Selecting Models with Best Subset Selection, Forward Stepwise Selection, and Backward Stepwise Selection

Data Analysis Assignment Lesson 3

David Chen

Feburary 4th, 2020

Introduction

In this report, we will attempt to apply model selection methods to a simulated data set. After generating 100 observations of a response variable `Y`, we used the best subset selection, forward stepwise selection, and backward stepwise selection methods to determine the best model based on (Mallow's) C_p, BIC (Bayesian information criterion), and adj-R^2 criterion. Ultimately, all methods resulted in the same selected models, although the C_p and BIC criterion selected a four variable model while the adj-R^2 criteria selected a five variable model.

Data

For this analysis, we will be simulating all the data. First, we generate 100 observations of a predictor `X` and a noise vector ϵ based on a standard normal distribution $N(0, 1)$. In order to generate a response variable, `Y`, we will apply the following model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

The beta coefficients were arbitrarily selected, chosen as the following:

1. $\beta_0 = 14$
2. $\beta_1 = 3$
3. $\beta_2 = -2$
4. $\beta_3 = -7$

Examining the individual variables in Figure 1, we find that the predictors appear as expected – `X` is randomly scattered around 0, X^2 has positive values only, and X^3 is condensed around 0 (since most of the values were between -1 and 1).

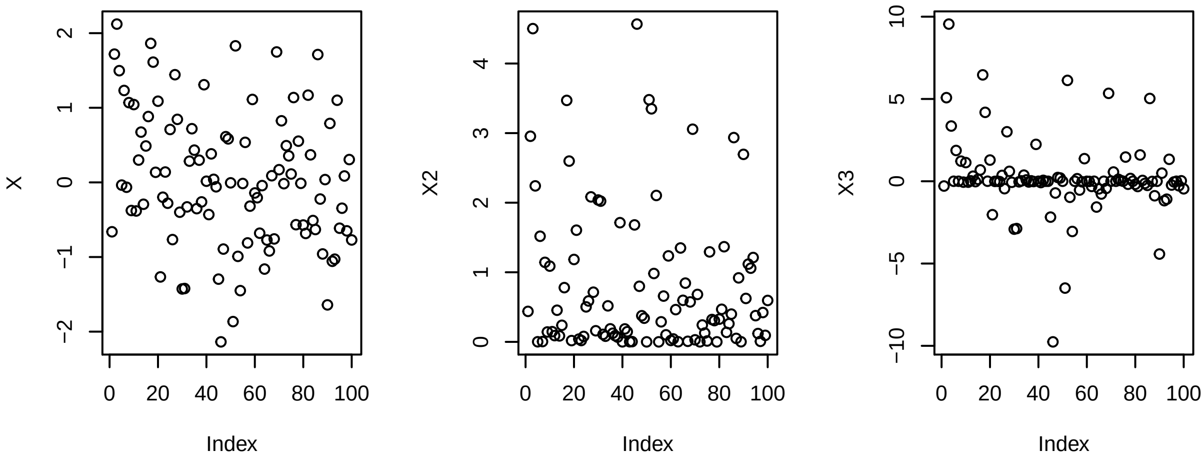

Figure 1: Basic Plots of the Predictors

Examining the relationship between the predictors and `Y` in Figure 2, we observe that only X^3 has a true linear relationship. Moving forward, we should expect X^3 to be the most influential predictor.

