**ISL Lab-4 & 5**

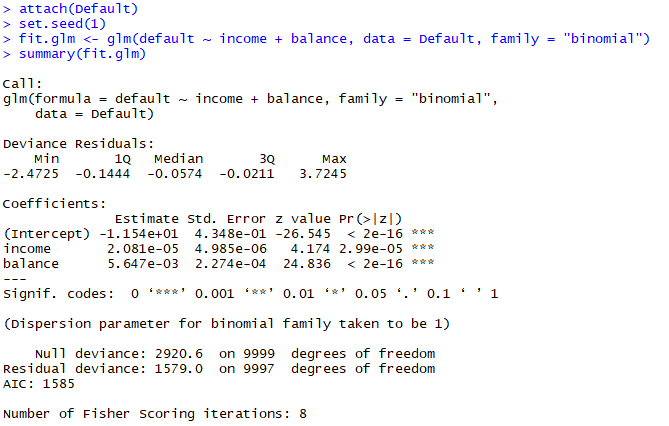
**Linga Sai Yuvesh Venketa Kotiala**

**16242113**

**2. In Chapter 4, we used logistic regression to predict the probability of default using income and balance on the Default data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.**

**(a) (10 points) Fit a logistic regression model that uses income and balance to predict default.**

Ans.



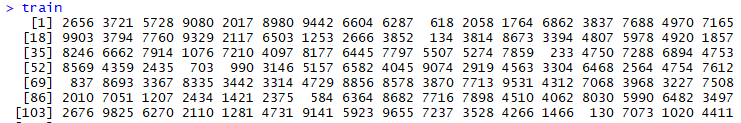
**(b) (10 points total) Using the validation set approach, estimate the test error of this model. In order**

**to do this, you must perform the following steps:**

**i. (2.5 points) Split the sample set into a training set and a validation set.**

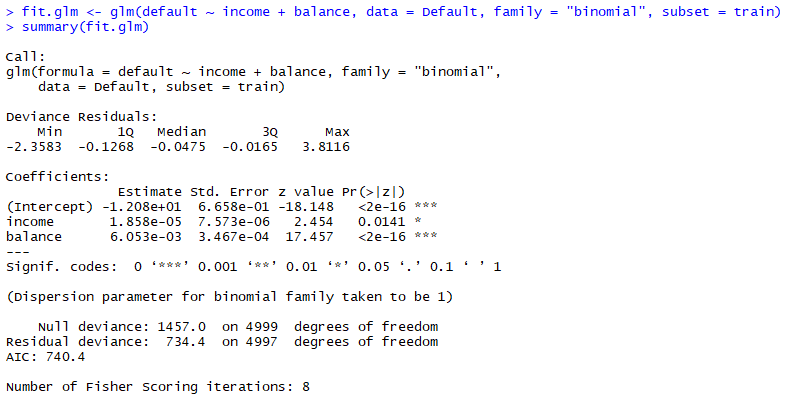
**Ans.**





**ii. (2.5 points) Fit a multiple logistic regression model using only the training observations.**

**Ans.**

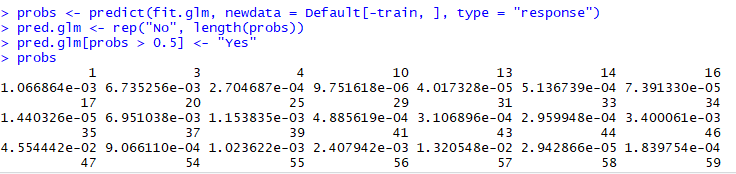


**iii. (2.5 points) Obtain a prediction of default status for each individual in the validation set**

**by computing the posterior probability of default for that individual, and classifying the**

**individual to the default category if the posterior probability is greater than 0.5.**

**Ans.**



**iv. (2.5 points) Compute the validation set error, which is the fraction of the observations in the**

**validation set that are misclassified.**

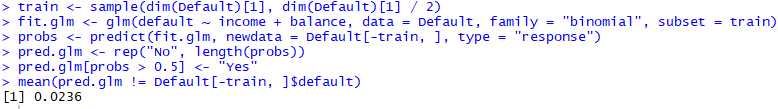
**Ans.**

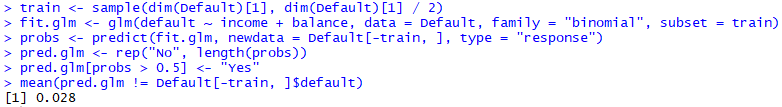


We have a 2.86% test error rate with the validation set approach.

