Module 10 Assignment

Jared Yu

1. Use any math/stat software (e.g., [www.numbergenerator.org/randomnumbergenerator](http://www.numbergenerator.org/randomnumbergenerator)) of your choice to find a random number generator to randomly select 20 rows of Table B.2. Then do Problem 10.4 (a), (b), (c), (d), (e), page 367 of Textbook, using your generated data.
   1. Use forward selection to specify a subset regression model.

Ans:

The randomly chosen rows for this problem are: 1, 2, 4, 5, 6, 7, 9, 10, 11, 12, 14, 17, 18, 19, 21, 22, 23, 25, 26, and 28.

The “stepAIC()” function in R was used for this task. It calculated from an intercept-only model the AIC value, and then proceeded to add a regressor one-at-a-time to the model. The results can be seen below in Table 1. It can be seen that the regressors are added iteratively one-by-one, and the AIC continues to decrease, and so regressors are continually added until the final model is the full model that includes all five original regressors.

Table The table below shows the result of using the stepAIC() function in R to calculate the forward selection model.

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

* 1. Use backward elimination to specify a subset regression model.

Ans:

The same “stepAIC()” function was used again in R to calculate the backward elimination model for the data. It starts with the full model and then calculates the sum of squares, residual sum of squares, and the AIC for each case where it potentially removes any of the regressors. The result can be seen below in Table 2 (*Note: This function seems to prefer AIC, and so this metric will be used continuously from here on.*). It can be seen that starting from the full model with all regressors (marked None), it continuously had the lowest AIC (and hence the best) score amongst all other possible models with one regressor removed. Removing would bring the AIC to a close level, but not low enough to remove it from the current model. Therefore, the backward elimination method has chosen the same model as in part a) (i.e., the full model with all regressors).

Table The below table shows the results of using the stepAIC() function in R to calculate the backward elimination model.

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

* 1. Use stepwise regression to specify a subset regression model.

Ans:

The same “stepAIC()” function was used again in R to calculate the stepwise regression model for the data. The result is identical to what is seen in part a), so the above Table 1 shows the same output. Table 1 however doesn’t show the steps during stepwise regression, where the function checks first to see if all regressors should be retained. These can be seen in the Code Appendix (within the section marked “Problem 1 (c)”). As with part a), the final model chosen is the full model with all five regressors.

* 1. Apply all possible regressions to the data. Evaluate , , and for each model. Which subset model do you recommend?

Ans:

Using the leaps package and its “regsubsets()” function, the all possible regressions method was used to find the “best” subset. The package only provides by default the and values, so the is calculated based on the values which is by default given. Based on these three metrics, two have selected the full model, and one has selected a reduced model that excludes . Based on the majority vote, and considering the choices previously by the forward and backward methods, I would go with the full model. There is a nuance though, for example is utilized here, rather than . However, when using that method also, it likewise chooses the full model over any other model.

Table The table below shows the results based on the regsubsets() function in R. It shows the best performing model and corresponding metric when utilizing , , and .

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

* 1. Compare and contrast the models produced by the variable selection strategies in part a-d.

Ans:

There was a roughly unanimous decision amongst the different subset selection methods. Using forward selection, backward elimination, stepwise regression, and all possible regressions method, the chosen model has continuously been the full model, that is, . The only exception is with regard to the metric in part d). This is the only time that an alternative model was recommended. Doing some additional investigation, the metric also chooses the same model as , while also chooses the full model. Therefore, it is possible that the model excluding is also viable, but it is uncertain since this model was not chosen nearly unanimously.

1. Use any math/stat software (e.g., [www.numbergenerator.org/randomnumbergenerator](http://www.numbergenerator.org/randomnumbergenerator)) of your choice to find a random number generator to randomly select 20 rows of Table B.1. Then do Problem 11.1 (a), (b), page 386 of Textbook, using your generated data.
   1. Calculate the PRESS statistic for this model. What comments can you make about the likely predictive performance of this model?

Ans:

* 1. Delete half the observations (chosen at random), and refit the regression model. Have the regression coefficients changed dramatically? How well does this model predict the number of games won for the deleted observations?

Ans: