

MATH 424: Homework Chapter 6:

Model Selection Methods

Due on Thursday, November 2, 2017

Kafai 11:10am

Jonathan Dombrowski

Contents

| | |
|---|----------|
| Q 1 | 3 |
| 1.a Stepwise Methodology | 3 |
| 1.b Mallows' Cp Selection Methodology | 5 |
| Q 2 | 7 |
| 2.a Stepwise | 7 |
| 2.b Mallows Cp | 7 |

Q 1

1. Write-up, in detail, the steps taken to do the following regressions in model building. Make sure you clearly state the CRITERIA for model selection.
 - a) Stepwise selection
 - b) Mallows Cp selection
2. Use the HW6.txt data and your answers to question 1 on how SAS/R is proceeding with the model selection for question 1, parts a, and b. The provided data set has the following variables, y , x_1 , x_2 , x_3 , x_4 , and x_5 .

1.a Stepwise Methodology

The StepWise method requires starting with a y , and then creating models with y, x_i for $\forall i \in n$, where n is the number of predictors in the full model.

After creating models, you test each beta as follows:

$$H_0 : \hat{\beta}_1 = 0$$

$$H_a : \hat{\beta}_1 \neq 0$$

For the stepwise method, the threshold for inputting and removing a term is $\alpha = 0.15$. After completing the tests, the $\hat{\beta}_i$ with the highest passing t-value is the one to be selected, and will seed the next iteration of models, we will call this $\hat{\beta}_1$. If no value passes, life is easy, quit and go home. For example, below is the appended t-values for the first 5 single-term models, followed by an expanded summary of the x_1 model.

```
(Intercept)           x1
  2.171654    7.736978
(Intercept)           x2
  4.544396    2.492404
(Intercept)           x3
  3.196224    4.221919
(Intercept)           x4
  1.709245    3.868098
(Intercept)           x5
  2.9021807   0.8381174

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept) 14.37632   6.61999   2.172   0.0385 *
x1          0.75461   0.09753   7.737 1.99e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1
```

After finding the most statistically significant term you will then use it for the base of the next iteration, and one by one, creating new models and appending the remaining terms to the model, respectively

The stopping criteria for this algorithm is that you have either ran out of variables, or that all of the variables you have left to add fail to reject the null hypothesis in determining to whether or not to add the term.

After performing the exact same t-test as mentioned in the previous step, and get a subset of x's with t-values that produce rejections for the second added term, which we will call $\hat{\beta}_2$. Before continuing the iterations and moving on to the next regressions, we test $\hat{\beta}_1$ with the same criteria on the model with the highest $\hat{\beta}_2$ t-value, to ensure that adding the possible $\hat{\beta}_2$ does not adversely affect $\hat{\beta}_1$. In this first case, we only add the term to the model if all of the β 's reject the null hypothesis.

If all of the $\hat{\beta}_2$ candidates fail to reject, then the model is complete and we can stop without adding $\hat{\beta}_2$. The

third case is that there is one or more $\hat{\beta}_2$ candidates, which is to say that they rejected the null hypothesis, yet the tests on the $\hat{\beta}_1$ retest failed. In this case, we remove the x_1 term from the model and retry the previous step with the next most-likely candidate.

```
(Intercept)           x1           x2
 2.1406590    6.5362166   -0.3860845
(Intercept)           x1           x3
 1.397899    5.431563    1.571324
(Intercept)           x1           x4
 1.4232380    5.3542807   0.4697813
(Intercept)           x1           x5
 1.27559502   7.45989926  0.01402383
```

This last process repeats until there are either no more terms, or the stopping clause of no more terms satisfy the requirements of having a t-value that rejects the null hypothesis at $\alpha = 0.15$. At the end of this algorithm we will end up with a model that has only statistically significant $\hat{\beta}$'s left in the model.

Finally, after having $\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_3$, we build individual models for x_i ; $i = 1$ to n . Below is an output from R for the t-values from the t-test.

```
(Intercept)           x1           x3           x2
 1.5383686    5.2964340   1.7070208  -0.7985149
(Intercept)           x1           x3           x4
 1.2670494    4.7918486   1.4745203  -0.1569606
(Intercept)           x1           x3           x5
 0.870616464   5.269791625  1.541899619  0.006650698
```

We can see that there are no terms left with t-values within the rejection region. Therefore the second criteria for stopping the model creation has been completed and we can conclude that the model after manually completing stepwise selection is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_3$$

