**Homework 3**

Xiao-Qun Wang, Zili Ma

**Q1. We perform best subset, forward stepwise, and backward stepwise selection on a single data set. For each approach, we obtain p+1 models containing 0,1,2,⋯,p predictors. Explain your answers :**

1. **Which of the three models with *k* predictors has the smallest training RSS ?**

Best subset selection: The model with k predictors is the model with the smallest RSS among all the models with k predictors. Forward stepwise selection: The model with k predictors is the model with the smallest RSS among the p−k models which augment the predictors in with one additional predictor.   
Backward stepwise selection: The model with k predictors is the model with the smallest RSS among the k models which contains all but one of the predictors in .So, the model with k predictors which has the smallest training RSS is the one obtained from best subset selection as it is the one selected among all k predictors models.

1. **Which of the three models with *k* predictors has the smallest test RSS ?**

It is hard to say, the best subset selection may have the smallest test RSS because it takes into account more models than the other methods. However, the other methods might also pick a model with smaller test RSS by sheer luck.

1. **True or False :**
2. **The predictors in the *k*-variable model identified by forward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by forward stepwise selection.**

True. The model with (k+1) predictors is obtained by augmenting the predictors in the model with k predictors with one additional predictor.

1. **The predictors in the *k*-variable model identified by backward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by backward stepwise selection.**

True. The model with k predictors is obtained by removing one predictor from the model with (k+1) predictors.

1. **The predictors in the *k*-variable model identified by backward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by forward stepwise selection.**

False. There is no direct link between the models obtained from forward and backward selection.

1. **The predictors in the *k*-variable model identified by forward stepwise are a subset of the predictors in the (*k*+1)-variable model identified by backward stepwise selection.**

False. There is no direct link between the models obtained from forward and backward selection.

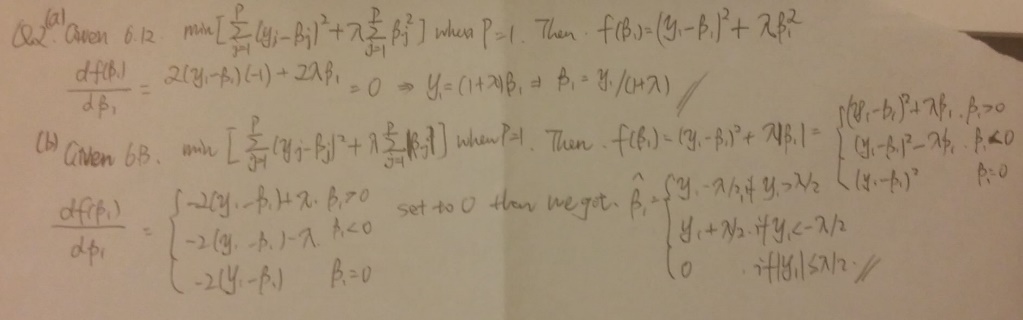
1. **The predictors in the *k*-variable model identified by best subset are a subset of the predictors in the (*k*+1)-variable model identified by best subset selection.**

False. The model with (k+1) predictors is obtained by selecting among all possible models with (k+1) predictors, and so does not necessarily contain all the predictors selected for the k-variable model.

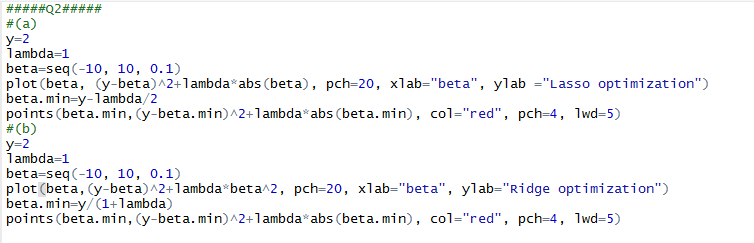
**Q2. We will now explore (6.12) and (6.13) further.**

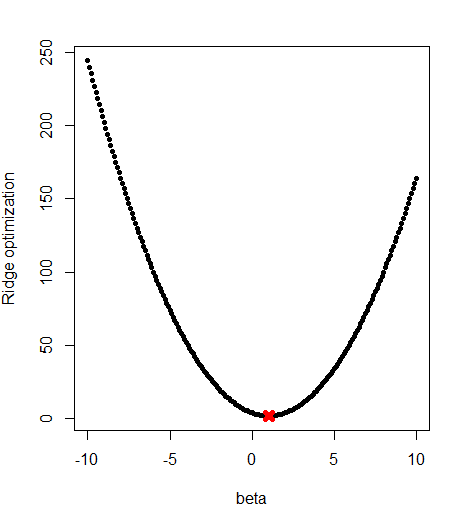
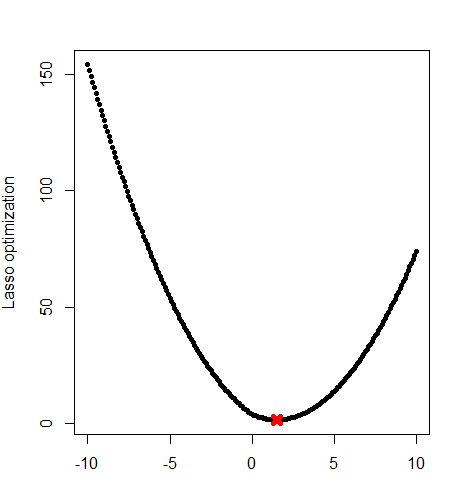
1. **Consider (6.12) with p = 1. For some choice of y1 and λ > 0, plot (6.12) as a function of β1. Your plot should confirm that (6.12) is solved by (6.14).**
2. **Consider (6.13) with p = 1. For some choice of y1 and λ > 0, plot (6.13) as a function of β1. Your plot should confirm that (6.13) is solved by (6.15).**

**Ans:**



By following codes, we can verify the formulas give us the right β1 values to optimize the method of Lasso and Ridge, respectively.





**Q3. In this exercise, we will generate simulated data, and will then use this data to perform best subset selection.**

1. **Use the *rnorm()* function to generate a predictor X of length n = 100,**

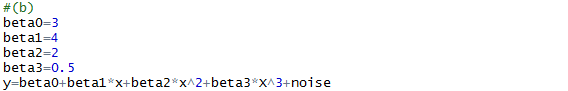
**as well as a noise vector of length n = 100.**



1. **Generate a response vector Y of length n = 100 according to the model**

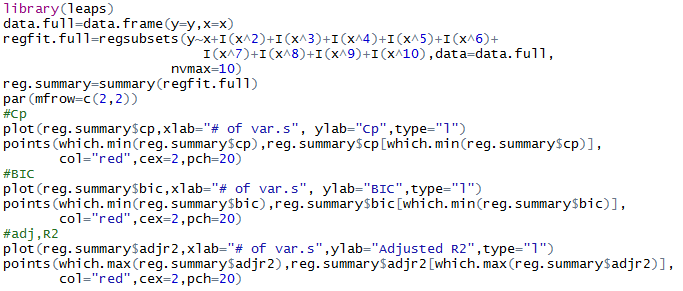
**Y = β0 + β1X + β2X2 + β3X3 + ε**

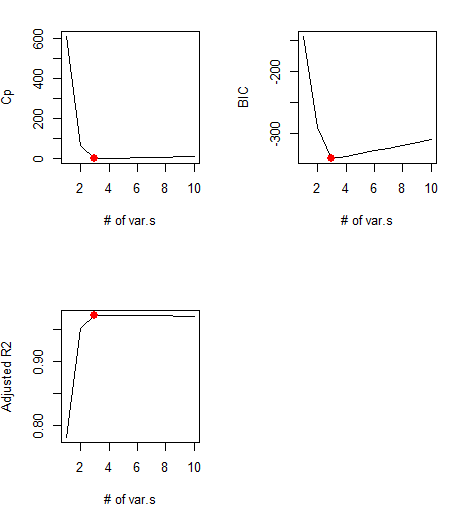
**where β0, β1, β2, and β3 are constants of your choice.**



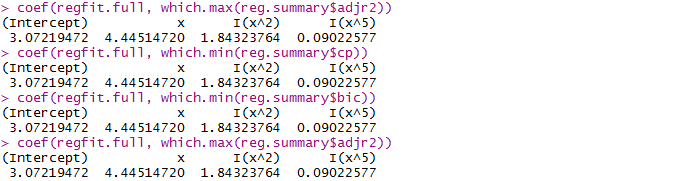
**Ans:** we set beta0 = 3, beta1 = 4, beta2 = 2, beta3 = 0.5.

1. **Use the *regsubsets()* function to perform best subset selection in order to choose the best model containing the predictors X, X2, . . . , X10. What is the best model obtained according to C*p*, BIC, and adjusted R2? Show some plots to provide evidence for your answer, and report the coefficients of the best model obtained. Note you will need to use the *data.frame()* function to create a single data set containing both X and Y .**





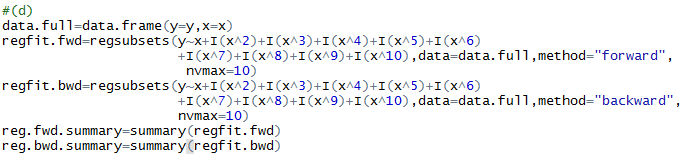
**Ans:** All the three index indicate that 3 variables (x, x2, x5) are the best fitting model here. The chosen best variables and their coefficients are shown as follows.



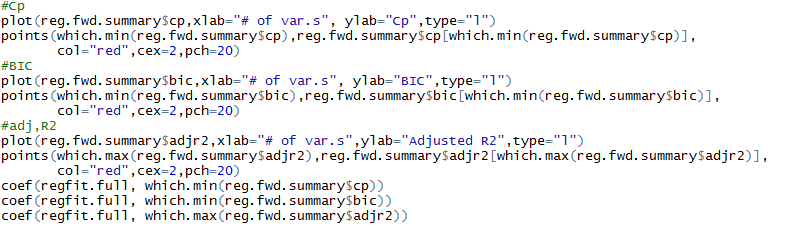
1. **Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c)?**

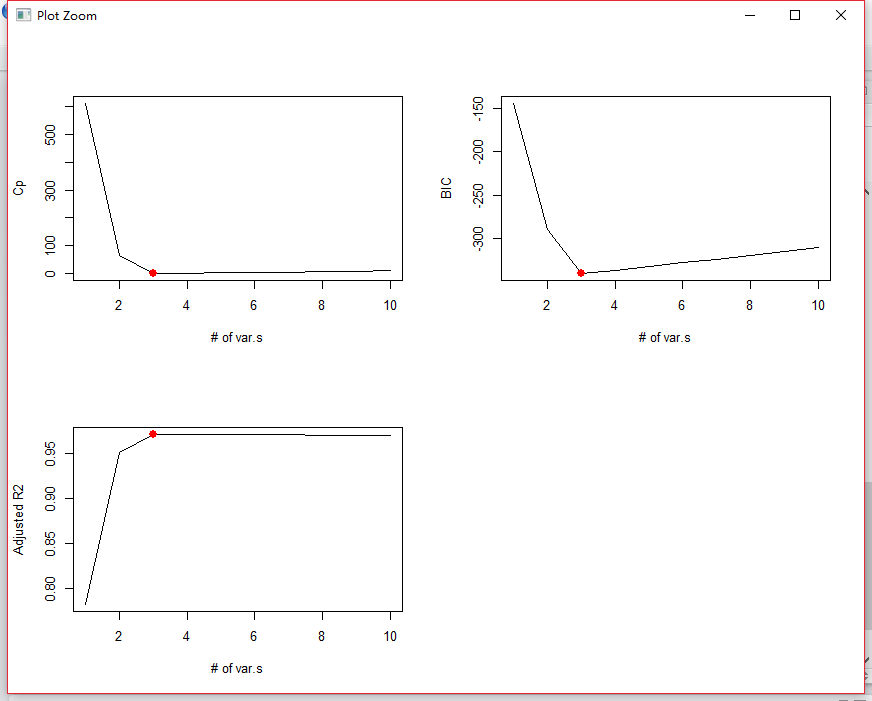
**Ans:**

**Codes for forward and backward selection.**

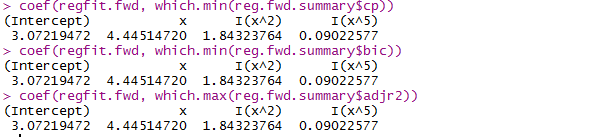


**Forward method summary:**

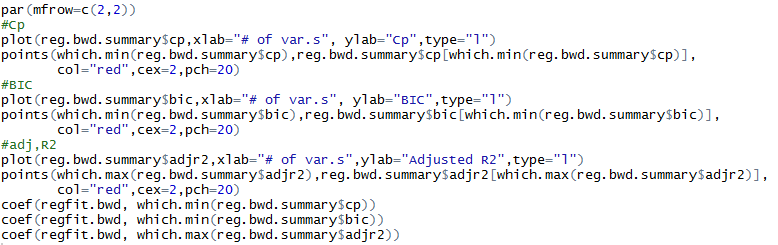


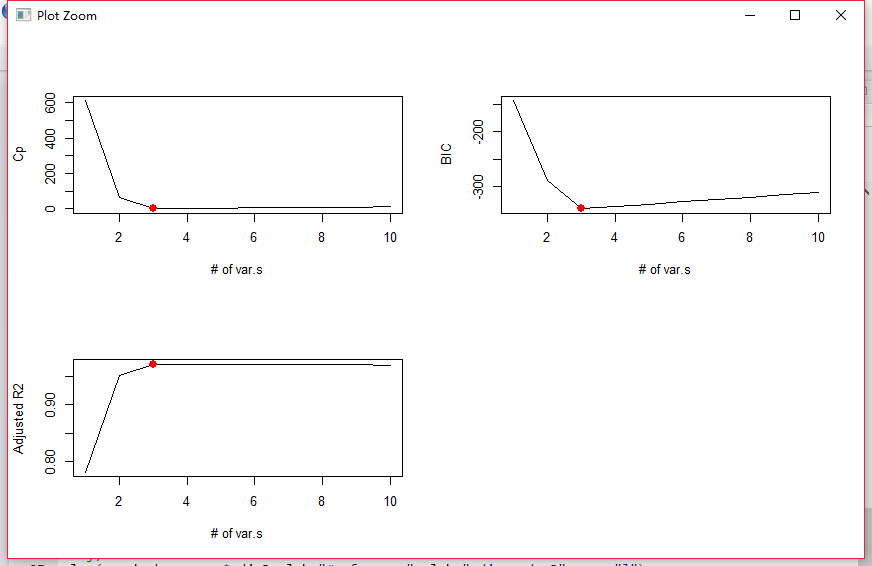


All the three index indicate that 3 variables (x, x2, x5) are the best fitting model here. The chosen best variables and their coefficients are shown as follows.

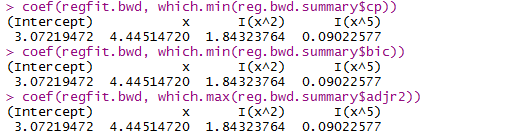


**Backward method summary:**



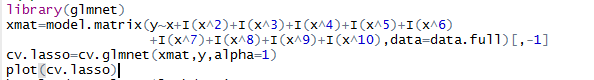


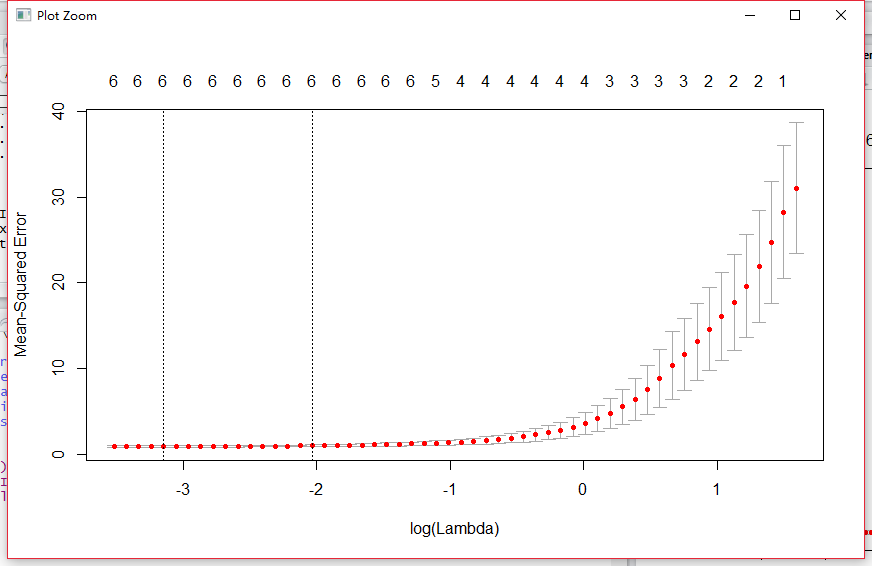
All the three index indicate that 3 variables (x, x2, x5) are the best fitting model here. The chosen best variables and their coefficients are shown as follows.



In these variable selection methods, all three methods select same variables (x, x2, x5).

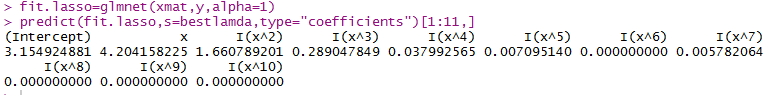
1. **Now fit a lasso model to the simulated data, again using *X, X*2*, . . . , X*10 as predictors. Use cross-validation to select the optimal value of *λ*. Create plots of the cross-validation error as a function of *λ*. Report the resulting coefficient estimates, and discuss the results obtained.**







**Ans:** Our model’s best lamda value is 0.0429 according to the cv value. By this value we can have the model with coefficients as follows. LASSO selects 6 variables (x,x2,x3,x4,x5,x7).



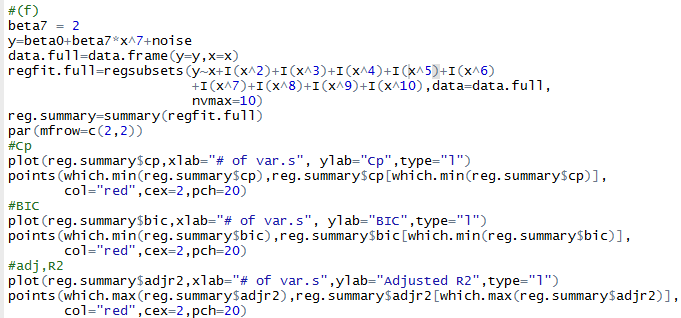
1. **Now generate a response vector *Y* according to the model**

