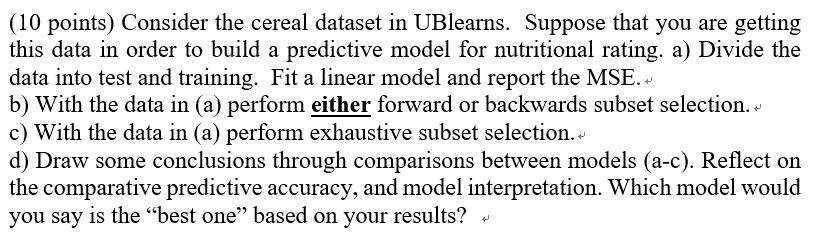
Question 1.



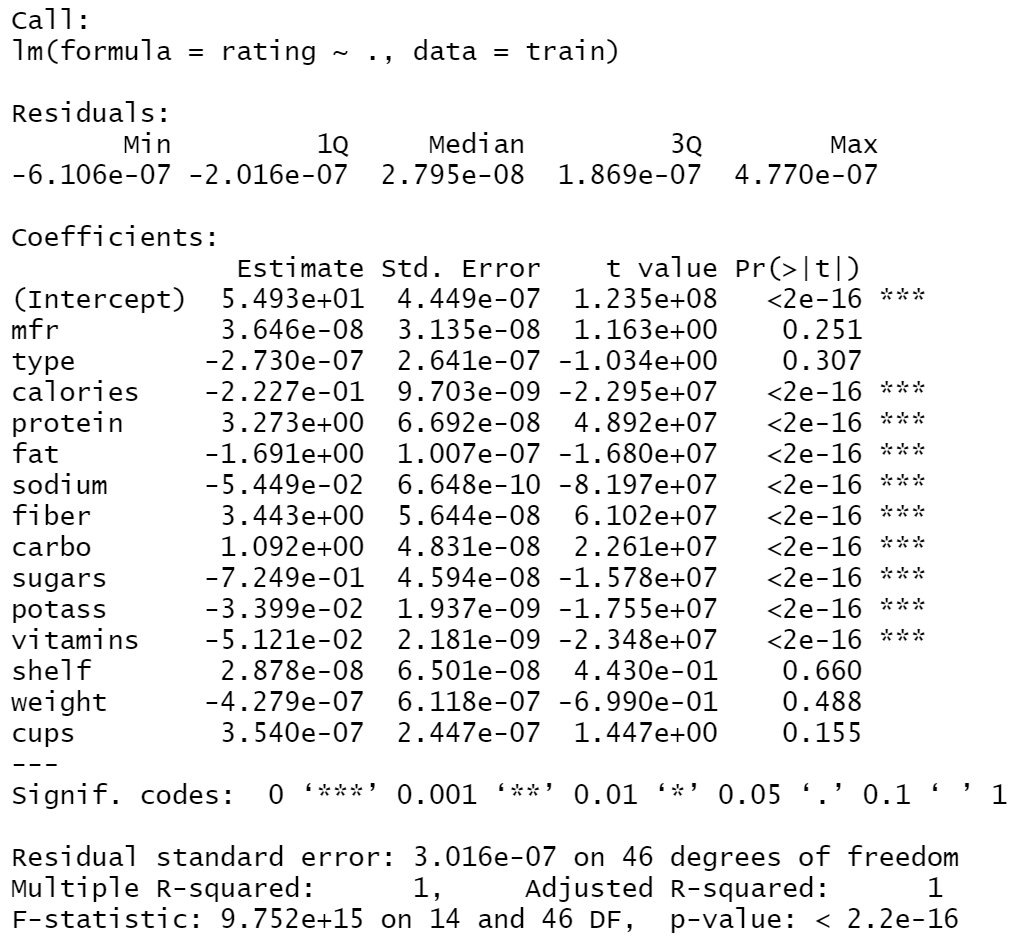
a)

Process:

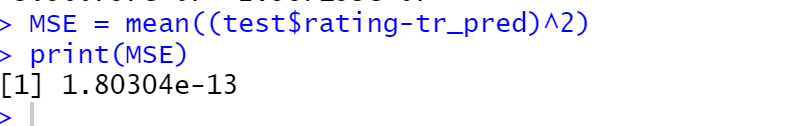
1. **Read data**: I read CSV data ‘cereal.csv’ with function ‘read.csv’
2. **Remove categorical data**: I removed categorical data because it is not useful when I fit this data with the Linear model.
3. **Make Train and Test data**: In my opinion, since there are only a few data in ‘cereal’, 80% of the data have become Train data.
4. **Fitting the Linear model with the data**: I used the ’lm’ function to make the Linear model.
5. **Calculate the MSE of the Linear model**: I put test data into the model and calculated the predicted data. With this data, I calculate the MSE of this model. MSE-> Mean of (Rating in the test – predicted Rating)^2

