Model Selection in R for Ozone Levels in Los Angeles 1976

Yuqing Zhou and Mazin Rafi

ABSTRACT This paper aims to present an analysis of atmospheric ozone concentration in the Los Angeles Basin in 1976 by building a simple, yet effective model. The procedure begins with an examination of three major variable screening methods in order to fit the regression model with the best predictor variables, keeping in mind adjusted R^2 and AIC values. After selecting the best model from these procedures, we perform a leave-one-out cross validation and a 5-fold cross validation test to measure the performance of the model on new test data sets. Based on the principle of parsimony and various model selection techniques, our results show that the best fit model should use the regressors temperature, humidity, inversion base height, and vis.

INTRODUCTION

In accordance with the principle of parsimony, we want the simplest model with the least assumptions and variables but with the greatest explanatory power. Therefore, as part of the model building process, it is necessary to select the most significant independent variables for modeling the mean response, E(y). Starting with stepwise regression, we identify the response variable y and the set of potentially important independent variables. There are three types of stepwise regression: forward selection, backwards elimination, and sequential regression. Forward selection assumes that no variables are in the model originally and significant regressors are added into the model one at a time. Backward elimination assumes the opposite and insignificant regressors are removed one at a time until all variables are examined. Sequential regression is a combination of the methods above that test at each step for variables to be included or excluded. After stepwise regression, we continue to examine another popular method, all possible regression selection. This algorithm fits all regressions involving one regressor, two regressors, three regressors, and so on. The selection criterion is recorded for each regression. Then, we determine which subset of variables is the best. Lastly, we assess model performance with a 5-fold cross validation test. This is to ensure that the model yields consistent and accurate results. We conduct this investigation using the ozone dataset from the “faraway” package. The methods and materials will involve stepwise selection procedures from the “olsrr” and “MASS” packages in R. In addition to analyzing the variables with the methods described above, we also have a control in which all the variables are included regardless of significance.

MATERIALS AND METHODS

Overview of the dataset: Faraway package ozone is a study of the relationship between atmospheric ozone concentration and meteorology in the Los Angeles Basin in 1976.

Format:

A data frame with 330 observations on the following 10 variables.
O3 Ozone conc., ppm, at Sandbug AFB (Our response variable, or Y)
vh a numeric vector (X_1)
wind wind speed (X_2)
humidity a numeric vector (X_3)
temp temperature (X_4)
ibh inversion base height (X_5)
dpg Daggett pressure gradient (X_6)
ibt a numeric vector (X_7)
vis visibility (X_8)
doy day of the year (X_9)

Principle of Parsimony:

The principle of parsimony is based on a common dilemma found in model-building. When given the data and predictors, it is difficult to choose between having few regressors or having many regressors. Having too few regressors will cause a lot of variance within a model, and having many regressors is harder to implement and maintain. The principle of parsimony states that having the fewest number of regressors that accurately predict the model is the best choice. For our model selection, if we are given two models that both have equally strong predicting power, the model with the least number of predictor variables should be chosen. Picking a model with more predictor variables could cause issues such as multicollinearity and overfitting.

Ordinary Least Square Regression (controlled)

This is multiple regression without any examination.

Stepwise Regression Methods

Forward Selection

Forwards selection is a variable screening process that is helpful for datasets with many variables and for datasets where multicollinearity can be a problem. It should be noted that if a variable is added to the model, it cannot be removed.

1. First, begin the model with no predictor variables. ($E(y)= \beta_0$)
2. Then, conduct a t-test to determine if the Null Hypothesis (H_0: $\beta_i=0$) for every variable. In other words, find a model for $E(y)= \beta_0+\beta_1X_1$, $E(y)= \beta_0+\beta_2X_2$, ...$E(y)= \beta_0+\beta_9X_9$, and calculate their p-values and general fit.
3. Pick/find the model with the best model fit (highest R^2 or the lowest AIC values.)
    a. In this case, Temperature (X_4) was selected to be the one variable model with the best R^2 value. Our current model is $E(y)=\beta_0+\beta_4X_4$

    b. By AIC selection, Temperature (X_4) was also selected to be the best by having the lowest AIC value. The model is $E(y)=\beta_0+\beta_4X_4$ (see Forward Selection Results)

4. Add another variable to the current model. Test for all variables. (In other words, find a model for $E(y)=\beta_0+\beta_5X_5+\beta_1X_1$ … $E(y)=\beta_0+\beta_5X_5+\beta_9X_9$)
5. Calculate the R^2 values for each of the two-variable-models. Pick the model with the highest R^2 value.
    a. In this case, ibh (inversion base height or X_5) is the next best predictor and produces the best two variable model highest adjusted R^2 value and the lowest AIC value. Thus, the model is $E(y)=\beta_0+\beta_4X_4+\beta_5X_5$
6. Repeat this process for the rest of the variables. Stop until the model begins to lose validity; in other words, when the adjusted R^2 values cannot get any higher or the AIC values cannot get any lower. (The final model will be calculated and discussed in the Results and Discussion section respectively.)

Backward Elimination

Backwards elimination is a variable screening process similar to forward selection, but in the reverse. Rather than add variables to the model to create a better one, we shall start with the complete model and take predictors until it becomes the most valid.

1. First, begin the model with all the predictor variables in a single model. In this case. ($E(y)= \beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8+\beta_9X_9$)
2. Determine which of the β_i values is the most insignificant (low t-value or a very high p-value.)
3. Remove this variable from the model. Test this new model's validity (this can be done by calculating adjusted R^2, AIC, C_p, PRESS values, etc.)
    a. In this case, removal of the predictor dpg (Dagget pressure gradient or X_6) will allow the model to have the highest adjusted R^2 value. Our current model is: $E(y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_7X_7+\beta_8X_8+\beta_9X_9$
    b. By AIC selection, removal of the predictor dpg (Dagget pressure gradient or X_6) is best by having the lowest AIC value. The model is: $E(y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_7X_7+\beta_8X_8+\beta_9X_9$ (see Backward Elimination Results)
4. If the new model has less fit than the original, keep the original and do not remove any more variables. If the new model has more fit, proceed.
5. Calculate the adjusted R^2 values for each of the seven-variable-models. Pick the model with the best fit.
6. Repeat this process for the rest of the variables.
    a. Continue adding variables until the adjusted R^2 value begins to decrease, or until the AIC begins to increase. (The final model will be calculated and discussed in the Results and Discussion section respectively.)

Sequential Regression

Sequential is a technique that combines techniques between forward and backwards elimination. It is the best method compared to forward selection and backwards elimination but takes the longest. The first four steps are the same from forward selection. However, sequential requires that each t-value for every β_i in the model gets retested.

1. First, begin the model with no predictor variables. ($E(y)= \beta_0$)
2. Then, conduct a t-test to determine if the Null Hypothesis (H_0: $\beta_i=0$) for every variable. (In other words, find a model for $E(y)= \beta_0+\beta_1X_1$, $E(y)= \beta_0+\beta_2X_2$, ...$E(y)= \beta_0+\beta_9X_9$, and calculate their p-values and general fit.)
3. Pick/find the model with the best model fit (highest R^2 or adjusted R^2, the lowest AIC values/multicollinearity, etc.)
    a. In this case, $E(y)=\beta_0+\beta_5X_5$ (must verify)
4. Add another variable to the current model. Test for all variables. (In other words, find a model for $E(y)=\beta_0+\beta_5X_5+\beta_1X_1$ … $E(y)=\beta_0+\beta_5X_5+\beta_9X_9$) Find the best fit model.
    a. In this case, $E(y)=\beta_0+ \beta_4X_4 + \beta_5X_5$ (must verify)
5. Test the t-values of every β_i in the model once more.
    a. If the first β_i is determined to be insignificant, it is removed from the model and replaced with another unpicked β_i that will have a maximum t-value in the presence of the second β_i.
    b. If no other inclusion/exclusion works best, then the first model proposed in step 3 is the best fit.
6. Check for a third independent variable to include. Add this to the model.
7. Recheck the t-values for all variables so far.
8. Continue to add and recheck until no further independent variables can be found that yield significant t-values.