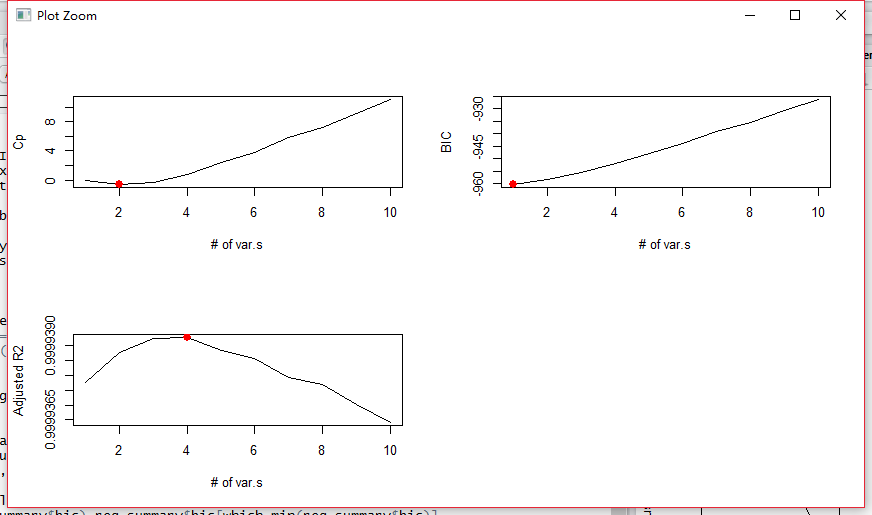
**Y *=* β0 *+* β7X7 *+ ε*,**

**and perform best subset selection and the lasso. Discuss the results obtained.**

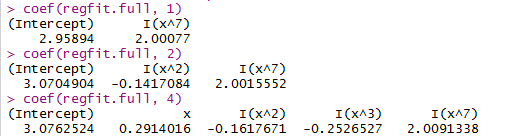
**Ans:**

**Best subset:**

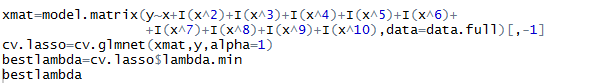


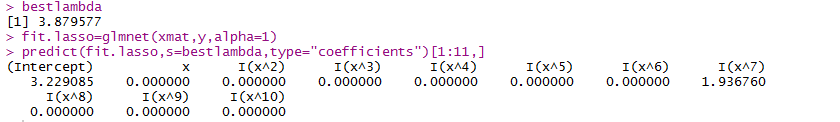


From best subsets method, three index do not give us the same number of variables. Cp implies 2 variables; BIC gives 1 variable, adjusted R2 uses 4 variables. The detail is shown below.



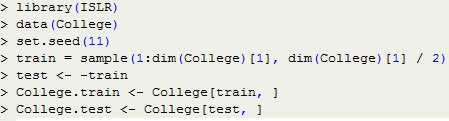
**LASSO:**

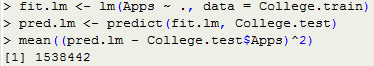




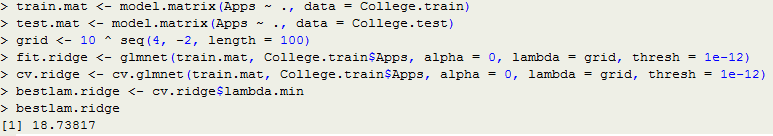
By the method of LASSO, we achieve the best model matching to our simulated question. In this case, LASSO performed better than best subset method.

**Q4 In this exercise, we will predict the number of applications received using the other variables in the “College” data set.**

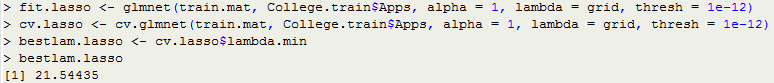
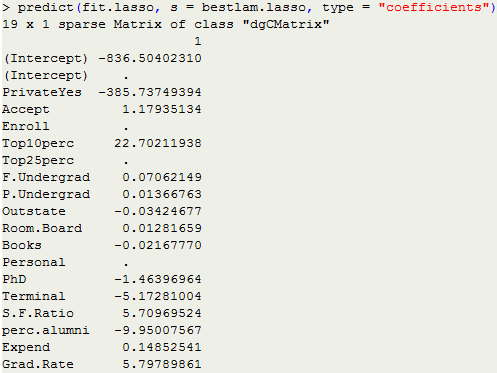
1. **Split the data set into a training and a test set.**
2. **Fit a linear model using least squares on the training set, and report the test error obtained.**