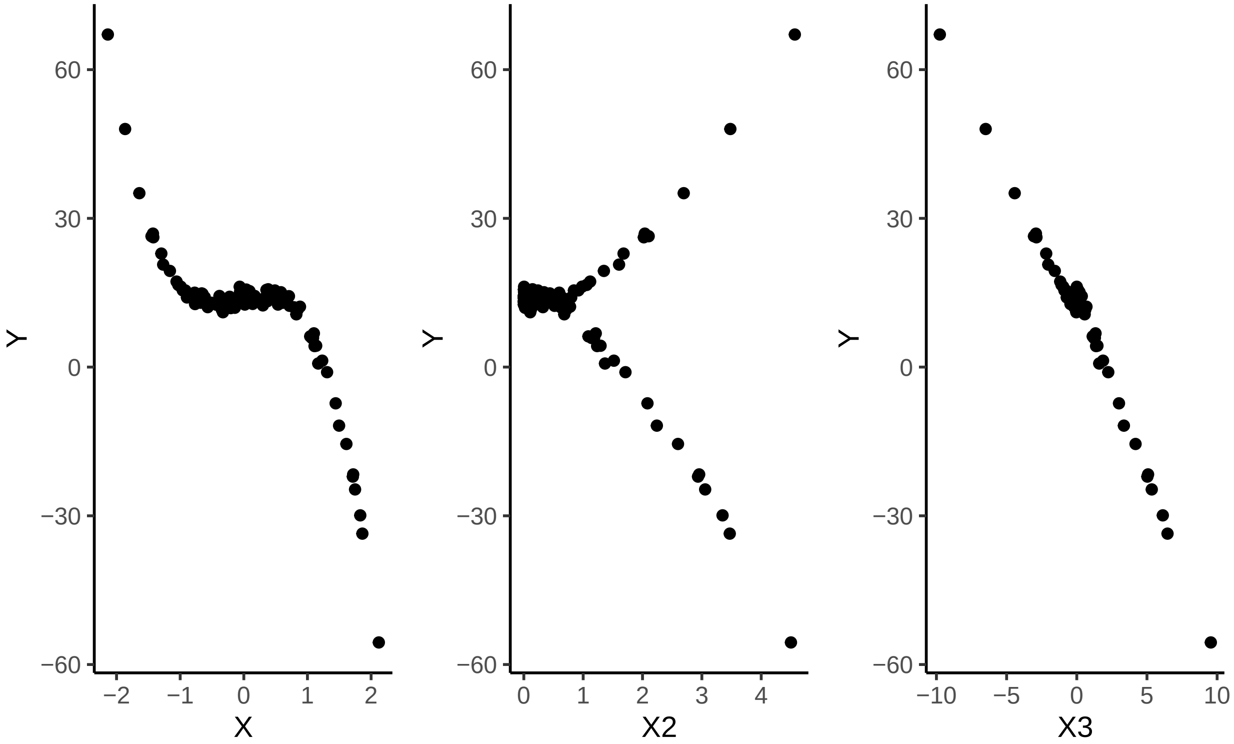

Figure 2: X and Y Relationships

Understanding these relationships, we proceed further with our analyses.

Analyses

For this analysis, we will be using three different model selection methods, using C_p, BIC, and adj-R^2 criterons to determine the optimal models. We will be using the best subset, forward, and backward stepwise selection methods on a model containing X, X^2, ..., X^10. First, we will begin with the best subset selection method.

Best Subset Selection

Using subset selection, 10 separate models are generated, each with an increasing number of variables. With the C_p, BIC, and adj-R^2 criterons, we are able to determine which model and number of predictors is best.

From Figure 3, we can observe that based on C_p, the five variable (predictor) model was selected, while BIC selected three variables. The adj-R^2 criterion seems to select the six variable model, although the values seem to level off after three variables. Recall that for C_p and Bic lower values are optimal, while adj-R^2 desires higher values.

