Homework Chapter 6

Jacob Thielemier

7 March 2024

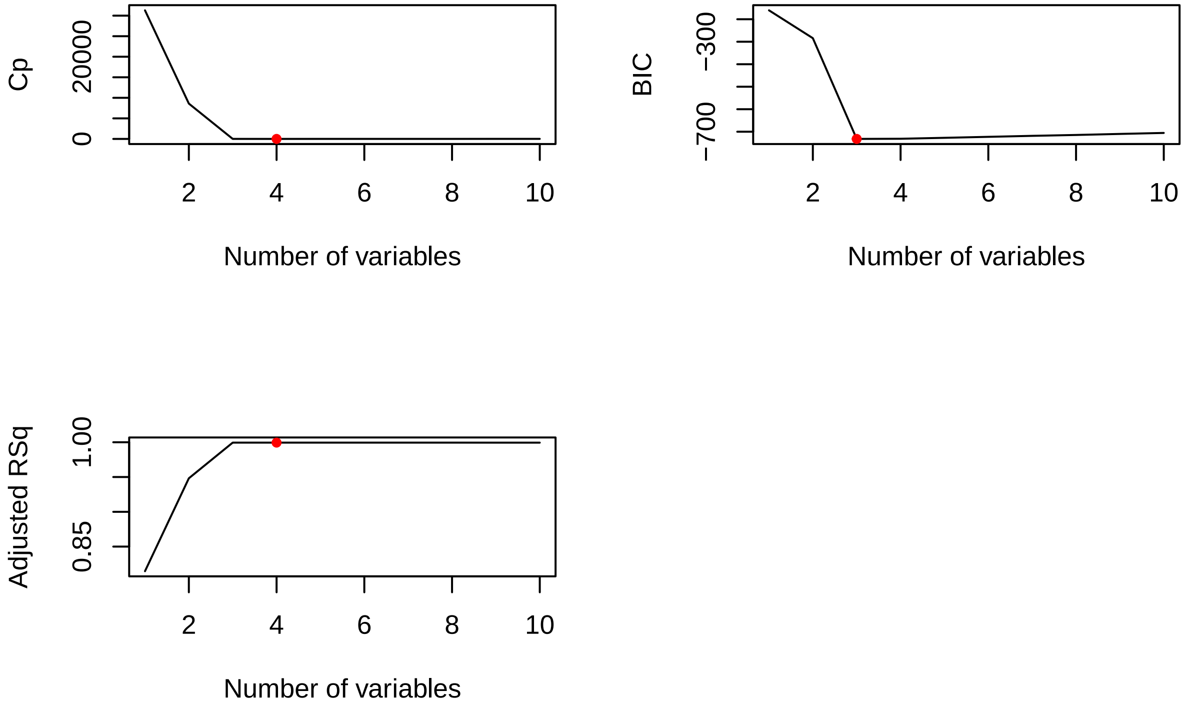
Question 1

- **Logistic Regression:** This is used for binary or binomial outcome data, where the response variable has two possible values (e.g., success/failure, 1/0). Logistic regression models the probability of the default category (often coded as 1).
- **Poisson Regression:** Suitable for count data or rates, Poisson regression is used when the response variable represents the number of times an event occurs in a fixed interval of time or space.
- **Multinomial Regression:** This is an extension of logistic regression to multiclass problems, where the response variable can take on more than two categories. It's useful for classifying subjects into multiple categories based on the features provided.
- **Cox Proportional Hazards Regression:** Used for survival analysis, this model helps in analyzing the time until the occurrence of an event of interest (e.g., death, failure), taking into account the effect of covariates on the time.
- **Multiple-response Gaussian:** This can be used when there are multiple continuous response variables that you want to predict from the same set of predictors. It fits a separate linear regression model for each response variable but does so simultaneously.
- **Multivariate Binomial with grouped data:** This is a special case for logistic regression where the response variable can have more than two outcomes, and these outcomes can be grouped in some way. It's useful when dealing with grouped or correlated binary outcomes.