<u>All Possible Regression Selection</u>

All subset regression tests all possible subsets of the set of potential independent variables. If there are K potential independent variables (besides the constant), then there are 2^kdistinct subsets of them to be tested. In this case, we have 512 possible models to test.

R^2_{adj}

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1} \right]$$

R^2 is the coefficient of determination, n is sample size, k is the number of estimators. It is a statistical measure that represents the proportion of the variance for a dependent variable that's explained by independent variable(s) in a regression model. Although we look for a large R^2, this is not a reliable criterion because R^2 can easily be forced to 1 by adding more independent variables to the model. As a general rule of thumb, adding too many unnecessary variables will violate the rule of parsimony. Instead, we will use $R^2_{adj.}$

R^2_{adj} explains the percentage of variation between the predictor and predicted variables while keeping in mind the number of estimators and sample size. Overall, it is much more accurate when describing how well fit a model is.

Mallow's C_p Statistic

$$C_p = \frac{SSE_p}{S^2} - N + 2P$$

Mallow's C_p Statistic is another technique that assesses the fit of a multiple regression model. A small Mallows' Cp value indicates that the model is relatively precise (has small variance) in estimating the true regression coefficients and predicting future responses. A Mallows' Cp value that is close to the number of predictors plus the constant indicates that the model is relatively unbiased in estimating the true regression coefficients and predicting future responses. Models with lack-of-fit and bias have values of Mallows' Cp larger than p.

RMSE or Root-Mean-Square-Error

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Where $\hat{y}_i$ represents predicted values and y_i represents actual values. RMSE can be thought of as the distance between observed values from the data and predicted values from the model. This means that if RMSE is low, the model is a good predictor for the data.

Akaike Information Criteria

$$AIC = n \cdot \ln\left(\frac{SSE}{n}\right) + 2k$$

AIC is a function of the number of observations n, the sum of squared errors (SSE), and the number of independent variables $k \leq p + 1$ where k includes the intercept. AIC looks at the unknown true likelihood function of the data and the fitted likelihood function of the model, meaning that a lower AIC signifies that a model is closer to the true model. To further verify that each of the selection processes work best, by using R, we shall be running the stepwise selection processes again while also keeping in mind the models that produce the lowest AIC values (this can be shown visually through the “forwardaic”, “backwardaic”, and “stepwiseaic” commands found in the MASS library.)

Bayesian Information Criteria

$$BIC = n \cdot \ln\left(\frac{SSE}{n}\right) + \frac{2(k+2)n\sigma^2}{SSE} - \frac{2n^2\sigma^4}{SSE^2}$$

BIC is a function of the number of observations n, the SSE, the pure error variance fitting the full model, and the number of independent variables k ≤ p + 1 where k includes the intercept. The penalty term is larger in BIC than in AIC. Like AIC, lower values in BIC imply a lower error variance when fitting the full model; in other words, we should be looking for models that produce low BIC values.

Schwarz Bayesian Criteria

$$SBC = n \cdot \ln\left(\frac{SSE}{n}\right) + k \ln n$$

SBC is a function of the number of observations n, the SSE, and the number of independent variables k ≤ p + 1 where k includes the intercept. The penalty term for SBC uses a multiplier of ln n for k instead of a constant 2 by incorporating the sample size n. Like AIC and SBIC, this measures an error of variance, so we should be looking for models that have low SBIC values.

K-fold Cross-Validation

This method evaluates the model performance on different subsets of the training data and then calculates the cross-validation error, which is the average of the k recorded prediction error that serves as the performance metric for the model. The basic idea is as follows:

1. Randomly split the data set into k-subsets (in this case 5 subsets)
2. Reserve one subset and train the model on all other subsets
3. Test the model on the controlled subset and record the prediction error
4. Repeat until all 5 subsets have served as the test set and compute the average prediction error.

There are exhaustive cross validation methods as well, but this requires incredibly long calculations and computation. One of the ways to test this is the leave-one-out cross validation. This applies when k=n, or in this case k=330. This allows the estimators to be completely unbiased for predicting Y; a downside to this is that splitting the data into very small subsets causes high variance. As a result, any calculated R^2 values from leave-one-out cross validation cannot be considered. Also, manual computation for leave-one-out cross validation can be quite long, since it requires 330 applications of the same learning method.

The choice of k involves a bias-variance trade-off. The typical choice for k is generally 5 or 10 because their test error rate estimates do not have high bias or variance. As k gets larger, the difference in size between the training set and the resampling subsets gets smaller. As this difference decreases, the bias becomes smaller. Although smaller k values tend to have more bias, the choice of 5 in this dataset works well due to the large sample size. Repeated k fold cross validation is splitting the data into k folds which can be repeated n numbers of times.

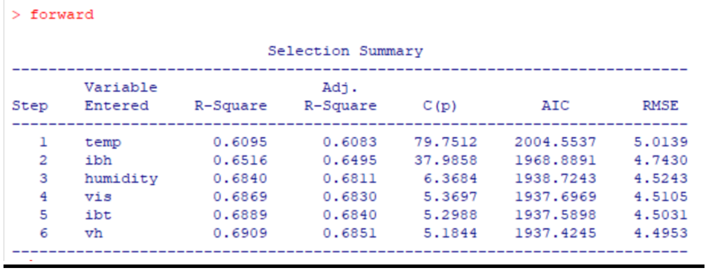
RESULT

View attached text file for complete code and output.

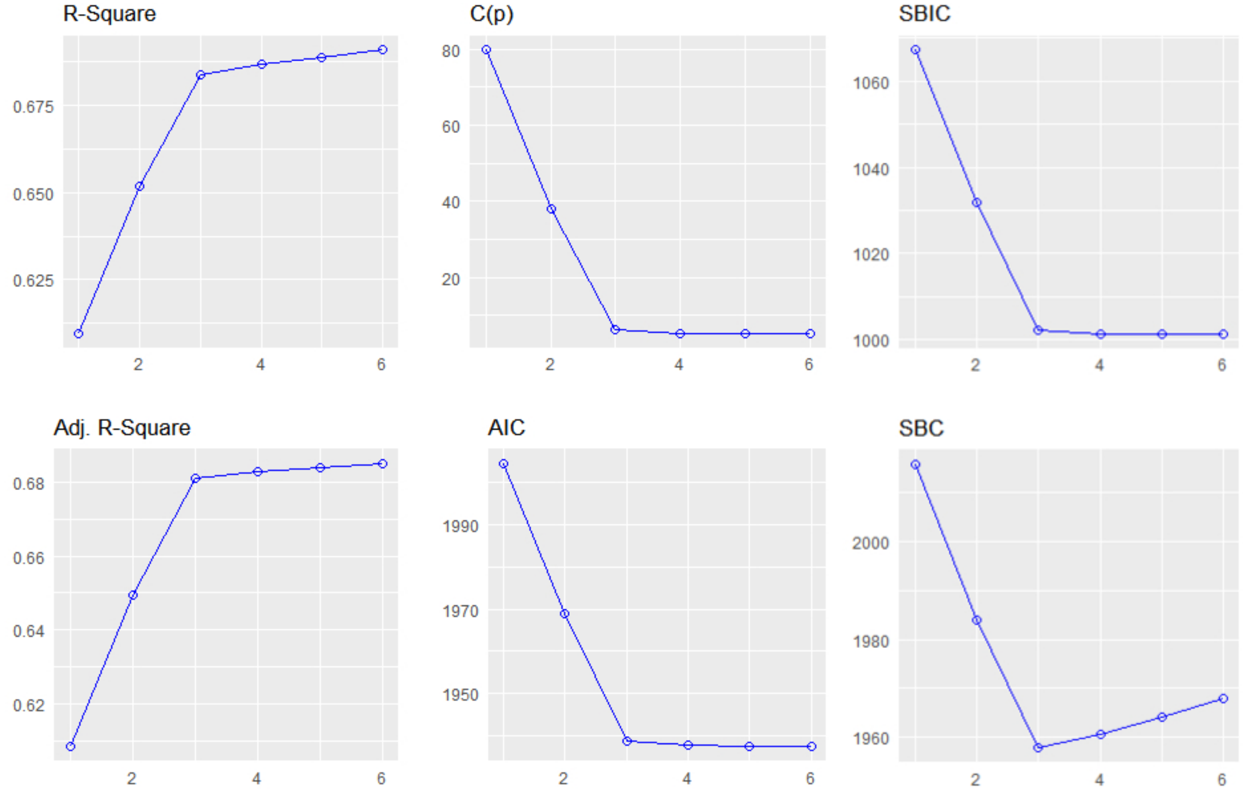
Forward Selection Graphs (General and AIC):

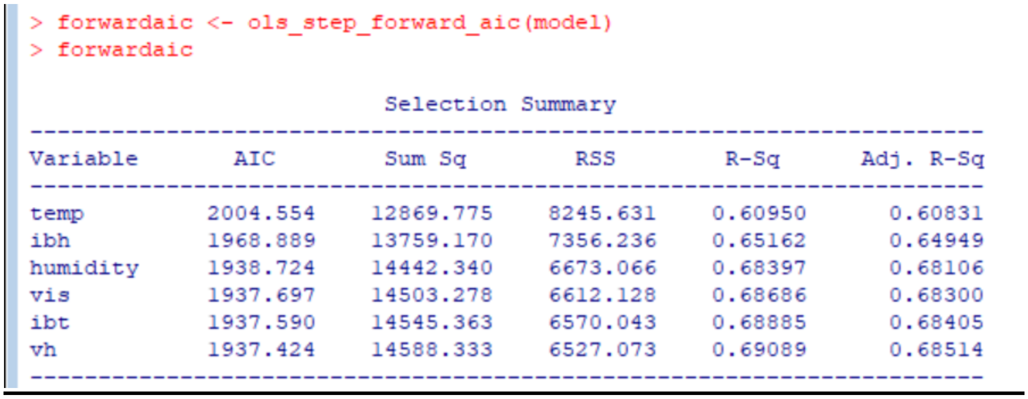
Based on p-values, Temperature was first determined to be the most statistically significant. Without checking the t-values of every preceding valuables, the following variables were added through

Forward Selection: temperature (X_6), ibh (X_5), humidity (X_3), vis (X_8), ibt (X_7), and vh (X_1). The graphs show that the addition of these predictors yield a higher adjusted R^2 values and lower C(p), AIC, SBIC, and SBC values.

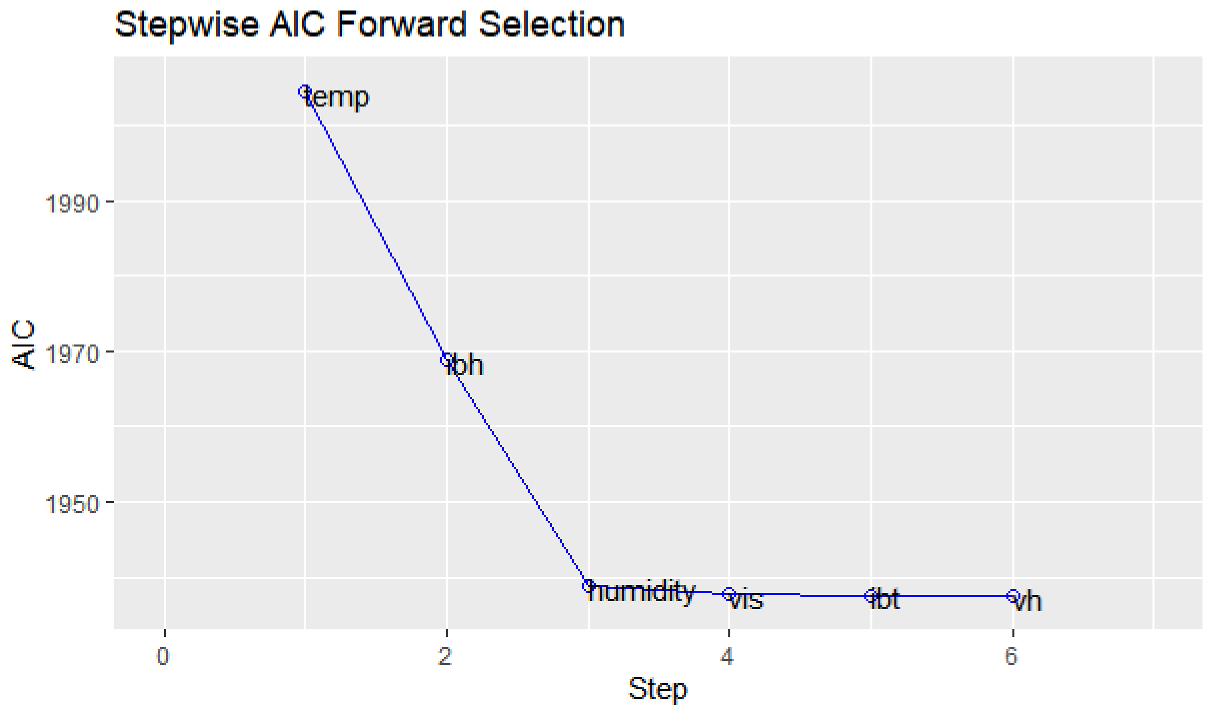


```
> forward

                              Selection Summary
--------------------------------------------------------------------------------
        Variable                   Adj.
Step    Entered        R-Square    R-Square     C(p)         AIC         RMSE
--------------------------------------------------------------------------------
   1    temp             0.6095      0.6083    79.7512    2004.5537    5.0139
   2    ibh              0.6516      0.6495    37.9858    1968.8891    4.7430
   3    humidity         0.6840      0.6811     6.3684    1938.7243    4.5243
   4    vis              0.6869      0.6830     5.3697    1937.6969    4.5105
   5    ibt              0.6889      0.6840     5.2988    1937.5898    4.5031
   6    vh               0.6909      0.6851     5.1844    1937.4245    4.4953
--------------------------------------------------------------------------------
```



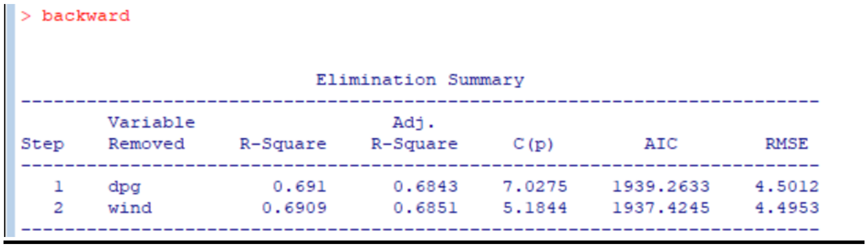


```
> forwardaic <- ols_step_forward_aic(model)
> forwardaic

                          Selection Summary
------------------------------------------------------------------------
Variable         AIC        Sum Sq        RSS       R-Sq      Adj. R-Sq
------------------------------------------------------------------------
temp          2004.554    12869.775    8245.631    0.60950      0.60831
ibh           1968.889    13759.170    7356.236    0.65162      0.64949
humidity      1938.724    14442.340    6673.066    0.68397      0.68106
vis           1937.697    14503.278    6612.128    0.68686      0.68300
ibt           1937.590    14545.363    6570.043    0.68885      0.68405
vh            1937.424    14588.333    6527.073    0.69089      0.68514
------------------------------------------------------------------------
```