**(c) Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.**

**Ans.**

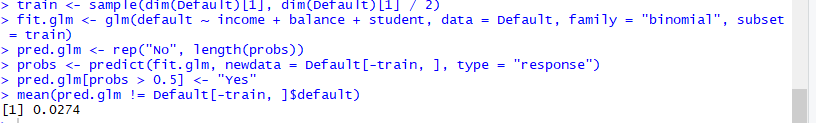




We see that the validation estimate of the test error rate can be variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

**(d) Now consider a logistic regression model that predicts the probability of “default” using “income”, “balance”, and a dummy variable for “student”. Estimate the test error for this model using the validation set approach. Comment on whether including a dummy variable for “student” leads to a reduction in the test error rate.**

Ans.



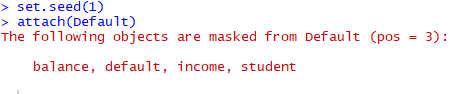
**3. (40 points) We continue to consider the use of a logistic regression model to predict the probability of default using income and balance on the Default data set. In particular, we will now compute estimates for the standard errors of the income and balance logistic regression coefficients in two different ways: (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed before beginning your analysis.**

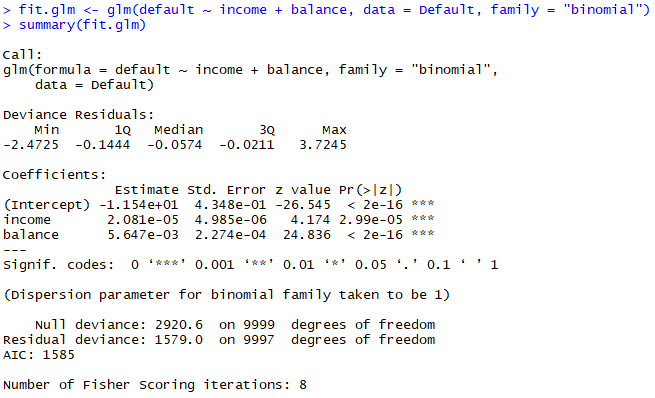
**(a) (10 points) Using the summary () and glm() functions, determine the estimated standard errors**

**for the coefficients associated with income and balance in a multiple logistic regression model**

**that uses both predictors.**

**Ans.**





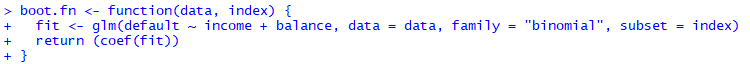
The glm() estimates of the standard errors for the coefficients β0, β1 and β2 are respectively 0.4347564, 4.985167210^{-6} and 2.273731410^{-4}.

**(b) (10 points) Write a function, boot.fn(), that takes as input the Default data set as well as an**

**index of the observations, and that outputs the coefficient estimates for income and balance in**

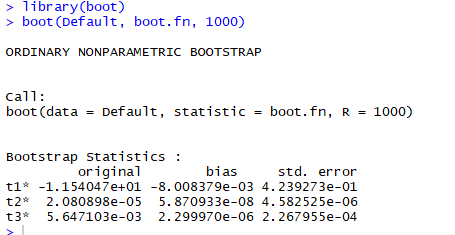
**the multiple logistic regression model.**

**Ans.**



**(c) Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for “income” and “balance”.**

**Ans.**



**(d) Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.**

**Ans.**

The estimated standard errors obtained by the two methods are close.