Question 2

Part (a, b, and c)

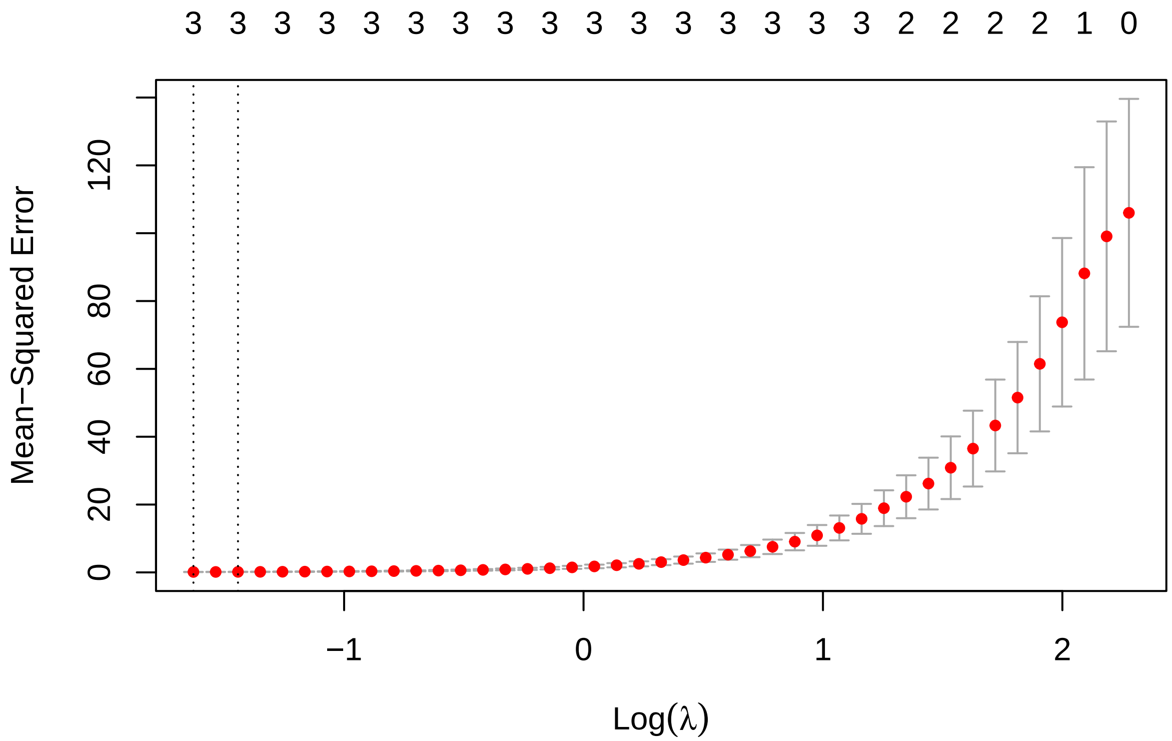
- C_p reduces substantially from the one and two variable model to the three variable model. It reduces slightly in the four variable model and rises in small increments after. BIC value is lowest for the three variable model. Adjusted R^2 increases to 0.999 in the three variable model from the two variable model value of 0.95.
- These metrics point to the three variable model as being the best choice. We can confirm this visually in the charts below.



Part (d)

- Using forward stepwise the statistical metrics are very similar to that for best subset selection.
- Using backwards stepwise the results are very similar to best subset and forward selection.
- Compared to 8(c) we see that all these metrics show that the three variable model with the squared and cubed term is the best.

Part (e)



- Higher values of λ result in an increase in the MSE. The best value of λ is 0.2.
- The lasso model creates a sparse model with four variables. The intercept and coefficients for X, X^2 and X^3 closely match the ones chosen in 8(b) while the value for x^4 is very small. This model provides an accurate estimation of the response Y.

Part (f)

- C_p is lowest for one variable model with the X^7 term. BIC value is lowest for the one variable model. Adjusted R^2 is 0.999+ for the one variable model.
- These statistical metrics point to the one variable model with the X^7 term as being the best choice.
- Lasso model using best value of λ results in a sparse model with one variable. It assigns a non-zero coefficient to the variable X^7 that explains the response Y, and assigns a zero to the rest.

Question 3

Part (a and b)

- The MSE from ridge regression with lambda=0 and lm() function are reasonably similar, and the slight discrepancy is likely due to approximation used by glmnet().

