CS5565

Alex Dowell

Stat Learning: Features and Mode Selection Lab

Part A)

Generate the list of selected features generated by the "forward" and "backward" selection methods for a new value of "nvmax" (number of max variables) between 9 and 18.

What did you observe as a result of this change?

```
1 subsets of each size up to 12
Selection Algorithm: forward
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks
1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
4  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
5  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    " "   "*"  " "
6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "   "*"  " "
7  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    " "   "*"  "*"
8  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   "*"  "*"
9  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
10  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
11  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
12  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
          LeagueN DivisionW PutOuts Assists Errors NewLeagueN
1  ( 1 )  " "     " "       " "     " "     " "    " "
2  ( 1 )  " "     " "       " "     " "     " "    " "
3  ( 1 )  " "     " "       "*"     " "     " "    " "
4  ( 1 )  " "     "*"       "*"     " "     " "    " "
5  ( 1 )  " "     "*"       "*"     " "     " "    " "
6  ( 1 )  " "     "*"       "*"     " "     " "    " "
7  ( 1 )  " "     "*"       "*"     " "     " "    " "
8  ( 1 )  " "     "*"       "*"     " "     " "    " "
9  ( 1 )  " "     "*"       "*"     " "     " "    " "
10  ( 1 ) " "     "*"       "*"     "*"     " "    " "
11  ( 1 ) "*"     "*"       "*"     "*"     " "    " "
12  ( 1 ) "*"     "*"       "*"     "*"     " "    " "
```

```
1 subsets of each size up to 12
Selection Algorithm: backward
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI CWalks
1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
4  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "  " "
5  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "  " "
6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "  " "
7  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "  "*"
8  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   "*"  "*"
9  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
10  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
11  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
12  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*"  "*"
          LeagueN DivisionW PutOuts Assists Errors NewLeagueN
1  ( 1 )  " "     " "       " "     " "     " "    " "
2  ( 1 )  " "     " "       " "     " "     " "    " "
3  ( 1 )  " "     " "       "*"     " "     " "    " "
4  ( 1 )  " "     " "       "*"     " "     " "    " "
5  ( 1 )  " "     " "       "*"     " "     " "    " "
6  ( 1 )  " "     "*"       "*"     " "     " "    " "
7  ( 1 )  " "     "*"       "*"     " "     " "    " "
8  ( 1 )  " "     "*"       "*"     " "     " "    " "
9  ( 1 )  " "     "*"       "*"     " "     " "    " "
10  ( 1 ) " "     "*"       "*"     "*"     " "    " "
11  ( 1 ) "*"     "*"       "*"     "*"     " "    " "
12  ( 1 ) "*"     "*"       "*"     "*"     " "    " "
```

Nvmax was set to 12 for "forward" and "backward" selection, which means the algorithm will include up to 12 predictor variables in the model. The summary now shows the best model for each number of predictors (1 to 12). As a result of changing the nvmax from 19 to 12, the algorithm now only considers models with up to 12 predictor variables instead of 19. The output becomes more compact, and you can focus on models with fewer predictor variables. This can be beneficial if you want to avoid overfitting and create a more interpretable model.

Perform the Ridge Regression for the given line again using a new value (replace '50') of Lambda equal to the last **3 (non-zero) digits** of your student ID.

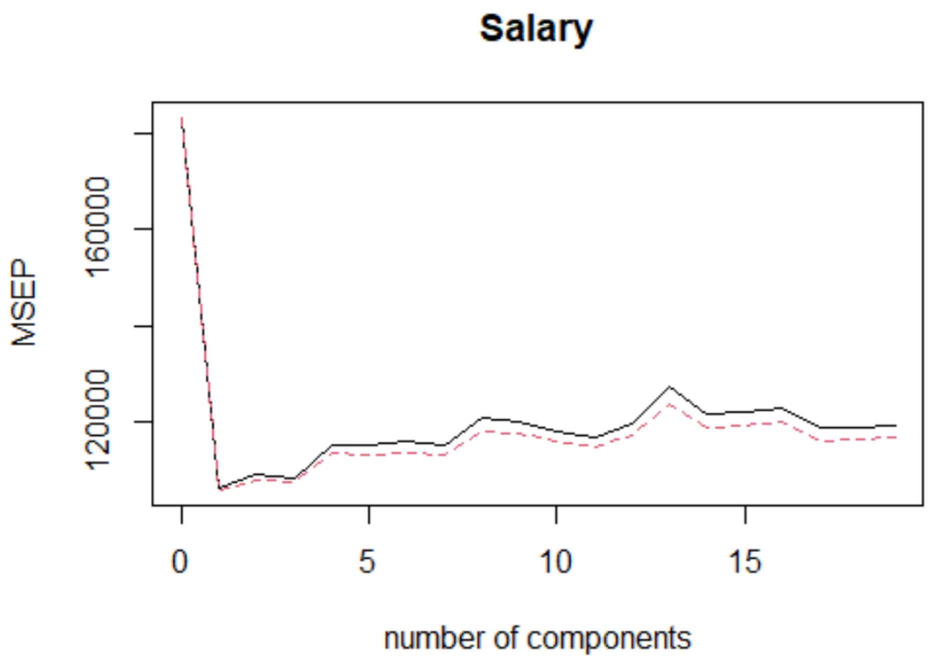
```
predict(ridge.mod, s = 50, type = "coefficients")[1:20, ]
```

How did this change impact the quality of your shrinkage method? What is the *theoretical* impact of selecting a larger value for Lambda for the L2 method?