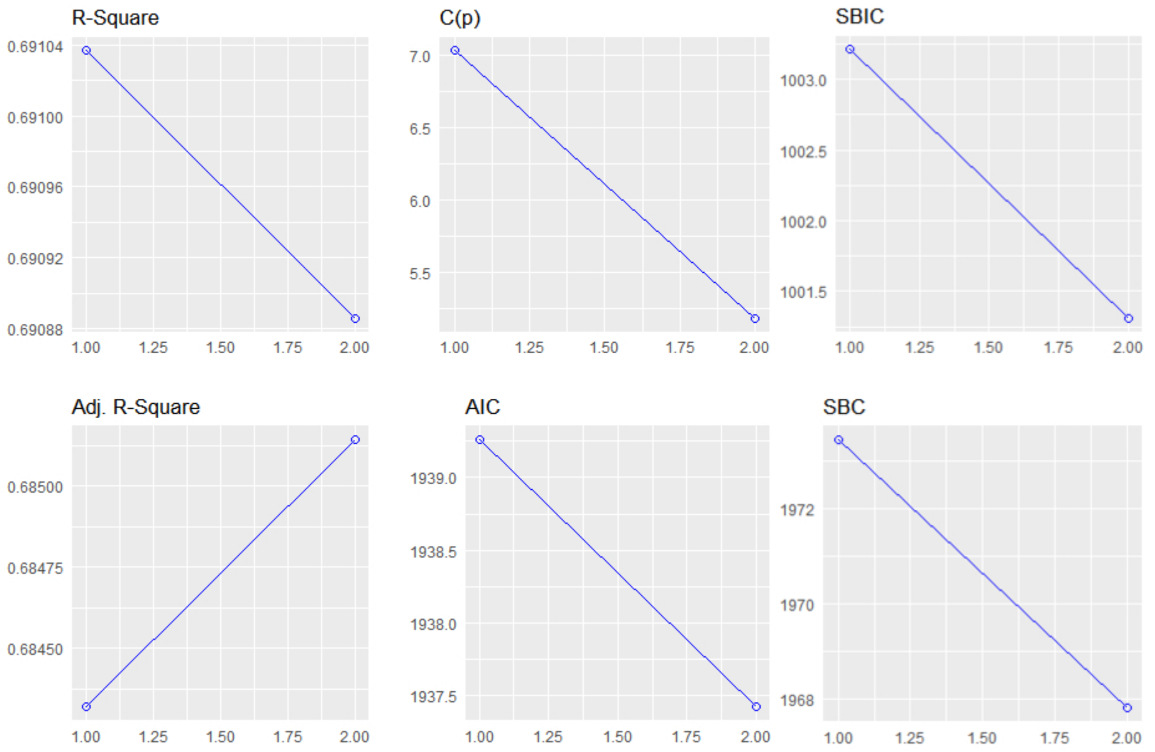


<u>Backwards Elimination Graphs (General and AIC):</u>

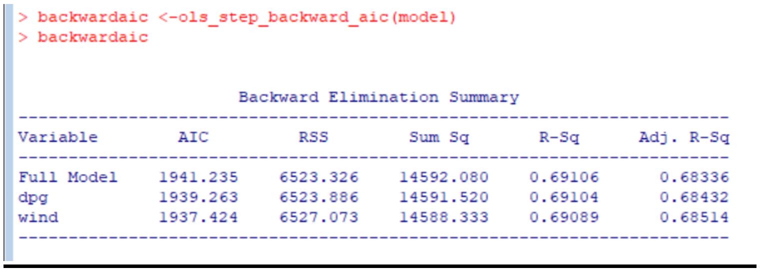
Based on p-values, Temperature was determined to be the most statistically significant. Without checking the t-values of every preceding valuables, the following variables were removed by Backwards Elimination: dpg/Daggett pressure gradient (X_6) and wind (X_2). The graphs show that the removal of these predictors yield a higher adjusted R^2 values and lower C(p), AIC, SBIC, and SBC values.



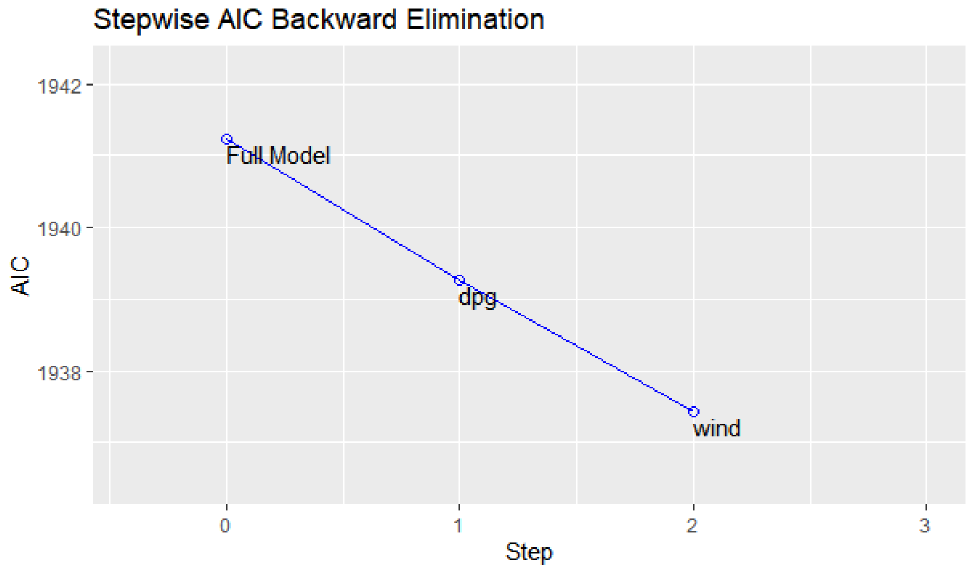
```
> backward

                          Elimination Summary
--------------------------------------------------------------------------
        Variable                  Adj.
Step    Removed     R-Square    R-Square     C(p)       AIC         RMSE
--------------------------------------------------------------------------
   1    dpg            0.691      0.6843    7.0275    1939.2633    4.5012
   2    wind          0.6909      0.6851    5.1844    1937.4245    4.4953
--------------------------------------------------------------------------
```



Stepwise AIC Backward Selection Graphs:



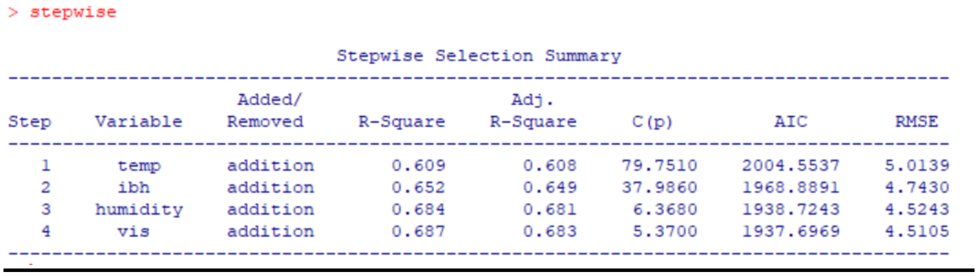
```
> backwardaic <-ols_step_backward_aic(model)
> backwardaic


                    Backward Elimination Summary
-------------------------------------------------------------------------
Variable        AIC         RSS        Sum Sq        R-Sq      Adj. R-Sq
-------------------------------------------------------------------------
Full Model    1941.235    6523.326    14592.080    0.69106      0.68336
dpg           1939.263    6523.886    14591.520    0.69104      0.68432
wind          1937.424    6527.073    14588.333    0.69089      0.68514
-------------------------------------------------------------------------
```