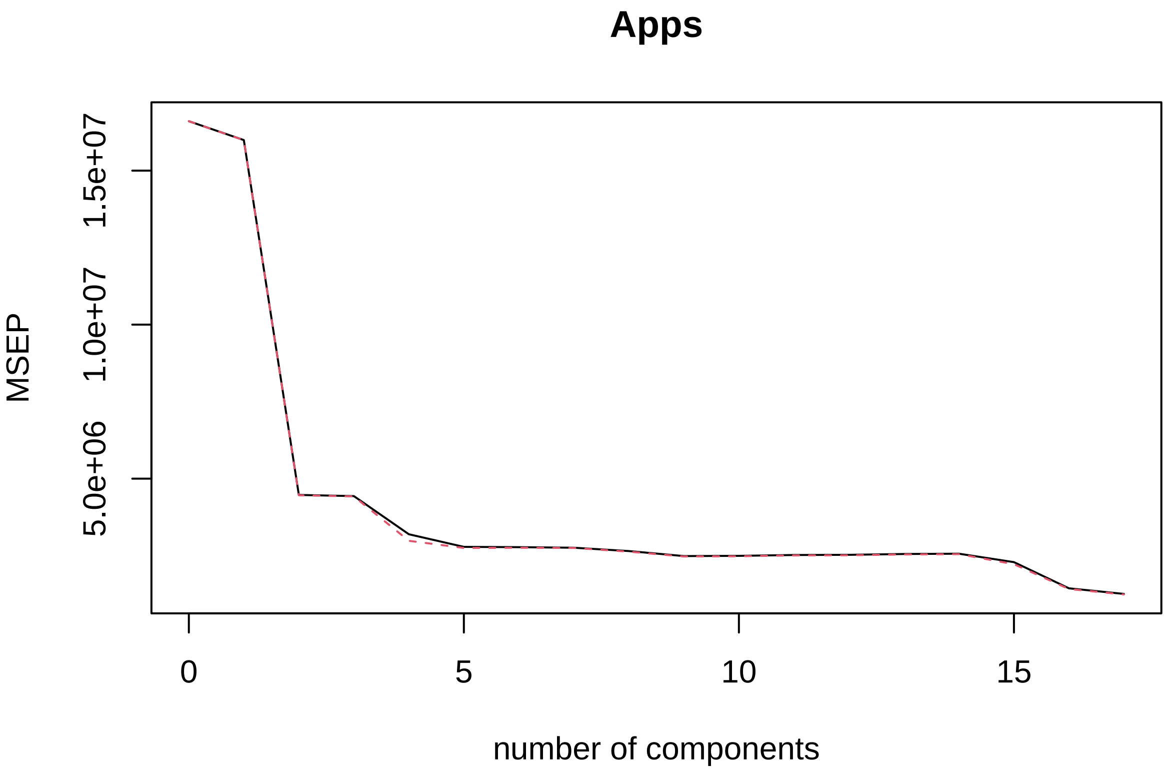
Part (c)

- Best value of λ is 387.9 and the test MSE is slightly lower than for least squares.

Part (d)

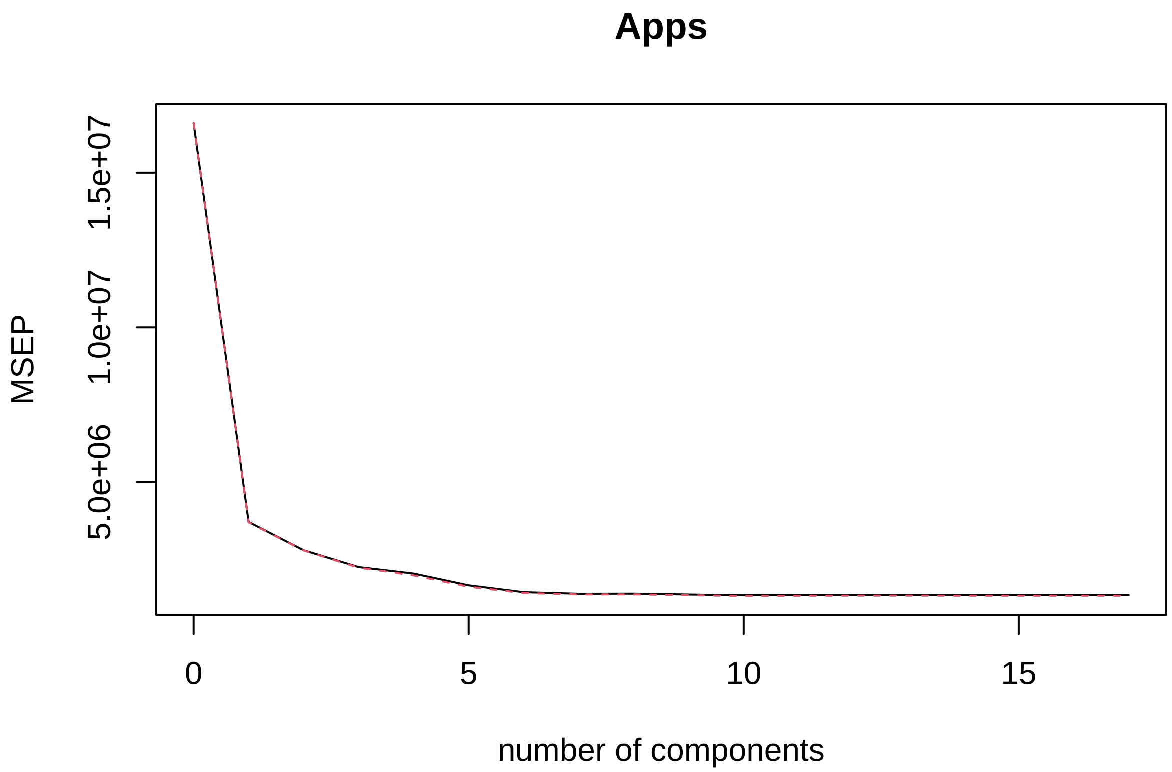
- Test MSE is slightly lower than least squares regression.
- There are no non-zero variables, but many variables are heavily shrunk.

