with the lowest AIC has the 3 parameters as shown below
    - CRIM ~ RAD + B + LSTAT
- 5FCV:
    - The model with the lowest MSE from the 5FCV has 2 parameters as shown below
    - CRIM ~ B + LSTAT
- The results show that each method for choosing the optimal model yields a different outcome. The 5FCV yields a lower number of predictors than the mathematical-adjustment approaches. Of the mathematical-adjustment approaches the adjusted R-squared, which chooses the most predictors, is not as well motivated in statistical theory as AIC or BIC. From the graphs below we can see that the estimated test errors seem to be somewhat close for the 2~6 variable models. Based on the one-standard error rule, if a set of models appear to be similarly good, we might as well choose the simplest model. In this case it would be the model with two variables, B and LSTAT.









c. *Does your chosen model involve all of the features in the data set? Why or why not?*

- It doesn't include all of the features in the data set. It doesn't include all the features because adding some features to the model increases the MSE. Therefore I chose the model that had the lowest MSE from the 5FCV by comparing the results with the other adjustment approaches and applying the one standard error rule.