The R output for the summary of this model is as follows :

```
Call:
lm(formula = y ~ x1 + x3)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.5568 -5.7331  0.6701  6.5341 10.3610 

Coefficients:
            Estimate Std. Error t value Pr(>t)
(Intercept)  9.8709    7.0612   1.398   0.174
x1          0.6435    0.1185   5.432 9.57e-06 ***
x3          0.2112    0.1344   1.571   0.128
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1

Residual standard error: 6.817 on 27 degrees of freedom
Multiple R-squared:  0.708, Adjusted R-squared:  0.6864 
F-statistic: 32.74 on 2 and 27 DF,  p-value: 6.058e-08
```

With the model equating to :

$$\hat{y} = 9.87 + 0.64x_1 + 0.21x_3$$

1.b Mallows' Cp Selection Methodology

The process for Mallows Cp selection method is similar to that of stepwise selection except that this method does not result in a definite ‘best model’; instead, it is subjective and is determined by the practitioners’ creation of the stopping rules. Without the definition of stopping rules, the best subset for each possible number of variables will be calculated. By *best* we mean to say that the subset of models with the lowest Cp value will be selected at each stage of the algorithm; each number of variables. Mallows Cp Statistic:

$$C_p = \frac{SSE_p}{MSE_k} + 2(p+1) - n$$

Evaluating all of the possible subset Cp values allows for one to see at which number of variables has the lowest SSE, therefore minimizing the error compared to the full model. Minimizing this ratio of error with the smallest model is the goal of this selection.

The proposed stopping rules are :

1. Stop adding terms when there are no more terms left to add
2. At every stage of the algorithm; after each iteration of calculating the Cp statistics, find the minimum Cp value model and record it. If the next iteration with p+1 terms has a higher minimum Cp value, return to the previous model with the lowest Cp value.

We are trying to build a model to minimize SSE_r vs. MSE_{full} , it makes no sense to add more terms to the model if it raises this ratio and degrades the accuracy of the model. There is no justification for adding another term to the regression if the minimum Cp statistic is raised. When R uses this method of selection, it does not return a definite model as the measure is subjective. To show an example of the stopping rules in action:

```
[1] "Number of Var,      Cp Stat"
[1] "Step 1 : Single Variable Models"
[1] 2.0000000 0.8701654
[1] 2.00000 43.00589
[1] 2.00000 25.51894
[1] 2.00000 28.95146
[1] 2.00000 56.25206
[1] "Step 2 : Two Variable Models"
[1] 3.0000000 2.722635
[1] 3.0000000 0.6188538
[1] 3.000000 2.652313
[1] 3.00000 2.86997
[1] 3.00000 25.95579
[1] 3.000 28.149
[1] 3.00000 44.24921
[1] 3.00000 22.31702
[1] 3.00000 26.91426
[1] 3.00000 30.52372
[1] "Step 3 : Three Variable Models"
```

```
[1] 4.000000 2.029551
[1] 4.000000 4.457372
[1] 4.000000 4.721477
[1] 4.000000 2.595548
[1] 4.000000 2.618812
[1] 4.000000 4.627902
[1] 4.000000 23.63883
[1] 4.00000 27.53214
[1] 4.00000 29.77383
[1] 4.00000 24.29786
[1] "The minimum value calculated is not lower than the previous step, so we can stop"
[1] "For sake of being verbose, we will calculate the other values"
[1] "Step 4 : Four Variable Models"
[1] 5.000000 4.011064
[1] 5.000000 4.027364
[1] 5.000000 6.434691
[1] 5.000000 4.590846
[1] 5.000000 25.61211
[1] "Step 5"
[1] 6 6
[1] "summary"
[1] "best model with respect to Cp Value"
[1] "y~x1+x3"
[1] 3.0000000 0.6188538
```

Thus completing the verbose version of this ruleset.

$$\hat{y} = 9.87 + 0.64x_1 + 0.21x_3$$

Q 2

2.a Stepwise

Running the R-package stepwise algorithm from the olsrr package yields the following:

```
Stepwise Selection Method
```

```
Candidate Terms:
```