Outputs:

1. Summary of the linear model.



1. Mean Square Error: 1.80304e-13



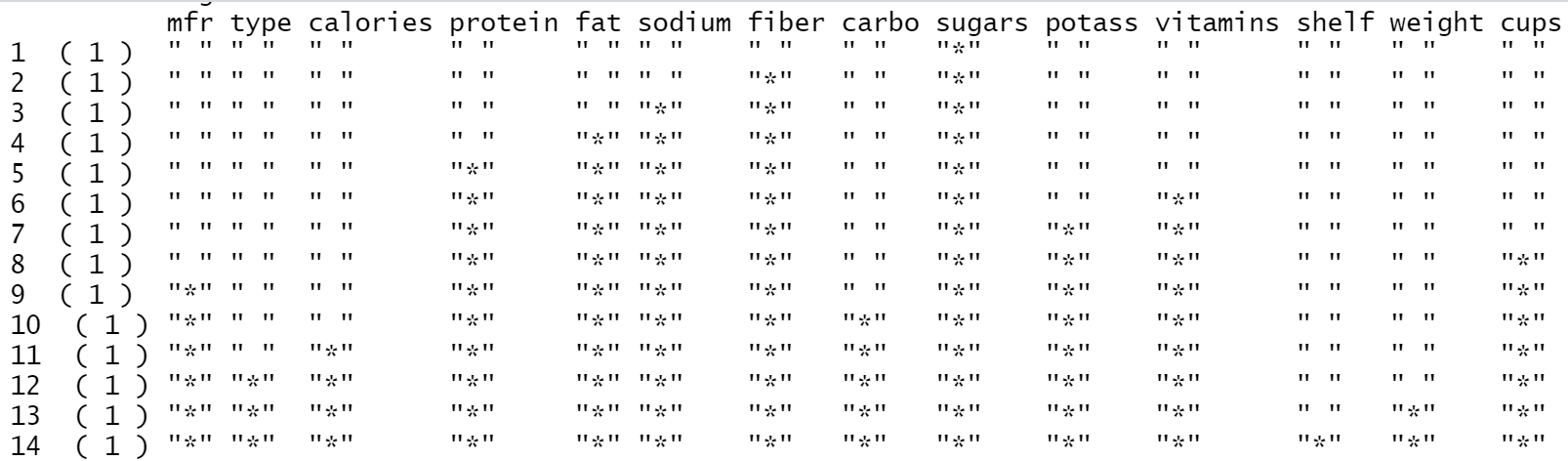
b)

Process:

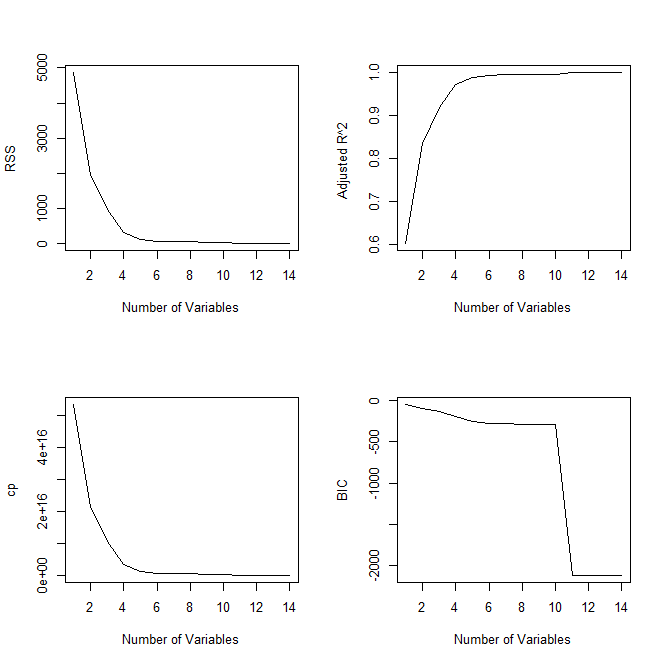
1. **Perform the forward subset section:** I used the function ‘regsubsets’ to distinguish important factors with the ‘forward’ method.
2. **Show the importance of factors in the model:** With function ‘summary’, I could see the importance factor in the model.
3. **Plot the relationship between the number of variables and ‘RSS’,’CP’,’BIC’ and ‘adjusted R^2’** I can guess what combination factors can mad the model accurately with these 4 plots and the summary of the ‘Forward Selection’.