****The test MSE is 1.538442\*10^6

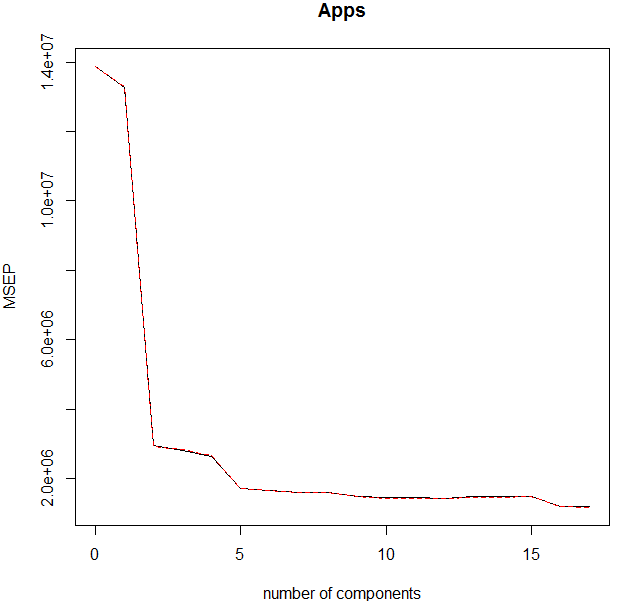
1. **Fit a ridge regression model on the training set, with *λ* chosen by cross-validation. Report the test error obtained.**

**  
**The test MSE is 1.608859\*10^6

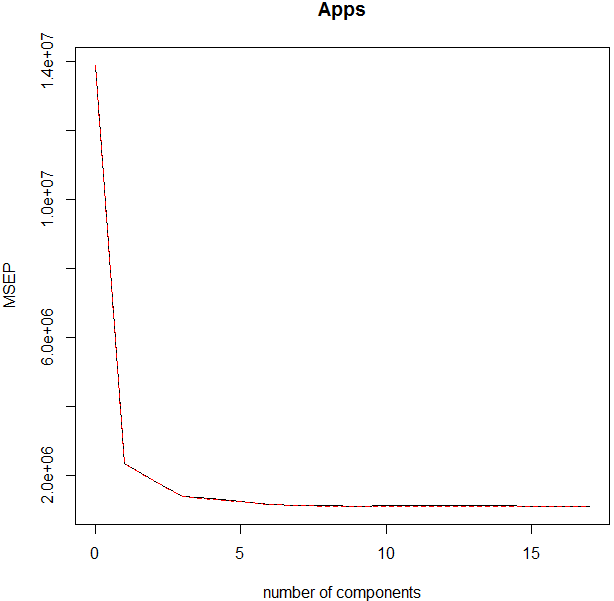
1. **Fit a lasso model on the training set, with *λ* chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.**

**  
**The test MSE is also higher for ridge regression than for least squares.  
****

1. **Fit a PCR model on the training set, with *M* chosen by cross-validation. Report the test error obtained, along with the value of *M* selected by cross-validation.**

**  
**The test MSE is also higher for PCR than for least squares.

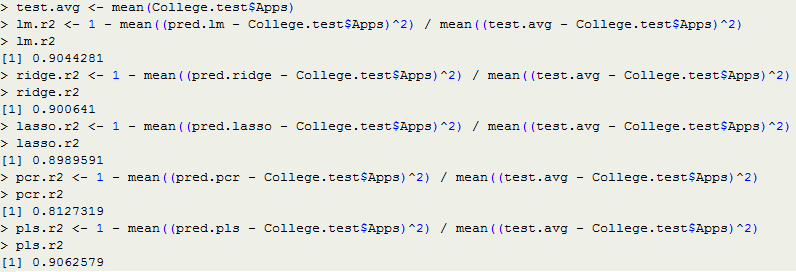
1. **Fit a PLS model on the training set, with *M* chosen by cross-validation. Report the test error obtained, along with the value of *M* selected by cross-validation.**

**  
**

****The test MSE is lower for PLS than for least squares.

1. **Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these five approaches ?**

We compute the test R^2 for all models.

****

As shown above, the test R2 for least squares is 0.9044281, the test R2 for ridge is 0.9000641, the test R2 for lasso is 0.8989591, the test R2 for pcr is 0.8127319 and the test R2 for pls is 0.9062579. All models, except PCR, predict college applications with high accuracy.