Part (e)



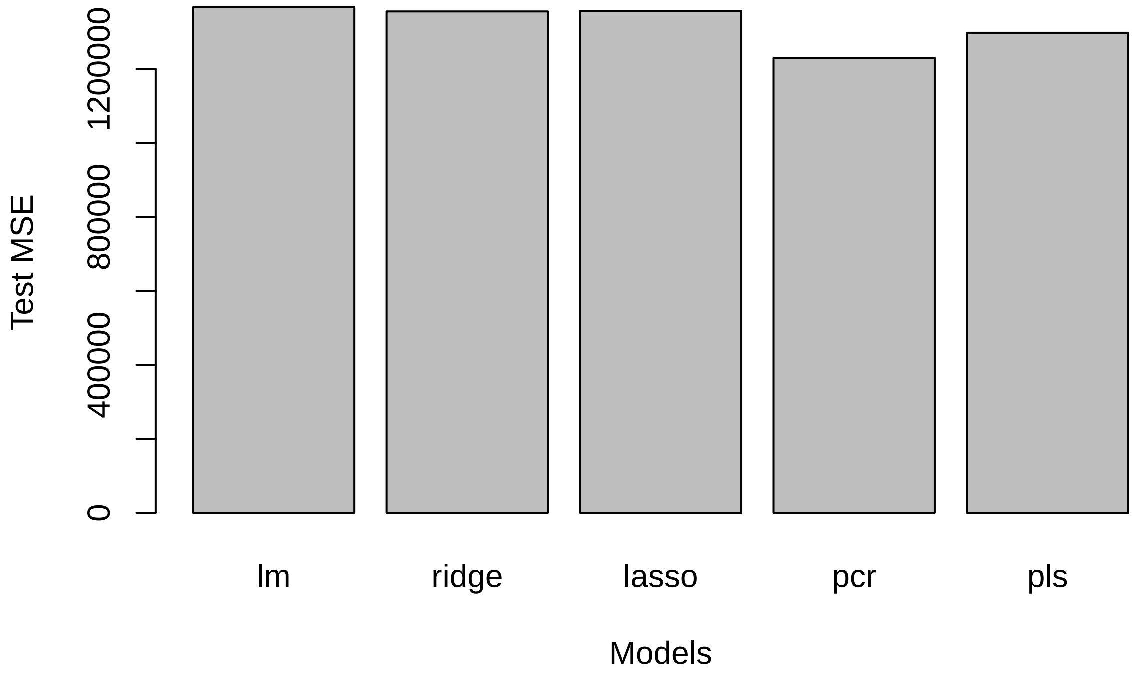
- From the graph we can observe that the MSE reduces rapidly as M increases and is lowest at M=16. However, the reduction from M=5 to M=16 is small when compared the reduction from M=1 to M=5.
- Test MSE for M=5 is 1945054, which is much larger than least squares. The best MSE is achieved when M=16, which gives a test MSE reasonably lower than least squared. Using M=17 gives the same result as least squares.

Part (f)