**5. (30 points total) In this exercise, we will generate simulated data, and will then use this data to perform**

**best subset selection.**

**(a) (5 points) Use the rnorm() function to generate a predictor X of length n = 100, as well as a**

**noise vector ϵ of length n = 100**

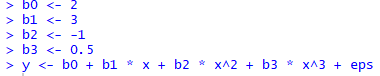


**(b) (5 points) Generate a response vector Y of length n = 100 according to the model**

**Y = β0 + β1X + β2X2 + β3X3 + ϵ,**

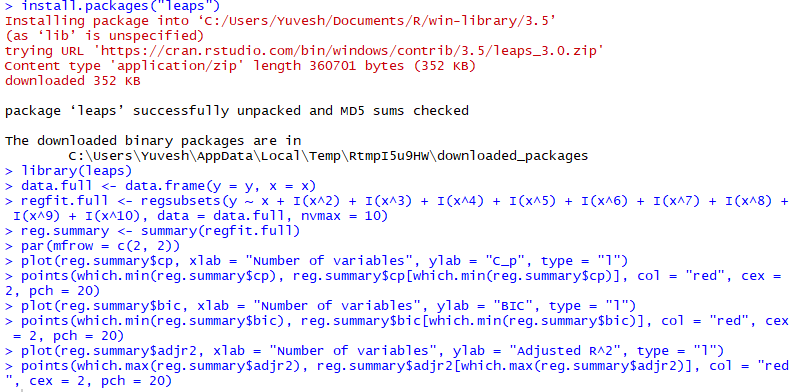
**where β0, β1, β2, and β3 are constants of your choice.**

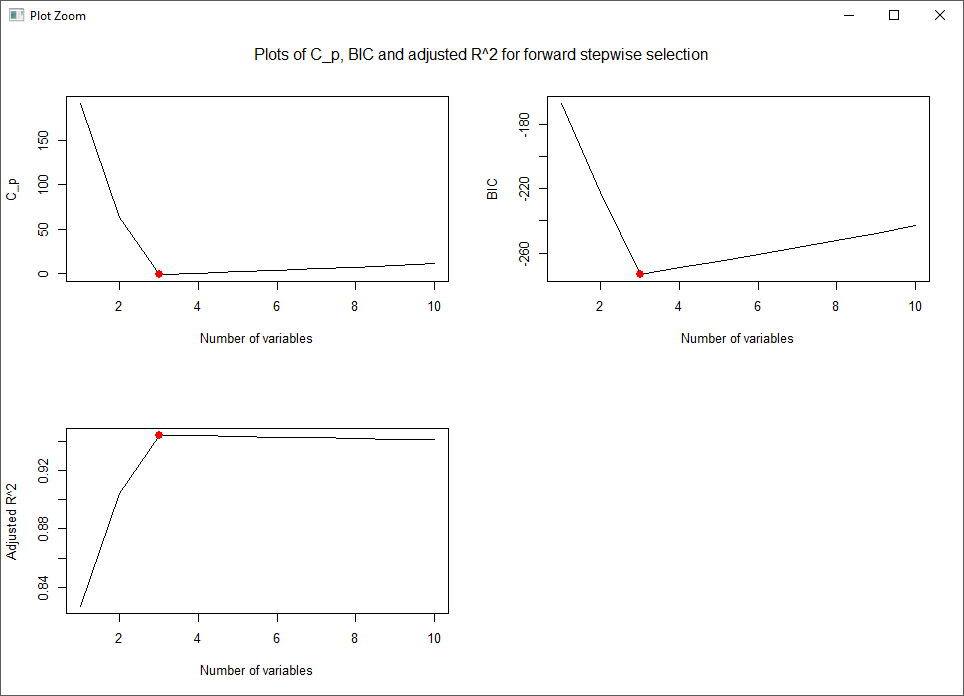
**Ans.**



**(c) (5 points) 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 Cp, 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.**





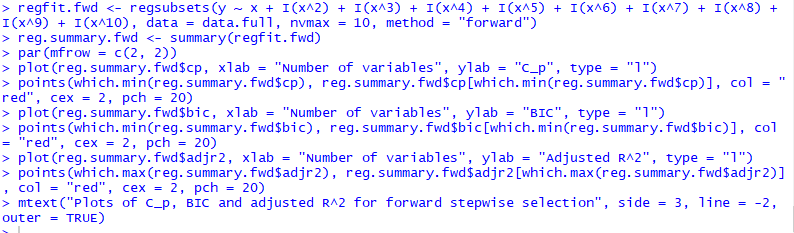
We find that, with Cp we pick the 3-variables model, with BIC we pick the 3-variables model, and with adjusted R2 we pick the 3-variables model**.**