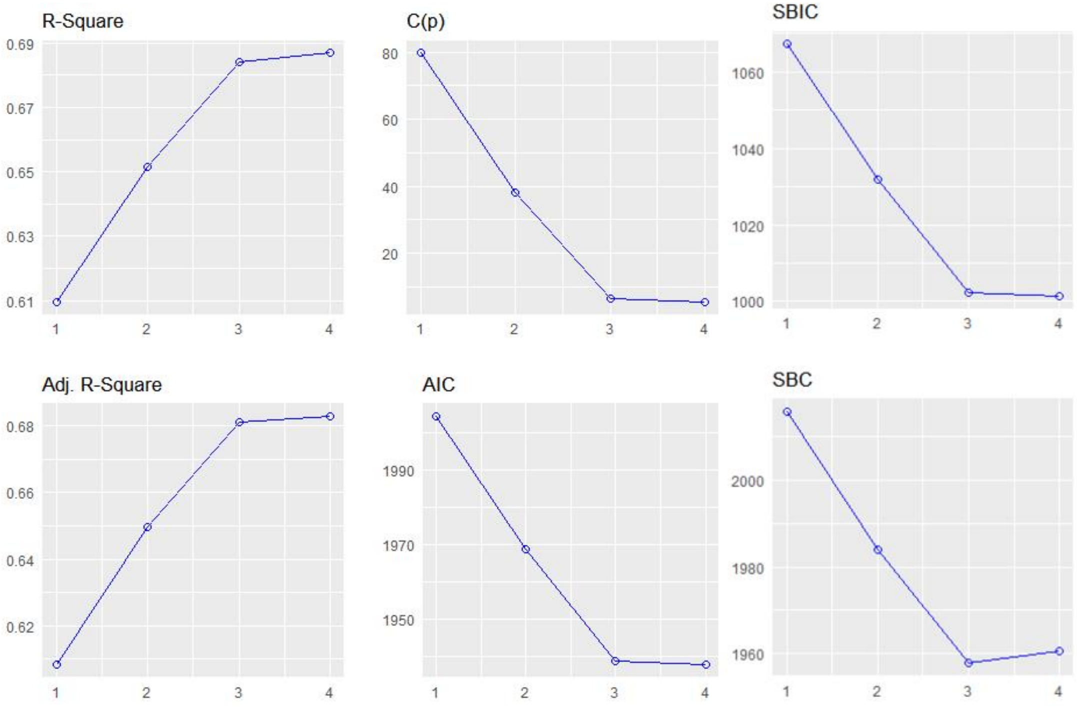
Sequential Regression Graphs:

Based on p-values, Temperature was first determined to be the most statistically significant. Without checking the t-values of every preceding valuables, the following variables were added to the model through Sequential Regression: temperature (X_6), ibh (X_5), humidity (X_3), and vis (X_8).

The Sequential Regression by AIC showed that temperature (X_6), ibh (X_5), humidity (X_3), vis (X_8), ibt (X_7), and vh (X_1) should be the added regressors. It is important to note that X_7 and X_1 are added in sequential regression by AIC, but not in general.



```
> stepwise

                              Stepwise Selection Summary
------------------------------------------------------------------------------------------
                         Added/                    Adj.
Step      Variable       Removed      R-Square    R-Square      C(p)         AIC        RMSE
------------------------------------------------------------------------------------------
   1        temp         addition       0.609       0.608     79.7510    2004.5537    5.0139
   2        ibh          addition       0.652       0.649     37.9860    1968.8891    4.7430
   3      humidity       addition       0.684       0.681      6.3680    1938.7243    4.5243
   4        vis          addition       0.687       0.683      5.3700    1937.6969    4.5105
------------------------------------------------------------------------------------------
```



R-Square

C(p)