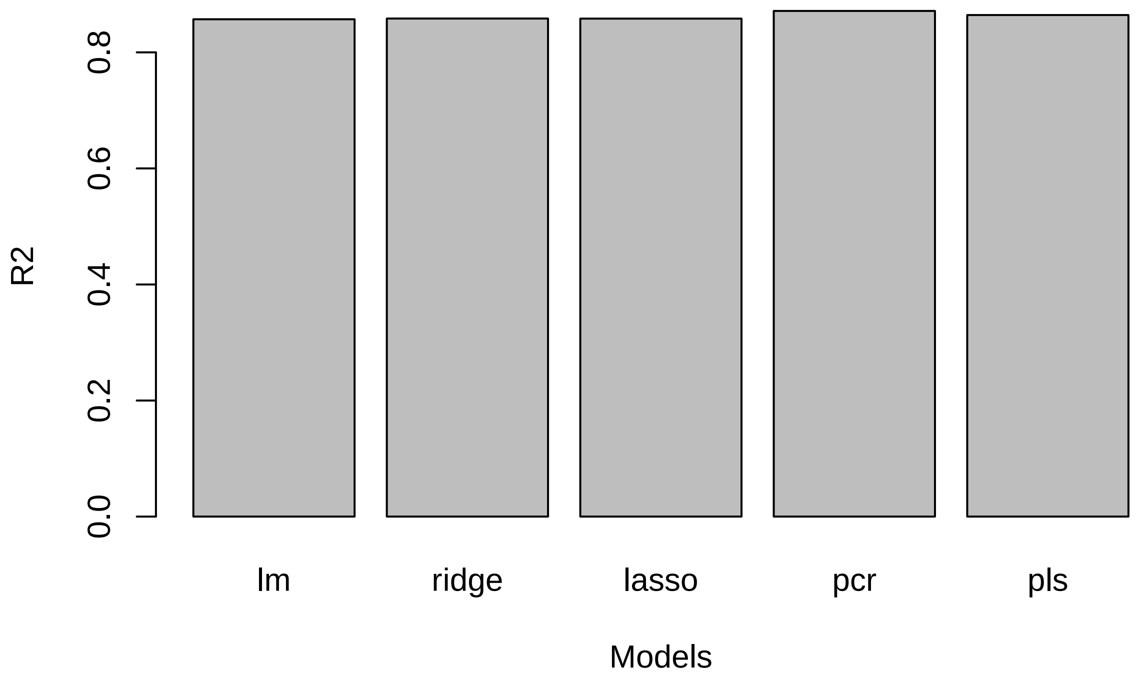


- The lowest MSE occurs when m~8.
- The test MSE is similar to PCR and slightly lower than least squares.

Part (g)



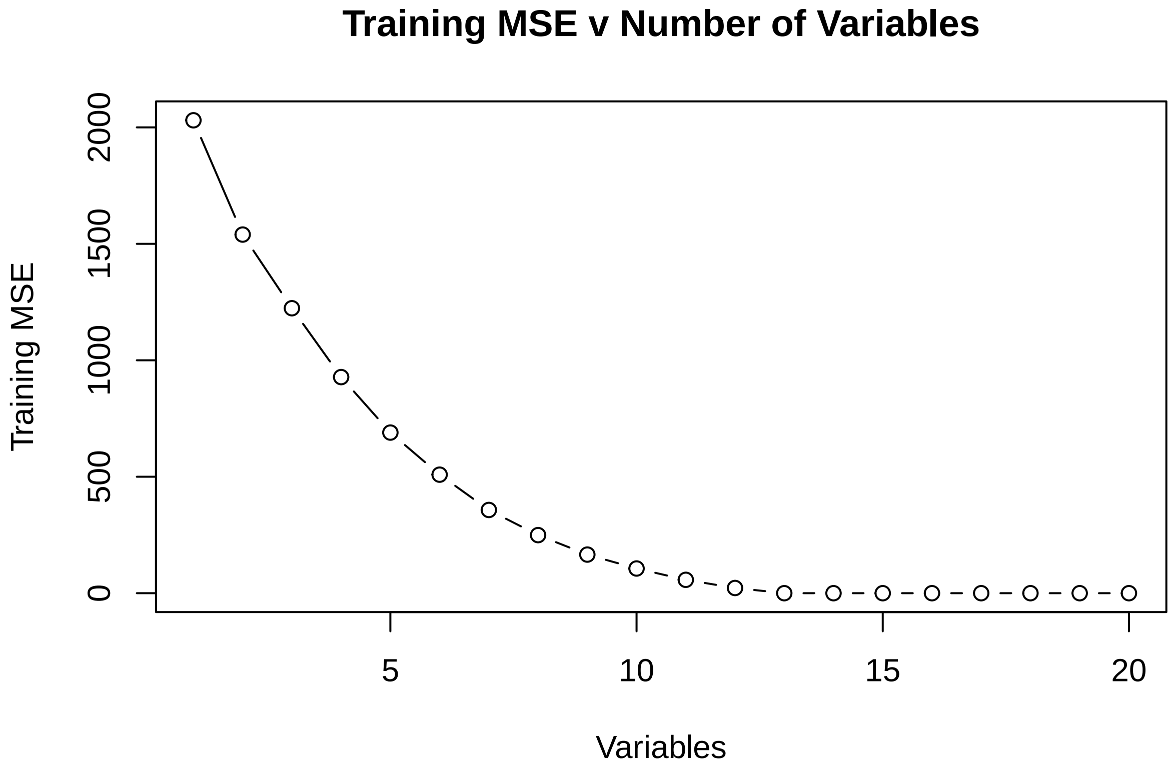
- All the models give reasonably similar results, with PCR and PLS giving slightly lower test MSE's.