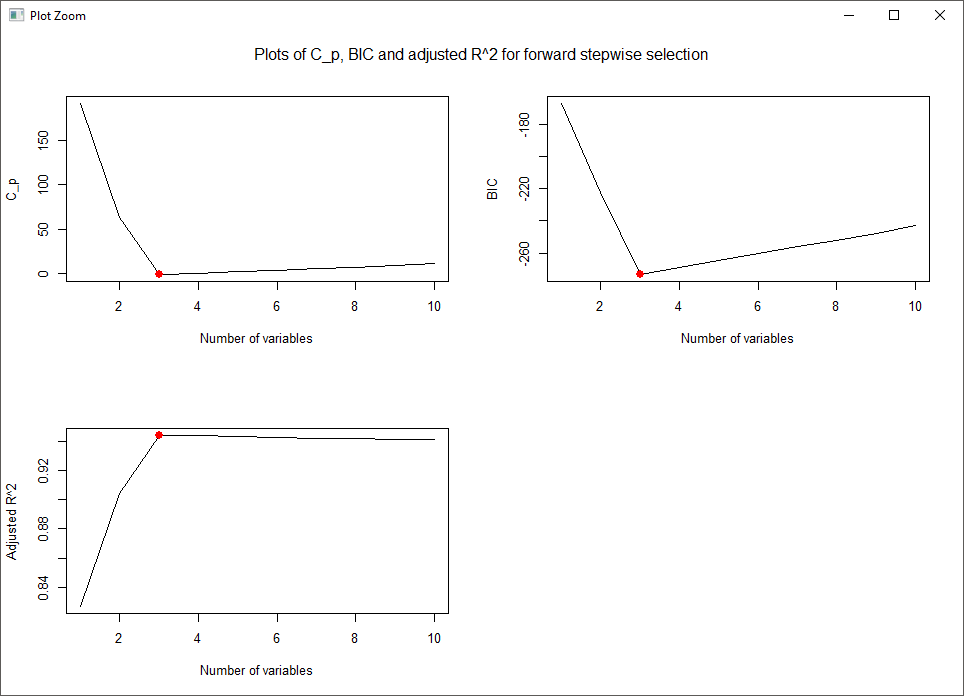


**(d) Repeat (c), using forward stepwise selection and also using backwards stepwise selection. How does your answer compare to the results in (c) ?**

**Ans.**

We begin with forward stepwise selection.