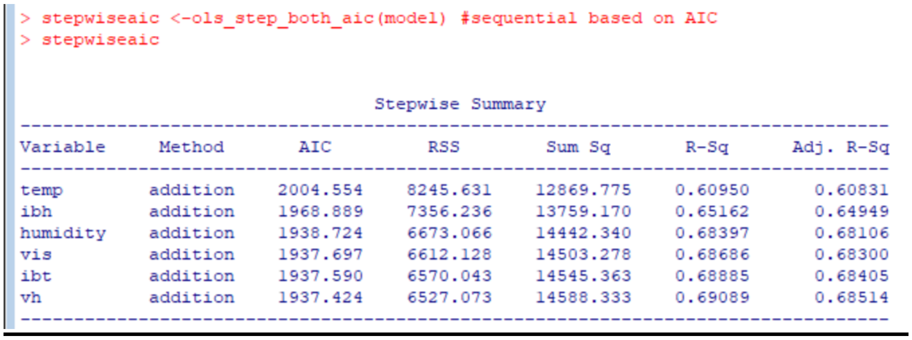
SBIC

Adj. R-Square

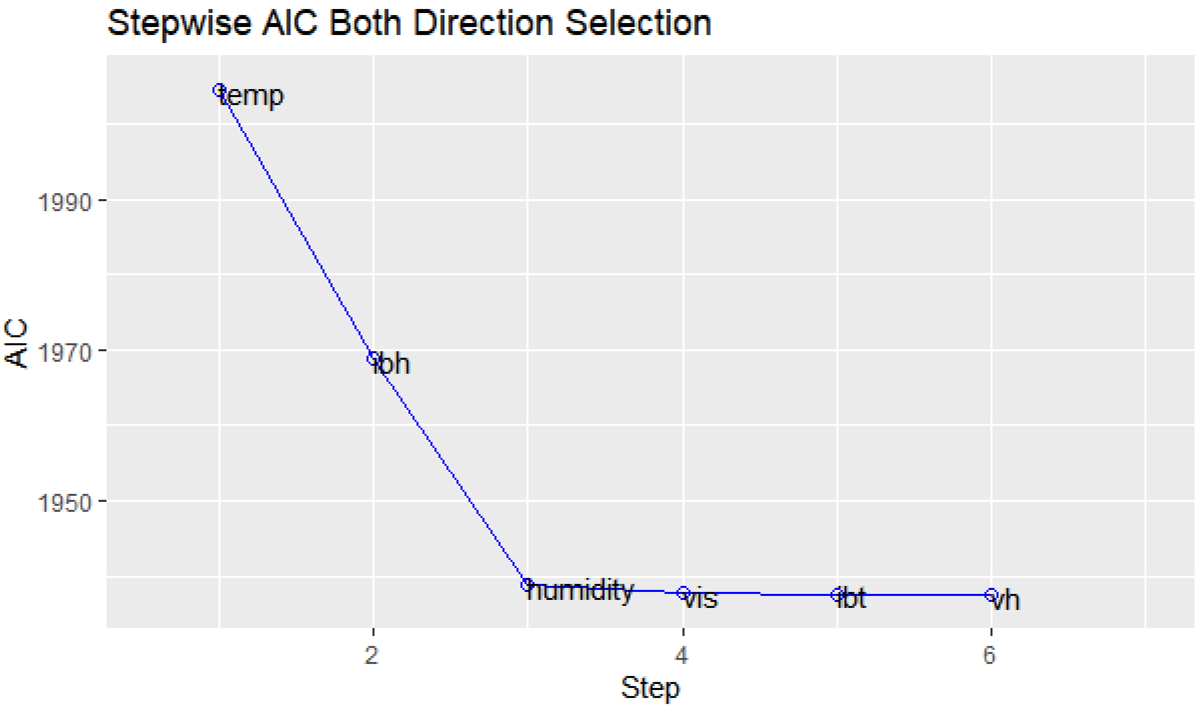
AIC

SBC

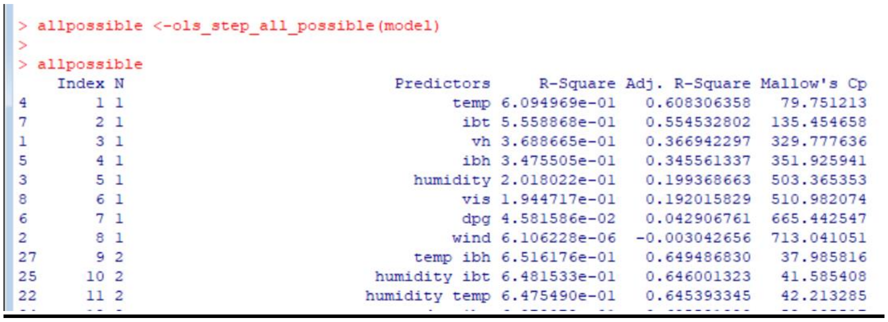
Stepwise AIC Both Direction Selection:



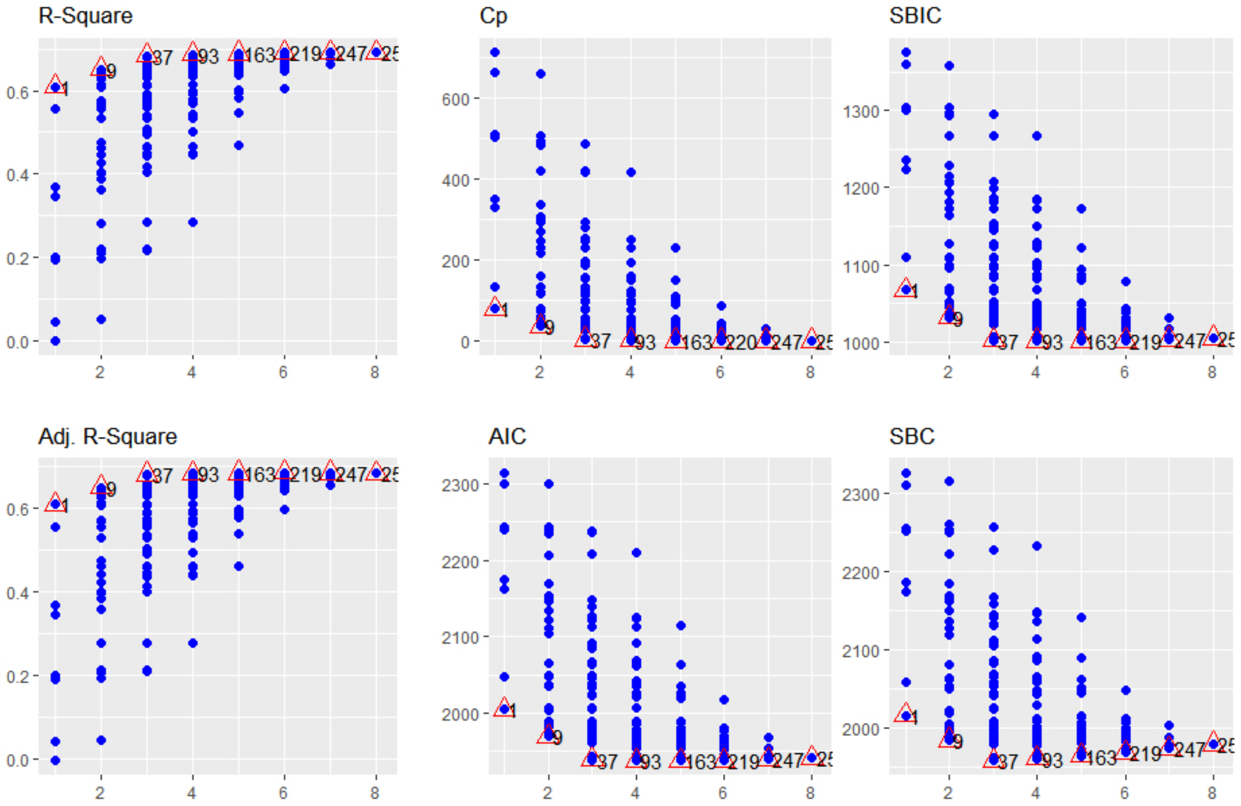
```
> stepwiseaic <-ols_step_both_aic(model) #sequential based on AIC
> stepwiseaic


                              Stepwise Summary
--------------------------------------------------------------------------------
Variable      Method       AIC          RSS          Sum Sq        R-Sq       Adj. R-Sq
--------------------------------------------------------------------------------
temp         addition    2004.554     8245.631     12869.775     0.60950     0.60831
ibh          addition    1968.889     7356.236     13759.170     0.65162     0.64949
humidity     addition    1938.724     6673.066     14442.340     0.68397     0.68106
vis          addition    1937.697     6612.128     14503.278     0.68686     0.68300
ibt          addition    1937.590     6570.043     14545.363     0.68885     0.68405
vh           addition    1937.424     6527.073     14588.333     0.69089     0.68514
--------------------------------------------------------------------------------
```