- Every model has a R2 metric of around 0.85 or above so we can be reasonably confident about the accuracy of the predictions.

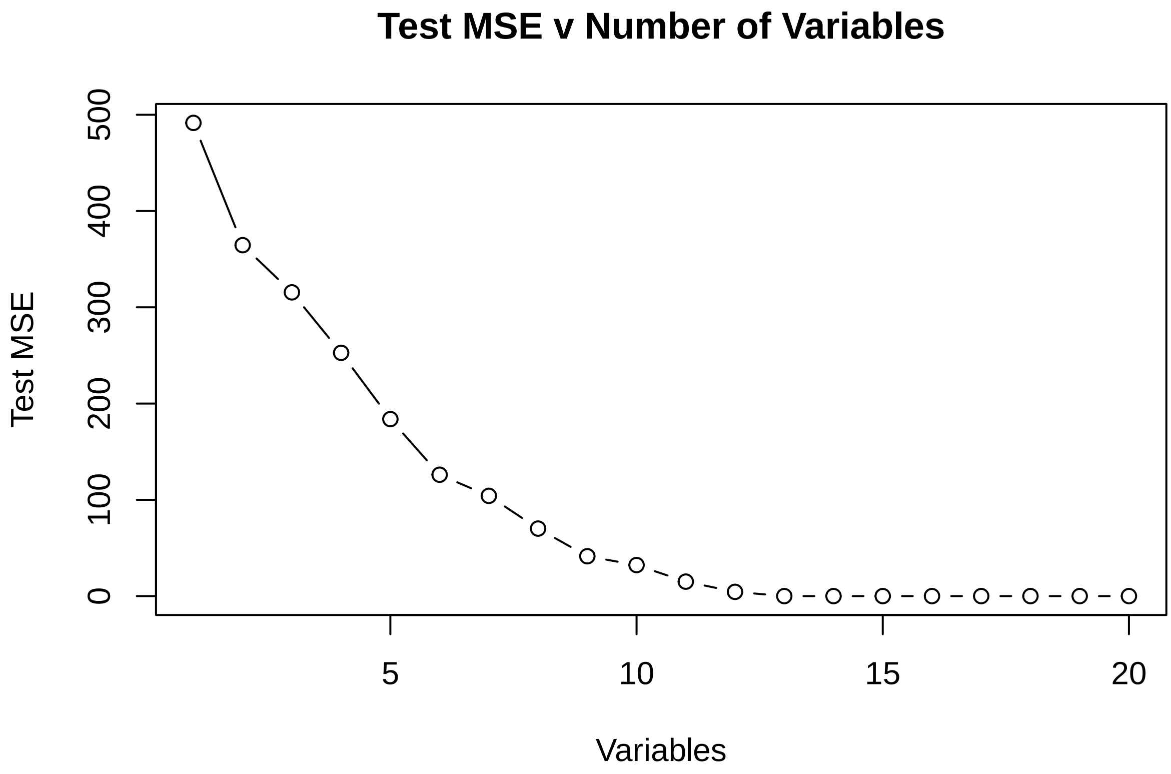
Question 4

Part (a, b, and c)



- Training MSE decreases monotonically as the number of variables increase. Minimum training MSE is at maximum number of variables: 20.

Part (d)



- Test MSE decreases rapidly as the number of variables increase, but the minimum is not at the max number of variables. Minimum test MSE is when number of variables: 13.