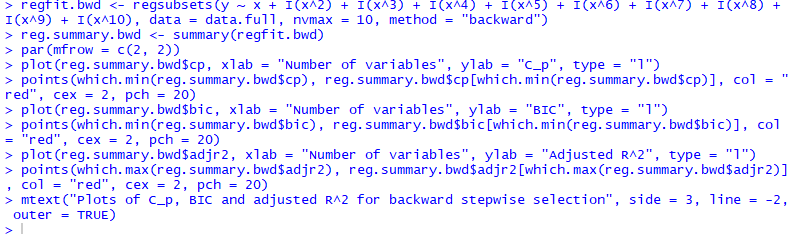


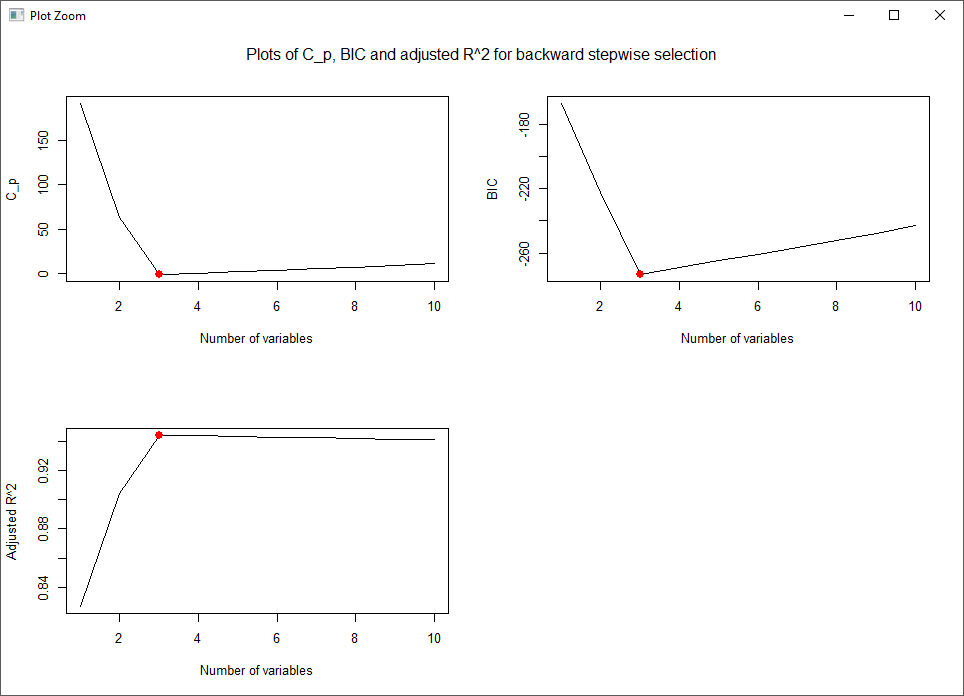


We find that, for forward stepwise selection, with Cp we pick the 3-variables model, with BIC we pick the 3-variables model, and with adjusted R2 we pick the 3-variables model.



Next, we proceed with backward stepwise selection.



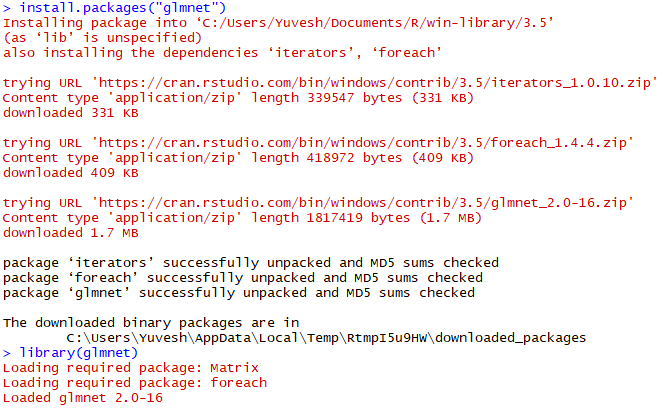


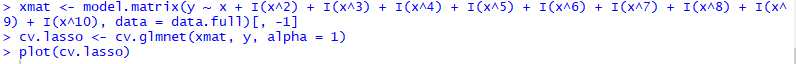
We find that, for backward stepwise selection, with Cp we pick the 3-variables model, with BIC we pick the 3-variables model, and with adjusted R2 we pick the 3-variables model.