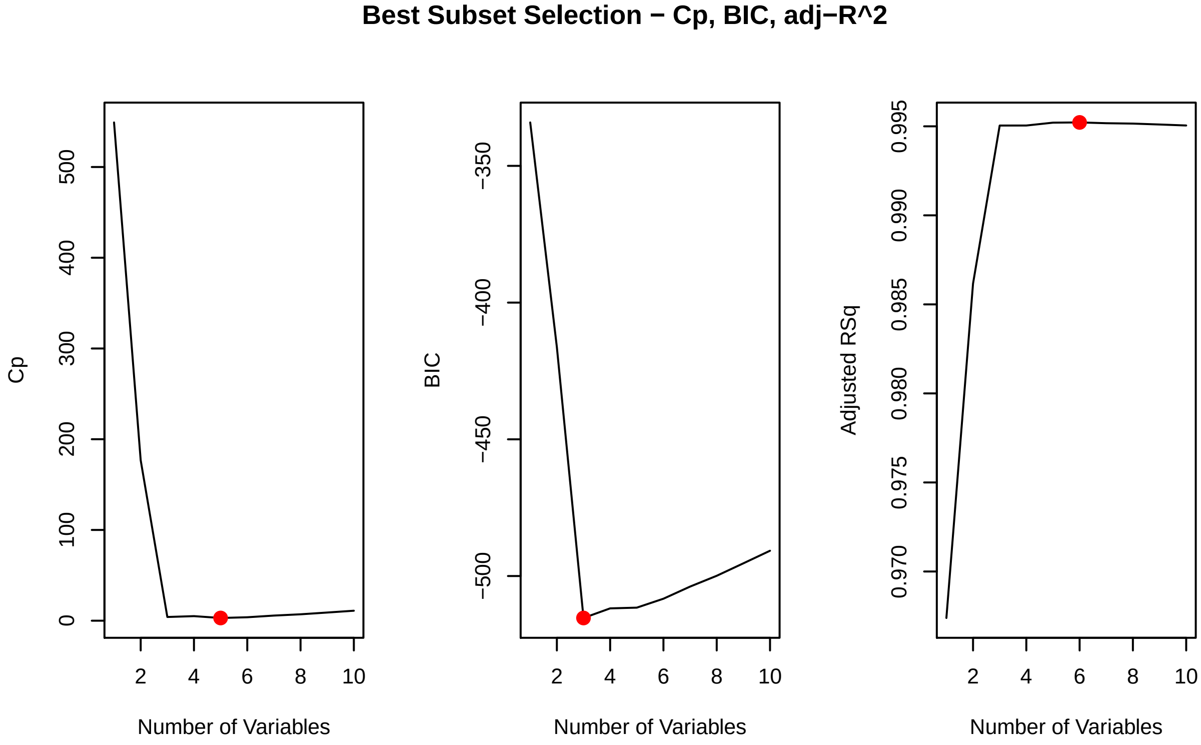

Figure 3: Best Subset Selection

From the C_p criterion, we observe the following five variable model:

$$Y \approx 13.747 + 2.897 * X - 7.050 * X^3 - 2.398 * X^4 - 0.902 * X^6 - 0.103 * X^8$$

The BIC Criterion results in the following three variable model:

$$Y \approx 13.965 + 2.855 * X - 1.945 * X^2 - 7.028 * X^3$$

The adj-R^2 criterion selected the following six variable model:

$$Y \approx 13.756 + 3.184 * X - 7.408 * X^3 - 2.415 * X^4 + 0.075 * X^5 - 0.917 * X^6 - 0.105 * X^8$$

Here, we note that the 3 variable model determined by the BIC criterion is close to the true model ($Y = 14 + 3X - 2X^2 - 7X^3$). It seems that the other two criterion had similiar intercepts and coefficients for X and X^3, but unfortunately the inclusion of other variables masked X^2.

From these models, we can interpret the coefficients to determine each predictor's association with `Y`. For example, for every unit increase in `X`, the value of `Y` decreases by 113.345 units.

Forward Stepwise Selection

Next, we conduct a forward stepwise selection. Recall that for a forward stepwise selection, the algorithm begins with a null model. It then gradually adds one variable at a time, selecting the variable that results in the highest adj-R^2 value. This process continues until all variables are included.

After conducting this analysis and plotting the criterions in Figure 4, we observe some slight differences from the best subset selection results. We note that the C_p criterion now selects the 3 variable model, just like the BIC criterion. This is an improvement, while the adj-R^2 model has a max value at 8 variables. Again, we do note that the adj-R^2 values seem to level off after 3 variables, so that model may have been selected as well.

It seems that the forward selection model has succesfully chosen the 3-variable model for all 3 criterion!