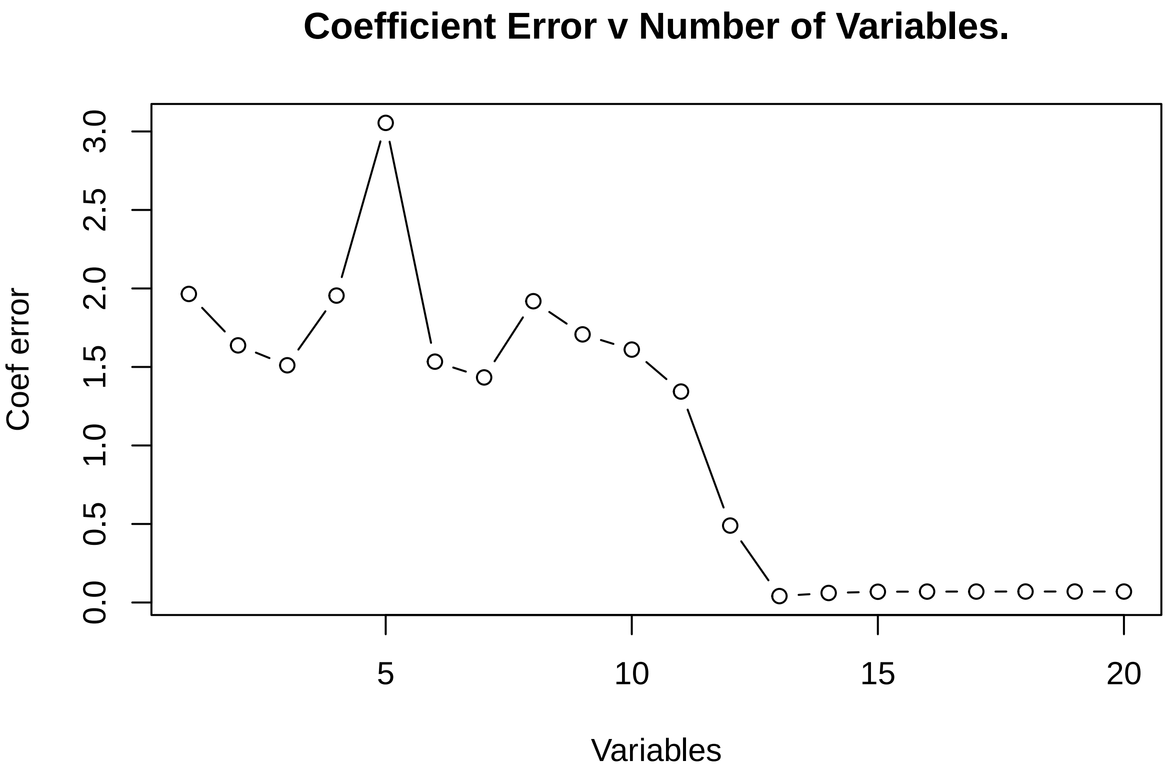
Part (e)

- Minimum Test MSE occurs at a model with 13 variables. The test MSE deceases until it reaches the minimum and then starts to rise afterwards.
- As the model flexibility increases, it is better able to fit the data set. This results in the Test MSE decreasing rapidly until it reaches a minimum. Further increases in model flexibility causes over fitting and this results in an increase in the Test MSE.

Part (f)

- The best model variables match the 13 non-zero variables from the original model, and their respective coefficients are very similar.

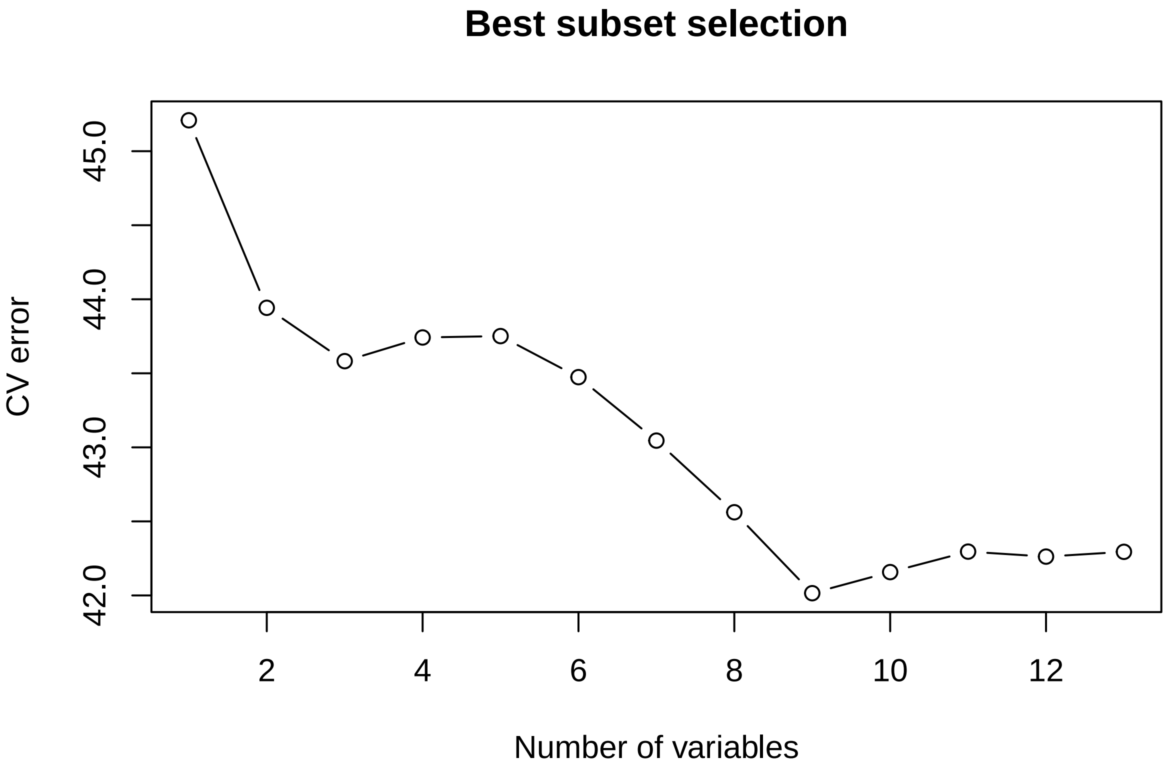
Part (g)