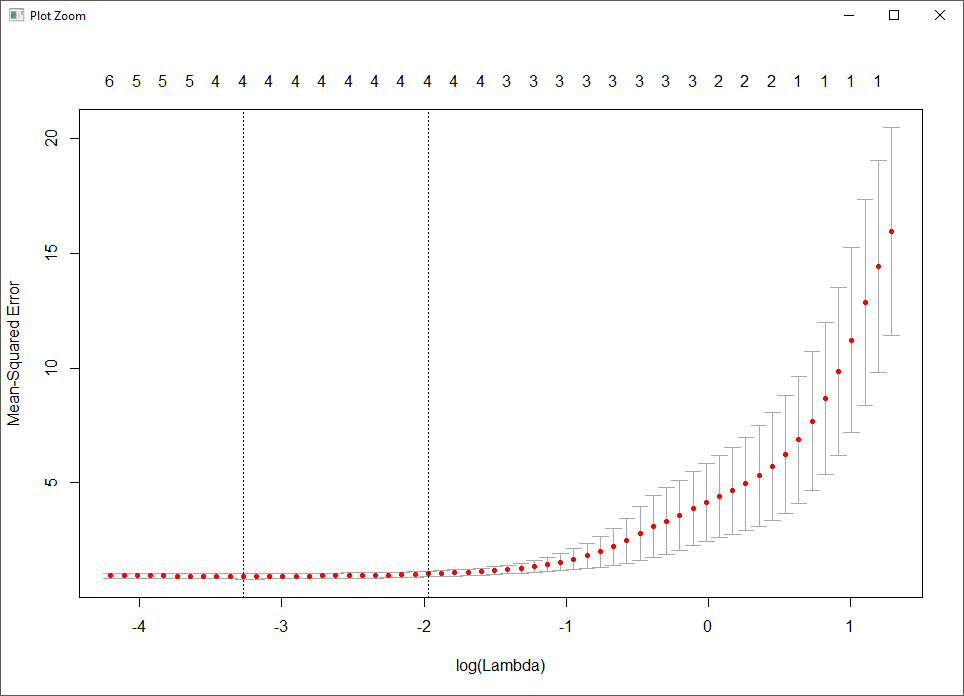


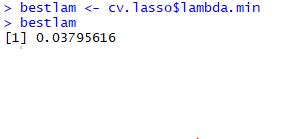
Here forward stepwise, backward stepwise and best subset all select the three variables model with X, X2 and X5.

**(e) Now fit a lasso model to the simulated data, again using X,X2,⋯,X10 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.**

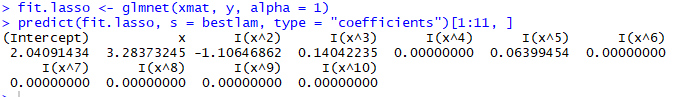








Now we refit our lasso model using the value λ= 0.0379562 chosen by cross-validation.



The lasso method picks X, X2, X3 and X5 as variables for the model.

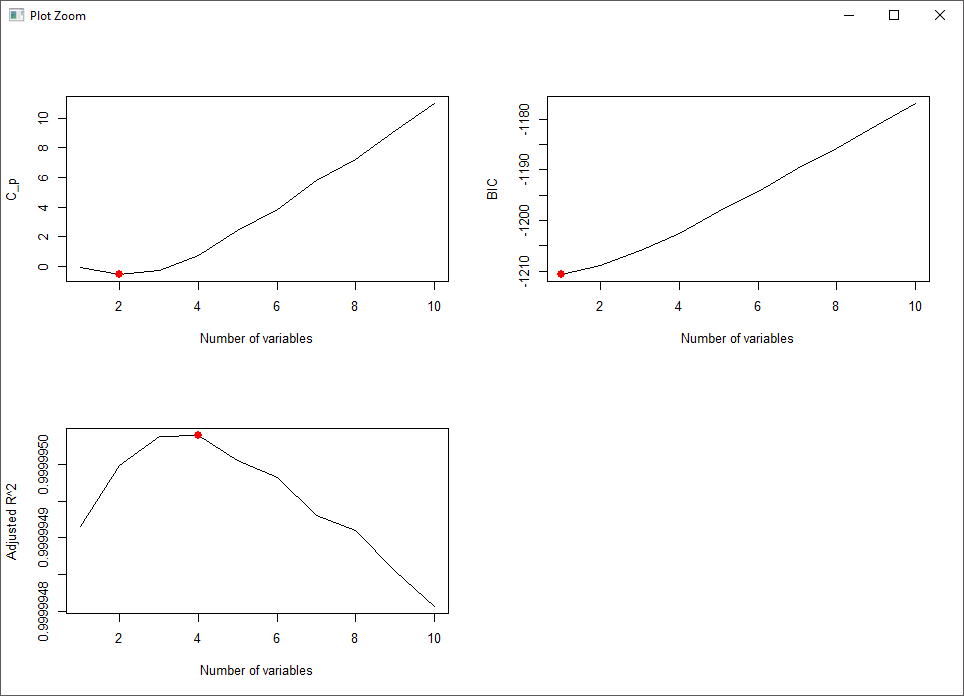
**(f) Now generate a response vector Y according to the model**

**Y=β0+β7X7+ε,**

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

**Ans.**

We begin with best subset selection.



We find that, with Cp we pick the 2-variables model, with BIC we pick the 1-variables model, and with adjusted R2 we pick the 4-variables model.

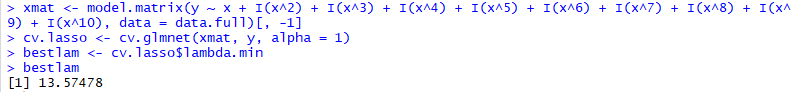


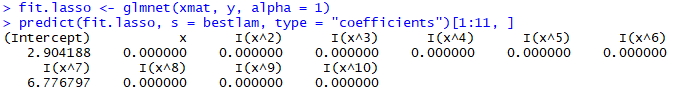




Here best subset selection with BIC picks the most accurate 1-variable model with matching coefficients.

Now we proceed with the lasso.



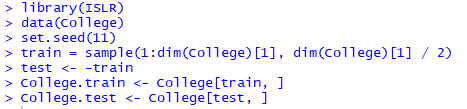


Here the lasso also picks the most accurate 1-variable model, but the intercept is quite off.

**6. (35 points total) In this exercise, we will predict the number of applications received using the other variables in the College data set.**

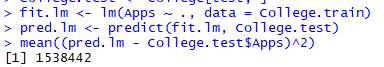
**(a) (5 points) Split the data set into a training set and a test set.**

**Ans.**



**(b) (5 points) Fit a linear model using least squares on the training set, and report the test error Obtained.**

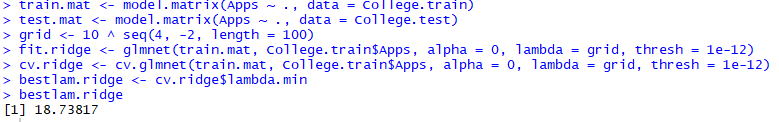
**Ans.**



The test MSE is 1.538442210^{6}.

**(c) (5 points) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained.**

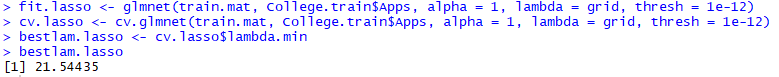
**Ans.**





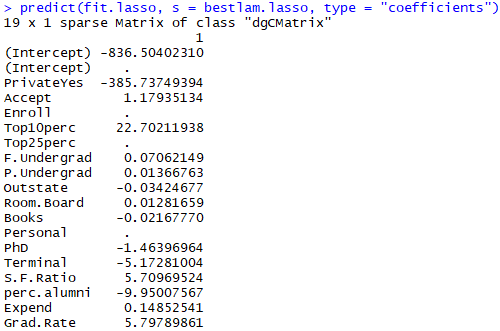
The test MSE is higher for ridge regression than for least squares.

**(d) (5 points) Fit a lasso model on the training set, with λ chosen by crossvalidation. Report the test error obtained, along with the number f non-zero coefficient estimates.Ans.**



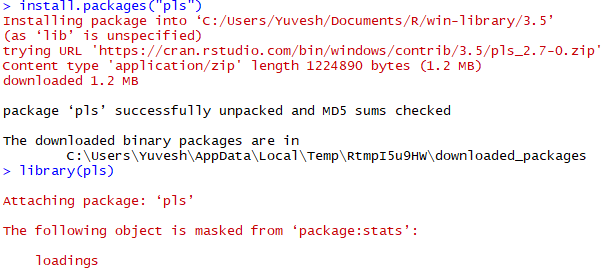


The test MSE is also higher for ridge regression than for least squares.