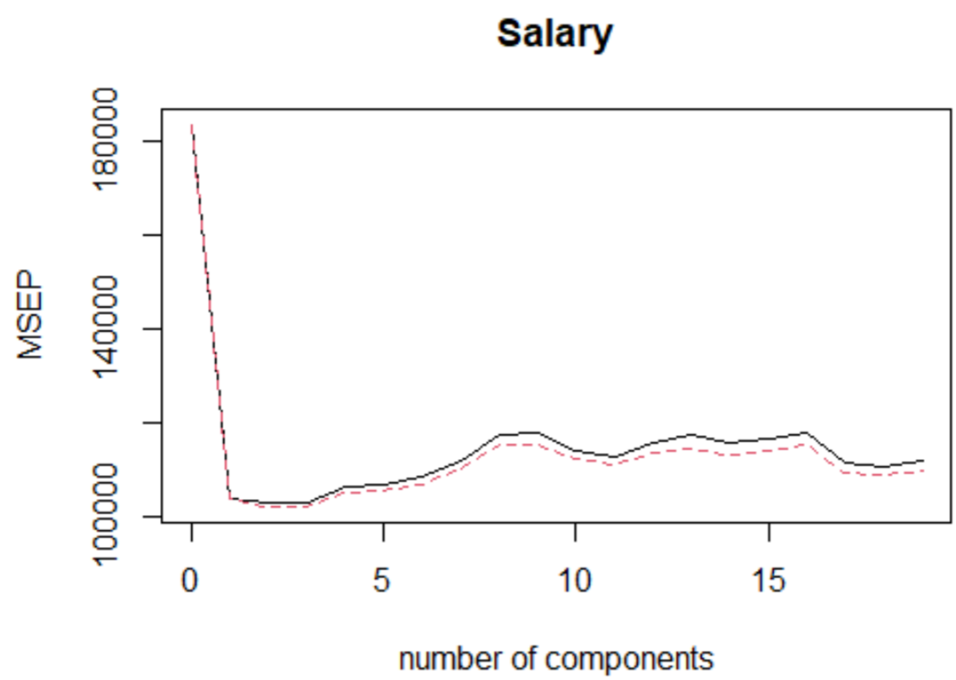
Increasing the value of Lambda in Ridge Regression, such as from 50 to 379, shrinks the magnitude of the coefficients further towards zero, as it prioritizes minimizing the coefficients to reduce the penalty term, helping to prevent overfitting. However, if Lambda is too large, the model may become overly simplistic and underfit the data, performing poorly on both training and testing datasets. In essence, selecting a larger Lambda value in the L2 method aims to balance preventing overfitting and avoiding underfitting. The optimal Lambda value can be determined using cross-validation, which helps select the regularization strength that achieves a balance between model complexity and generalization performance.

Change the **seed** on the first function used in Partial Least Squares to the last 3 (non-zero) digits of your student.

What changes did you observe? Capture a screenshot of your MSEP graph.



MSEP from the original seed(1)



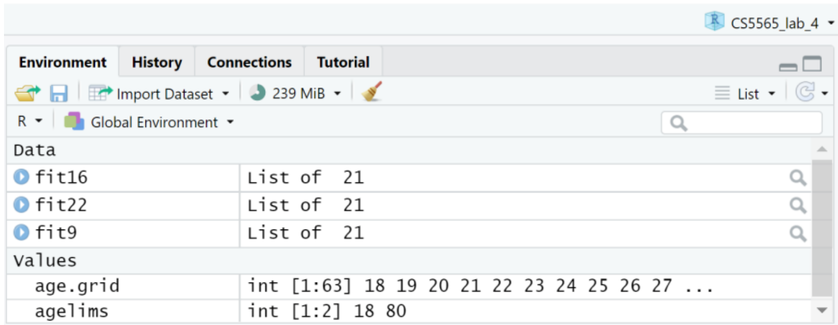
MSEP from seed(379)

Changing the seed affected the random aspects of the cross-validation process used in the Partial Least Squares (PLS) regression model. The seed value determines the initial state of the random number generator, so when the seed changed from 1 to 379, the random number sequence changed aswell. Consequently, this resulted in slightly different cross-validation folds or splits in the dataset.

These changes may lead to slightly different model performance metrics, such as mean squared error of prediction (MSEP). This can be observed when comparing the above MSEP plots above. Although both curves are both curves are within the same general MSEP range there is variation in them.

Part B)

of your plot and console for **wage** vs **age** for Smooth Spline with **df 16** (as given in the lab) and **two new fits** of **df 9** and **22**. (substitute the values as given in the assignment)
What would you conclude from the results generated from the fits given by df 9,16,22?



From the results generated by the fits with df = 9, 16, and 22, we can conclude that a lower degree of freedom (df = 9) results in a smoothing spline that is less flexible and might underfit the data, not capturing all the underlying patterns. In contrast, a higher degree of freedom (df = 22) leads to a more flexible smoothing spline that might overfit the data, capturing too much noise and fitting to random fluctuations. With df = 16, the smoothing spline seems to strike a balance between flexibility and overfitting, capturing the overall pattern in the data without fitting too closely to noise.