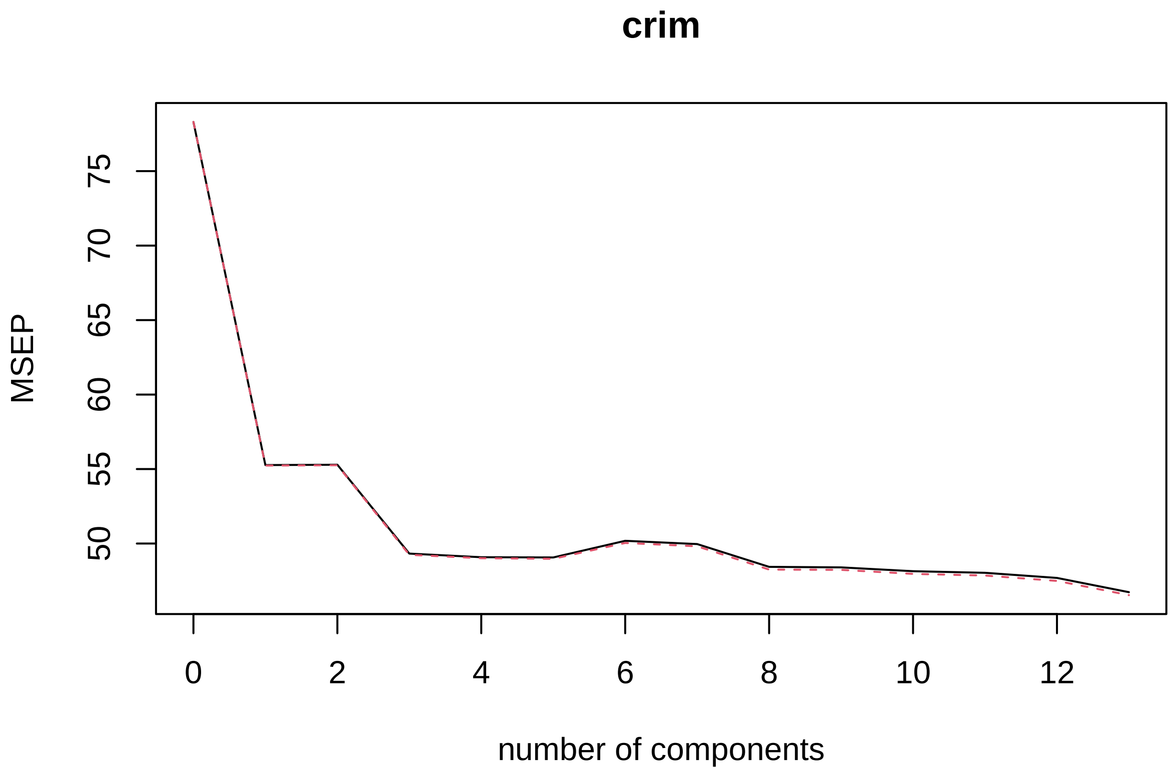
- The graph starts in a disjointed manner before the coefficient errors start reducing rapidly. Eventually, it does show a minimum at the same variable size as for the test MSE. Though, when using a different random seed the coefficient error chart does not always find a minimum at the same variable size as the test MSE chart. A model that gives a minimum for coefficient error does not always lead to a lower test MSE.

Question 4

Part (a)



- CV error is lowest for model with 9 variables. CV Error = 42.0151126.
- Test MSE of 31.6, with only age and tax being exactly zero, we have a best model with 10 variables.
- Lambda chosen by cross validation is close to zero, so both ridge regression and lasso test mse are similar to that provided by least squares.



- Using PCR gets a Test MSE of 33.7.

Part (b and C)

- I would choose the Lasso model, as it gives the lowest Test MSE.
- Lasso models are generally more interpretable for users and presentations.
- It results in a sparse model with 10 variables. Two variables whose effect on the response were below the required threshold were removed.