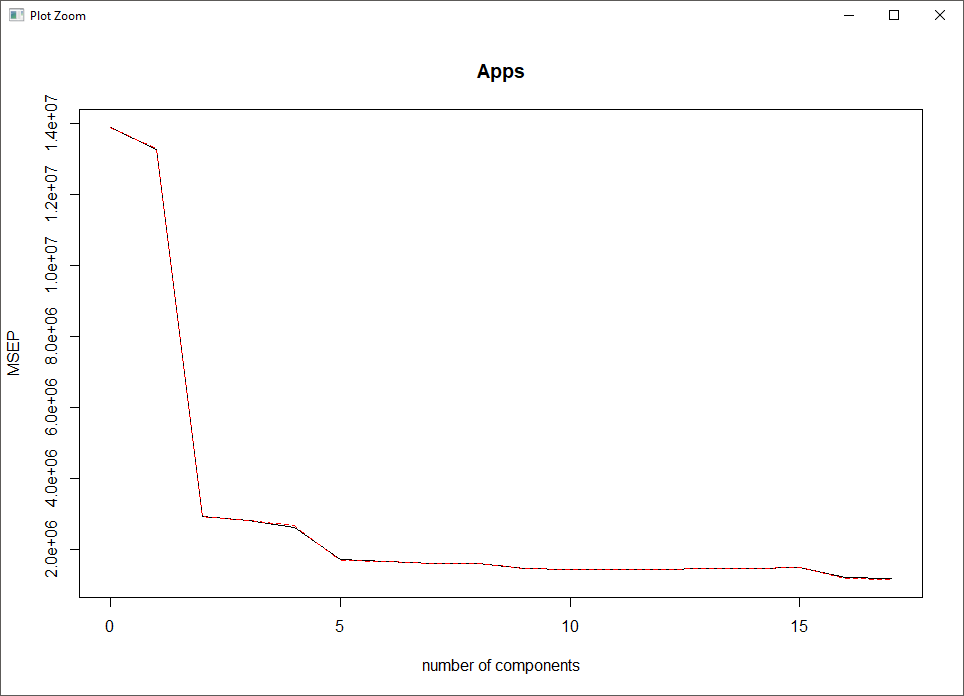


**(e) (5 points) Fit a PCR model on the training set, with M chosen by crossvalidation. Report the test error obtained, along with the value of M selected by cross-validation.**

**Ans.**







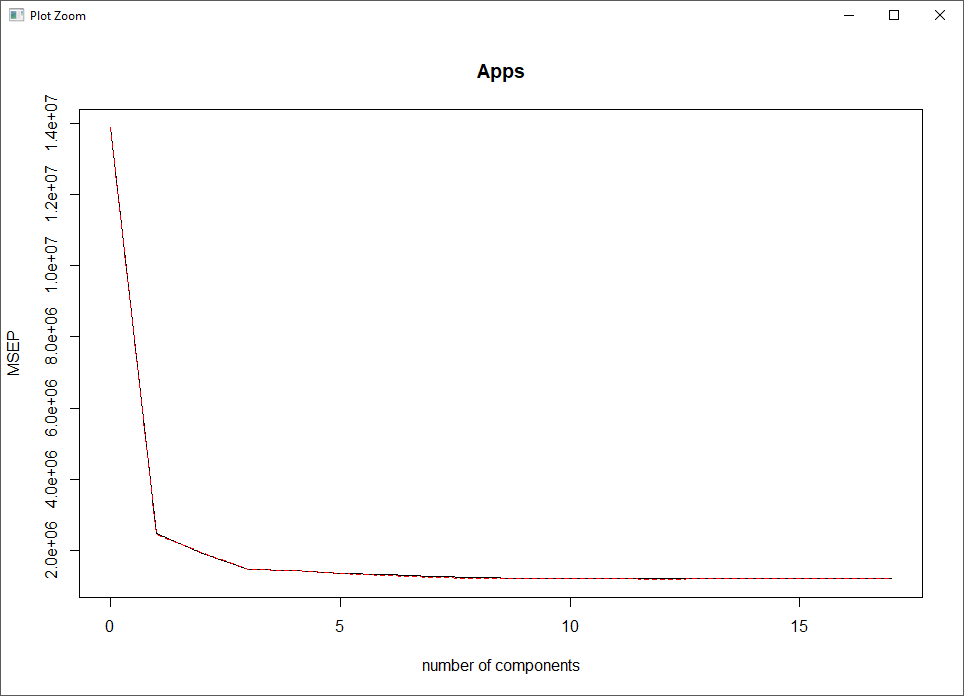


The test MSE is also higher for PCR than for least squares.

**(f) (5 points) 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**

**Ans.**



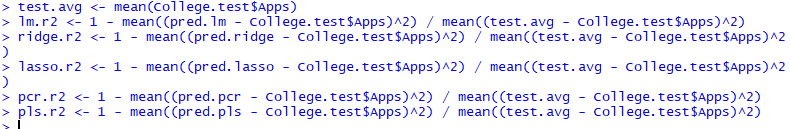




Here, the test MSE is lower for PLS than for least squares.

**(g) (5 points) 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?Ans.**

To compare the results obtained above, we have to compute the test R2 for all models.



So the test R2 for least squares is 0.9044281, the test R2 for ridge is 0.9000536, the test R2 for lasso is 0.8984123, 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.