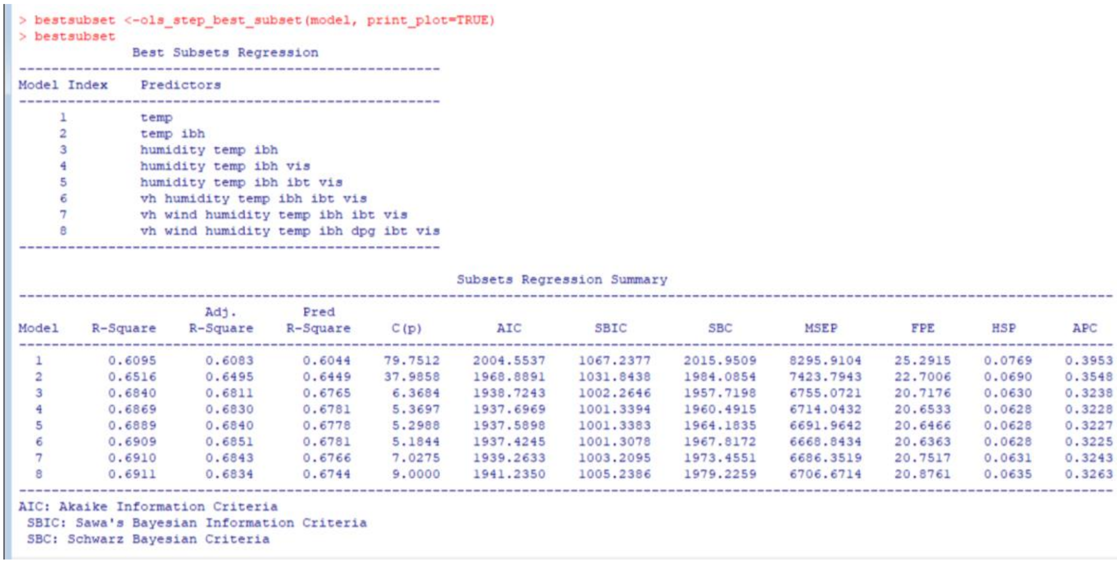
<u>All Possible Selections Models:</u>

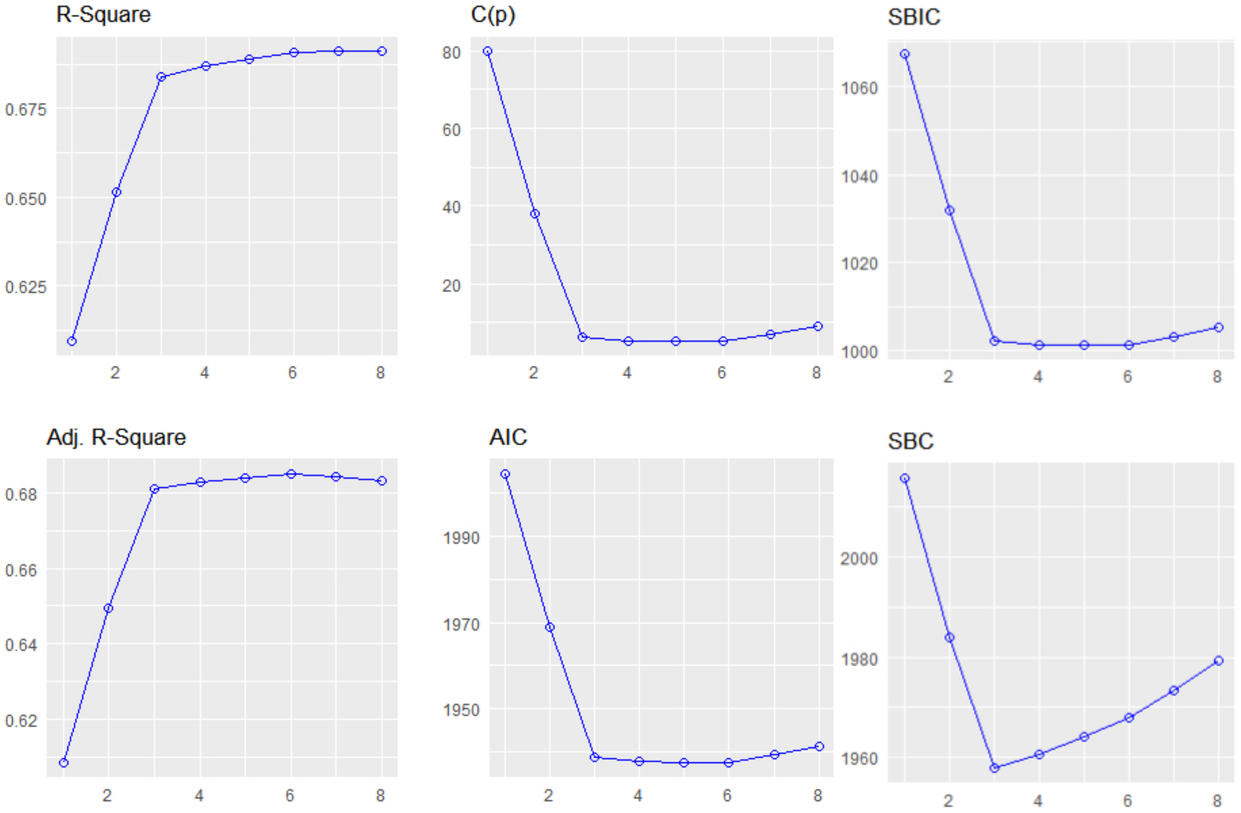


```
> allpossible <-ols_step_all_possible(model)
>
> allpossible
    Index N                               Predictors     R-Square Adj. R-Square Mallow's Cp
4       1 1                                     temp 6.094969e-01   0.608306358   79.751213
7       2 1                                      ibt 5.558868e-01   0.554532802  135.454658
1       3 1                                       vh 3.688665e-01   0.366942297  329.777636
5       4 1                                      ibh 3.475505e-01   0.345561337  351.925941
3       5 1                                 humidity 2.018022e-01   0.199368663  503.365353
8       6 1                                      vis 1.944717e-01   0.192015829  510.982074
6       7 1                                      dpg 4.581586e-02   0.042906761  665.442547
2       8 1                                     wind 6.106228e-06  -0.003042656  713.041051
27      9 2                                 temp ibh 6.516176e-01   0.649486830   37.985816
25     10 2                             humidity ibt 6.481533e-01   0.646001323   41.585408
22     11 2                            humidity temp 6.475490e-01   0.645393345   42.213285
```

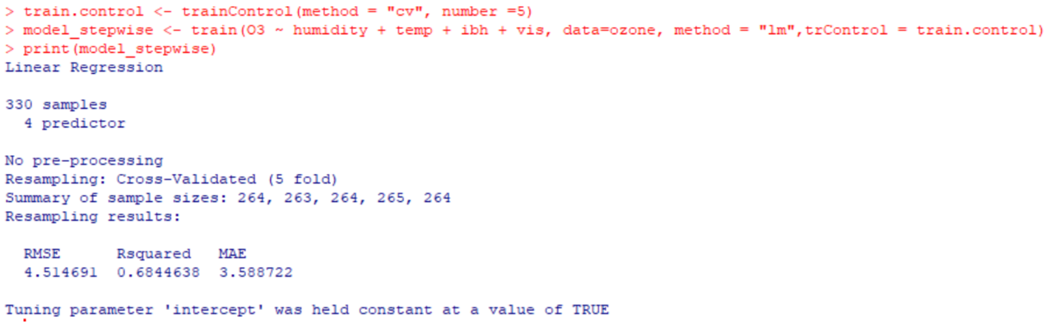


Best Subset Model Graphs:


```
> bestsubset <-ols_step_best_subset(model, print_plot=TRUE)
> bestsubset
               Best Subsets Regression
----------------------------------------------------
Model Index    Predictors
----------------------------------------------------
     1         temp
     2         temp ibh
     3         humidity temp ibh
     4         humidity temp ibh vis
     5         humidity temp ibh ibt vis
     6         vh humidity temp ibh ibt vis
     7         vh wind humidity temp ibh ibt vis
     8         vh wind humidity temp ibh dpg ibt vis
----------------------------------------------------

                                                   Subsets Regression Summary
------------------------------------------------------------------------------------------------------------------------------------
                       Adj.        Pred
Model    R-Square    R-Square    R-Square     C(p)        AIC         SBIC          SBC          MSEP         FPE        HSP       APC
------------------------------------------------------------------------------------------------------------------------------------
  1        0.6095      0.6083      0.6044    79.7512    2004.5537    1067.2377    2015.9509    8295.9104    25.2915    0.0769    0.3953
  2        0.6516      0.6495      0.6449    37.9858    1968.8891    1031.8438    1984.0854    7423.7943    22.7006    0.0690    0.3548
  3        0.6840      0.6811      0.6765     6.3684    1938.7243    1002.2646    1957.7198    6755.0721    20.7176    0.0630    0.3238
  4        0.6869      0.6830      0.6781     5.3697    1937.6969    1001.3394    1960.4915    6714.0432    20.6533    0.0628    0.3228
  5        0.6889      0.6840      0.6778     5.2988    1937.5898    1001.3383    1964.1835    6691.9642    20.6466    0.0628    0.3227
  6        0.6909      0.6851      0.6781     5.1844    1937.4245    1001.3078    1967.8172    6668.8434    20.6363    0.0628    0.3225
  7        0.6910      0.6843      0.6766     7.0275    1939.2633    1003.2095    1973.4551    6686.3519    20.7517    0.0631    0.3243
  8        0.6911      0.6834      0.6744     9.0000    1941.2350    1005.2386    1979.2259    6706.6714    20.8761    0.0635    0.3263
------------------------------------------------------------------------------------------------------------------------------------
AIC: Akaike Information Criteria
 SBIC: Sawa's Bayesian Information Criteria
 SBC: Schwarz Bayesian Criteria
```

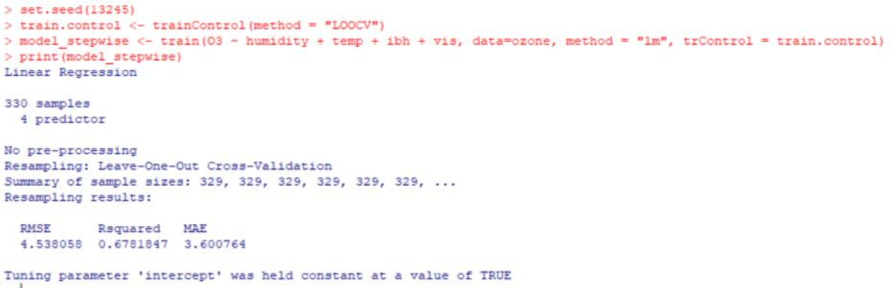


5 Fold Cross Validation



```
> train.control <- trainControl(method = "cv", number =5)
> model_stepwise <- train(O3 ~ humidity + temp + ibh + vis, data=ozone, method = "lm",trControl = train.control)
> print(model_stepwise)
Linear Regression

330 samples
  4 predictor

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 264, 263, 264, 265, 264
Resampling results:

  RMSE      Rsquared   MAE
  4.514691  0.6844638  3.588722

Tuning parameter 'intercept' was held constant at a value of TRUE
```

Leave One Out Cross Validation



```
> set.seed(13245)
> train.control <- trainControl(method = "LOOCV")
> model_stepwise <- train(O3 ~ humidity + temp + ibh + vis, data=ozone, method = "lm", trControl = train.control)
> print(model_stepwise)
Linear Regression

330 samples
  4 predictor

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 329, 329, 329, 329, 329, 329, ...
Resampling results:

  RMSE      Rsquared   MAE
  4.538058  0.6781847  3.600764

Tuning parameter 'intercept' was held constant at a value of TRUE
```

DISCUSSION

We begin our analysis by first looking at the stepwise regression methods. In forward selection regression, the variables that are included in the model are temperature (X_4), inversion base height (X_5), humidity (X_3), visibility (X_8), ibt (X_7), and vh (numeric vectors) (X_1). In backward elimination regression, the variables that are most insignificant to the model are daggett pressure gradient (X_6) and wind speed (X_2). In sequential regression, the variables that are significant to the model are temperature (X_4), inversion base height (X_5), humidity (X_3), and visibility (X_8). The rest of the variables that failed during the rechecks are insignificant. The all possible command lists all 512 possible models, as well as which ones fit best based on R^2 values, adjusted R^2 values, and Mallow's c(p). Because the 9-predictor model has been shown to be one of the least significant models, all of their combinations have been omitted from the output. The best subset command in "olsrr" picks the best one regressor, two regressor, three regressor etc., model based on R^2, adjusted R^2, Mallow's C(p), AIC, SBIC, SBC, and a few other criteria.

Here are the best possible models:

1 predictor model: Temperature

$\mathbf{E(y)=\beta_0+\beta_4X_4}$

2 predictor model: Temperature and inversion base height

$\mathbf{E(y)=\beta_0+\beta_4X_4+\beta_5X_5}$

3 predictor model: Humidity, temperature, and inversion base height

$\mathbf{E(y)=\beta_0+\beta_3X_3+\beta_4X_4+\beta_5X_5}$

4 predictor model: Humidity, temperature, inversion base height, and visibility.

$\mathbf{E(y)=\beta_0+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_8X_8}$

5 predictor model: Humidity, temperature, inversion base height, ibt, and visibility.

$\mathbf{E(y)=\beta_0+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_8X_8}$

6 predictor model: Vh, humidity, temperature, inversion base height, ibt, and visibility.

$\mathbf{E(y)=\beta_0+\beta_1X_1+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_8X_8}$

7 predictor model: Vh, wind speed, humidity, temperature, inversion base height, ibt, and visibility.

$\mathbf{E(y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_8X_8}$

8 predictor model: Vh, wind speed, humidity, temperature, inversion base height, ibt, and visibility.

$\mathbf{E(y)=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8}$

We also performed stepwise models (forward, backward, and sequential) based on Akaike Information Criteria. According to forward selection based on AIC, the most significant variables are temperature (X_4), inversion base height (X_5), humidity (X_3), visibility (X_8), ibt (X_7), and vh (numeric vectors) (X_1). As you can see, the result is very similar to forward selection in the previous step. In backward elimination based on AIC, the most insignificant variables are also daggett pressure gradient (X_6) and wind speed (X_2). In sequential regression based on AIC, the variables that are significant to the model are temperature (X_4), inversion base height (X_5), humidity (X_6), visibility (X_8), ibt (X_7), and vh (X_1). In this stepwise method, the variables ibt and vh were included whereas previously it was omitted.

Based on all of the results from the stepwise regression, as well as the values calculated by the best subset command, we bring the total down to two models:

I) Sequential Regression model (index # is 93):

$$\mathbf{E(y)=\beta_0+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_8X_8}$$

Adjusted R^2: 0.683003591   Mallow's Cp: 5.369743

II) Sequential AIC model (index # is 219):

$$\mathbf{E(y)=\beta_0+\beta_1X_1+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_7X_7+\beta_8X_8}$$

Adjusted R^2: 0.685143666   Mallow's Cp: 5.184380

In terms of highest adjusted R^2 and lowest AIC levels, it seems evident that model II is best. But we want the model to be both parsimonious as well. By having an adjusted R^2 only improve by ~.002 points does not vastly improve the model. There may also be problems with overfitting and multicollinearity. To keep the principle of parsimony in mind, model I will be chosen.

Usually sequential regression is the best out of the three because it recalculates and checks every time. So, we will use the sequential regression model for our k- fold cross validation. The leave-one-out method, or the k=n fold cross validation, yielded the following results:

```
RMSE       Rsquared   MAE
 4.538058  0.6781847  3.600764
```

The low RMSE and MAE values and high RSquared values indicate that the model has good fit.

Once again, k=5 was chosen since a smaller k value is best for the purpose of model building, as well as minimizing. The 5-fold cross validation yielded the following results:

```
 RMSE      Rsquared   MAE
  4.514691  0.6844638  3.588722
```

The low RMSE and MAE values, as well as the high R^2 value indicate that this model is a good fit. Compared to the leave one out method, 5-fold cross validation yields better values, and is a better compromise between variance and bias within the model. A good suggestion would be to repeat k-fold cross validation tests with different k-values. Repeated k-fold testing could be further tested on model II or other models like the three regressors model.

Final Conclusion:

The most useful first order model to predict Ozone Concentration at Sandbug AFB:

$$\mathbf{E(y)=\beta_0+\beta_4X_4+\beta_5X_5+\beta_6X_6+\beta_8X_8}$$

A model that has the regressors Temperature (X_4), humidity (X_3), inversion base height (X_5), and vis (X_8).

R^2: 6.868577e-01	Adjusted R^2: 0.683003591	Mallow's Cp: 5.369743
AIC: 1937.6969	SBIC: 1001.3394	SBC: 1960.4915

The model keeps in mind parsimony and all necessary regressors needed to accurately predict Ozone concentration at Sandbug AFB. Further k-fold cross validations can be conducted on the three regressors model (since this model yielded the lowest SBIC values) or 6 regressor model (as previously discussed.) One can also test to see if the inclusion of quadratic terms causes a significantly better fit than the first-order regression models proposed. Repeated k-fold cross validations should be performed on this model to verify that it is a good fit.

Literature Cited

Faraway, J. J. (2014). *Linear Models With R, Second Edition*. Taylor & Francis.

Hastie, T., Friedman, J., & Tisbshirani, R. (2017). The Elements of statistical learning: data mining, inference, and prediction. New York: Springer.

Mendenhall, W., & Sincich, T. (2020). *A second course in statistics: regression analysis*. Hoboken, NJ: Pearson.