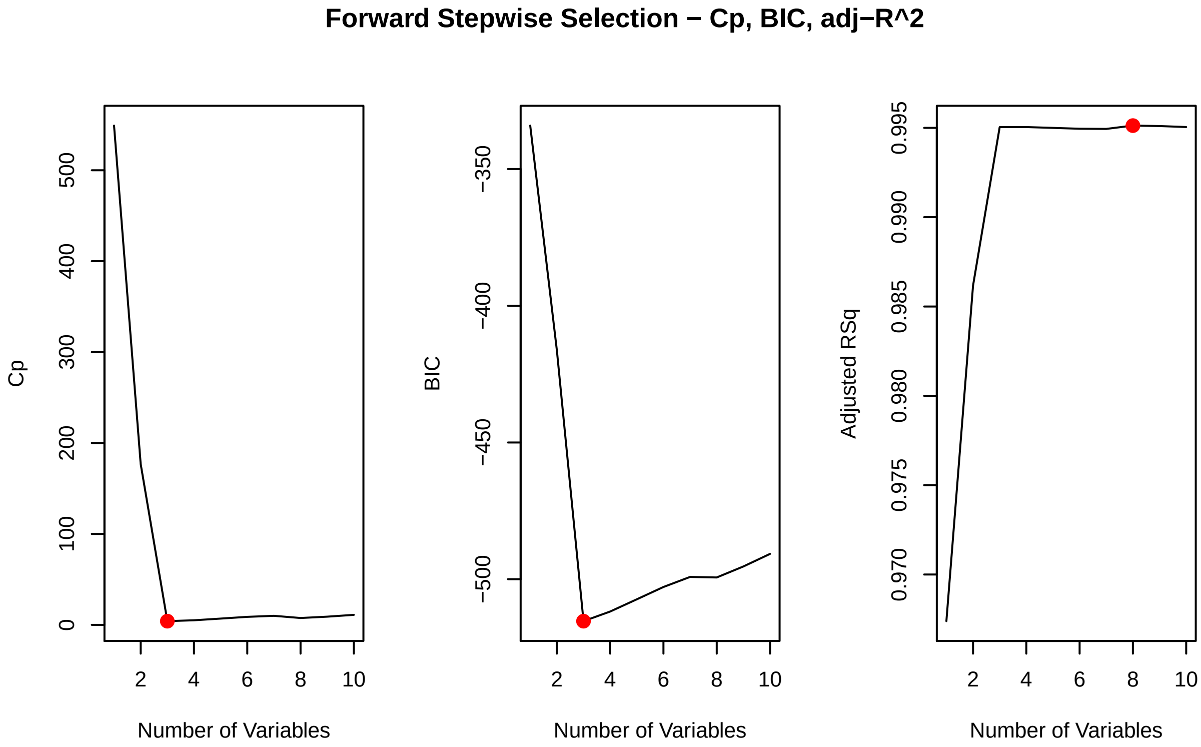

Figure 4: Forward Stepwise Selection

Backward Stepwise Selection

Lastly, we examine the backward stepwise selection method. Recall that for this method, the algorithm begins with a model including every predictor available. It then takes out one predictor at a time, selecting the variable that reduces the adj-R^2 the least. We then eventually reach the one variable model.

Repeating the same procedure as before, we find that the models selected are larger than the forward selection method. As shown in Figure 5, the C_p and BIC criterion select the five variable model, while the adj-R^2 selects the six variable model.

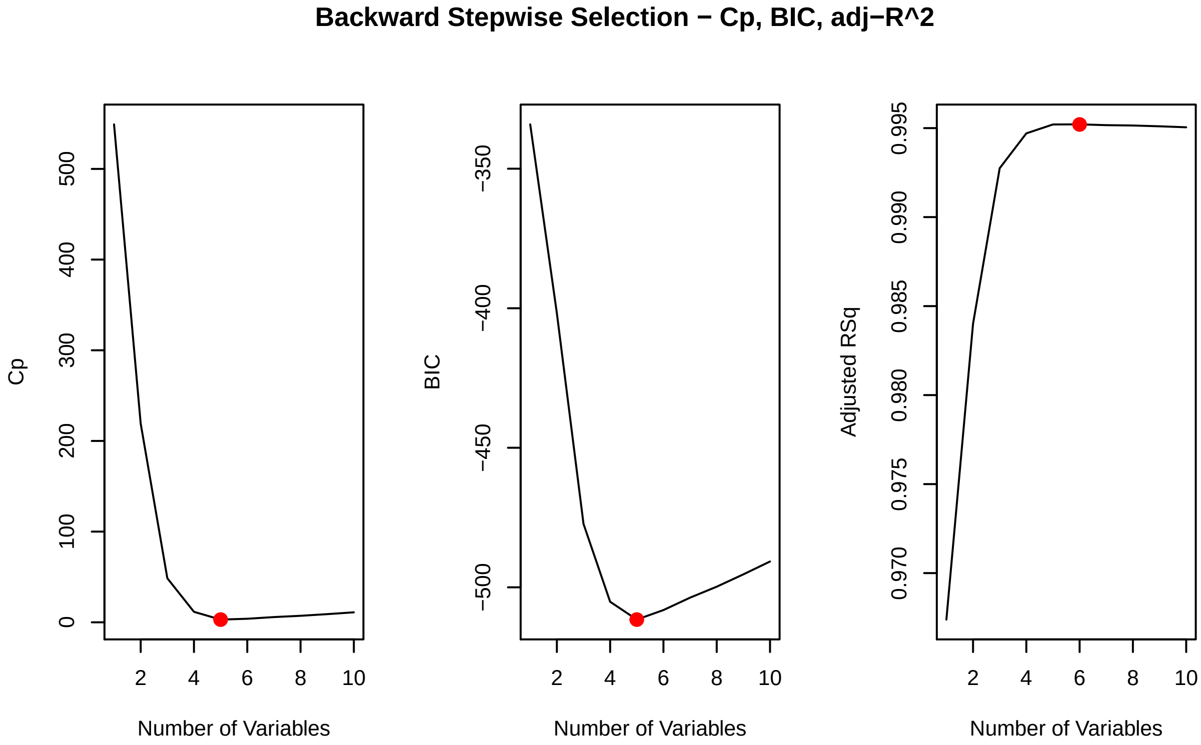

Figure 5: Backward Stepwise Selection

Conclusion

In conclusion, the best subset selection and forward stepwise selection methods were able to fairly accurately estimate the true, 3-variable model, while the backward stepwise selection method selected slightly more predictors than desired. Comparing the three different criterion (C_p, BIC, adj-R^2) the $C - p$ and BIC values were relatively consistent with each other, while the adj-R^2 value seemed to consistently favor a higher number of predictors. However, we do note that the adj-R^2 values consistently leveled off after 3 variables, and thus if we were to manually select the optimal number of predictors, we would have obtained the desired model. Thus, given that each method and classification was relatively close to the true model, this result inspires confidence in their ability to estimate the original coefficients.