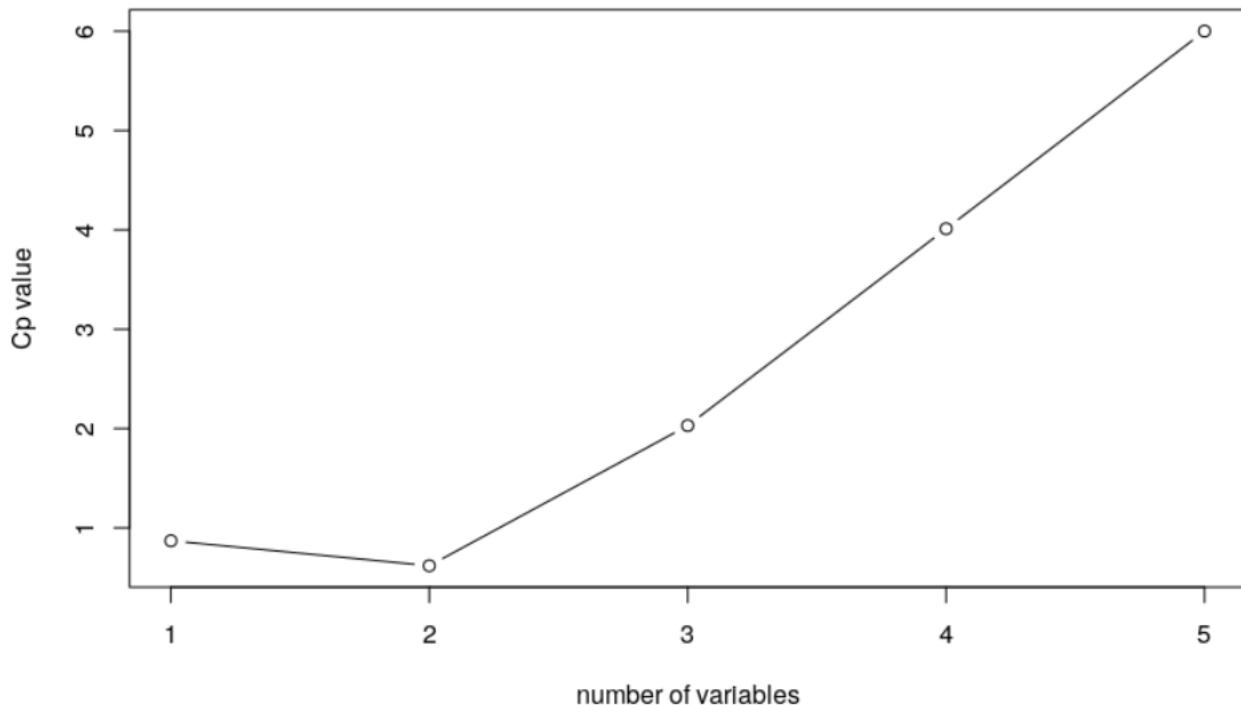
```
1 . V2  
2 . V3  
3 . V4  
4 . V5  
5 . V6
```

| Stepwise Selection Summary | | | | | | |
|----------------------------|----------|-------------------|----------|------------------|--------|----------|
| Step | Variable | Added/ Removed | R-Square | Adj. R-Square | C(p) | AIC |
| | RMSE | | | | | |
| 1 | V2 | addition | 0.681 | 0.670 | 0.8700 | 205.7638 |
| 6.9933 | | | | | | |
| 2 | V4 | addition | 0.708 | 0.686 | 0.6190 | 205.1387 |
| 6.8168 | | | | | | |

The algorithm yields that V2 and V4 (x1 and x3) are the components in the final model. Which is the same result we come to in the manual calculation.

2.b Mallows Cp

Running the packaged R method to do this analysis yields the same minimum values for each number of variables. As this method is exhaustive, it will calculate all values of all combinations of variables. The graph below shows the output, while the table below that shows the individual values.



Summary of the values:

```
[1] 0.8701654 0.6188538 2.0295511 4.0110637 6.0000000
```

A readout of the R-output using the package OSLRR:

| Index | N | Predictors | 'R-Square' | 'Adj. R-Square' | 'Mallow's Cp' |
|-------------------------|-------|------------|------------|-----------------|---------------|
| <int> | <int> | <chr> | <chr> | <chr> | <chr> |
| 1 | 1 | V2 | 0.68131 | 0.66993 | 0.87017 |
| 2 | 2 | V4 | 0.38897 | 0.36715 | 25.51894 |
| 3 | 3 | V5 | 0.34826 | 0.32499 | 28.95146 |
| 4 | 4 | V3 | 0.18158 | 0.15235 | 43.00589 |
| 5 | 5 | V6 | 0.02447 | -0.01037 | 56.25206 |
| 6 | 6 | V2 V4 | 0.70802 | 0.68639 | 0.61885 |
| 7 | 7 | V2 V5 | 0.68390 | 0.66048 | 2.65231 |
| 8 | 8 | V2 V3 | 0.68306 | 0.65959 | 2.72264 |
| 9 | 9 | V2 V6 | 0.68132 | 0.65771 | 2.86997 |
| 10 | 10 | V4 V5 | 0.45067 | 0.40998 | 22.31702 |
| # ... with 21 more rows | | | | | |

Which are the same as the minimum values calculated in the verbose manual calculation. We can conclude that the lowest SSE for the reduced model occurs when the variable number is set to zero. Without giving the program more specific objective instructions, this is the output of the R module. The final model calculated in the Cp selection method is $V1 V2 + V4 \approx \hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_3$ which is the same predicted model from all other parts of the homework for the optimal model for minimizing the Mallows' Cp statistic.