Outputs:

1. Summary of the ‘Forward Selection’



1. Graphs for number of variables and RSS, adjusted R^2, cp and BIC



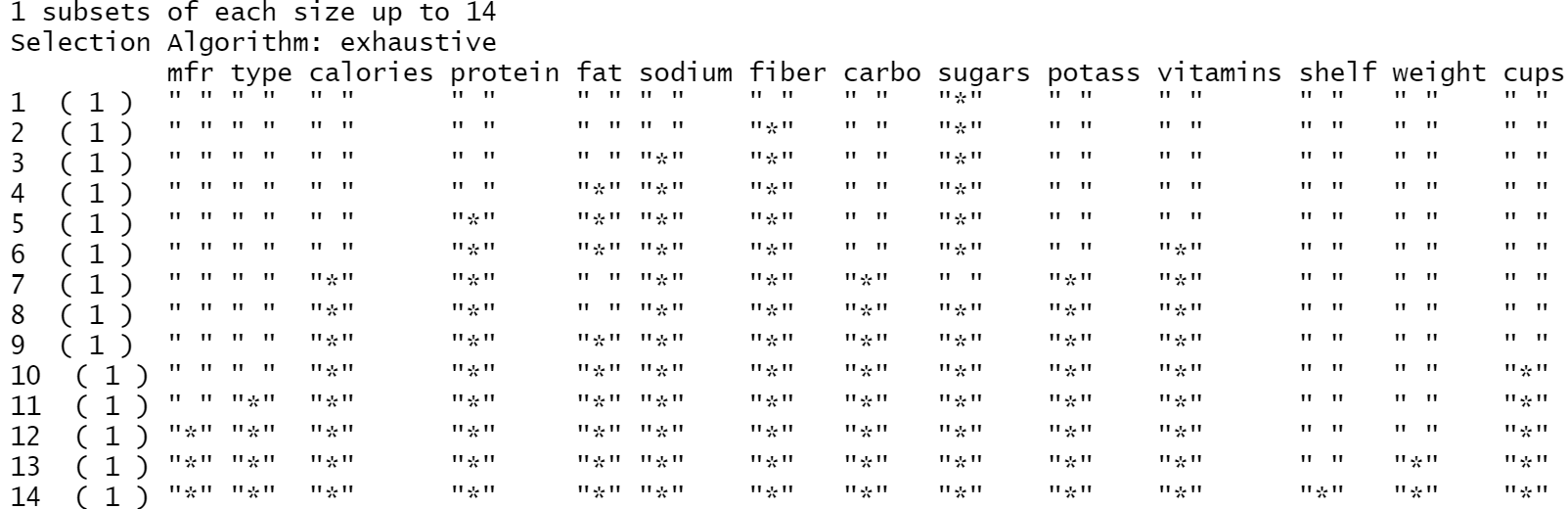
c)

Process:

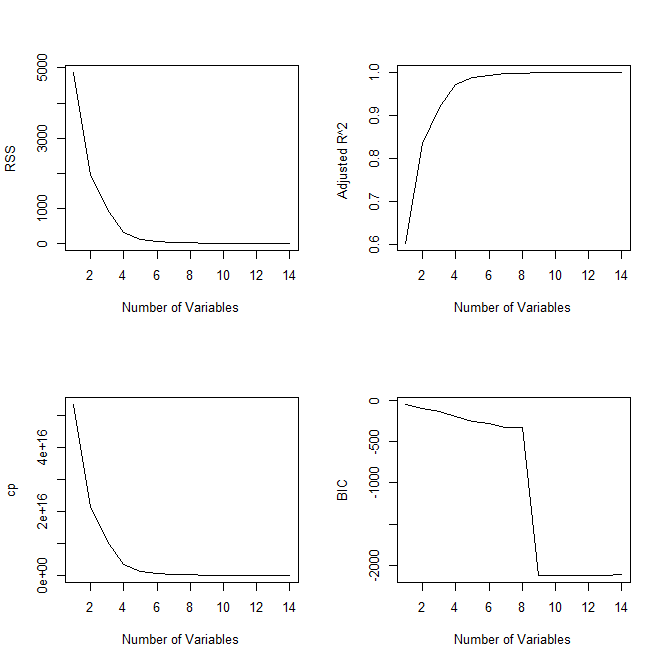
1. **Perform the exhaustive subset section:** I used the function ‘regsubsets’ to distinguish important factors with the ‘exhaustive’ method.
2. **Show the importance of factors in the model:** With function ‘summary’, I could see the importance factor in the model.
3. **Plot the relationship between the number of variables and ‘RSS’,’CP’,’BIC’ and ‘adjusted R^2’** : I can guess what combination factors can mad the model accurately with these 4 plots and the summary of the ‘Forward Selection’.

Out puts:

1. Summary of the ‘Exhaustive Subset Selection’



1. Graphs for the number of variables and RSS, adjusted R^2, cp, and BIC



d)

Consideration Before the Processing:

MSE is one of the best ways for comparing each models’ accuracy. With the Linear regression model, I got the MSE value with test data (In the Question a). In the question b and c. I made the forward subset selection model and exhaustive subset selection model with our train data. In conclusion, I decided to calculate the two model’s MSE and compare three models

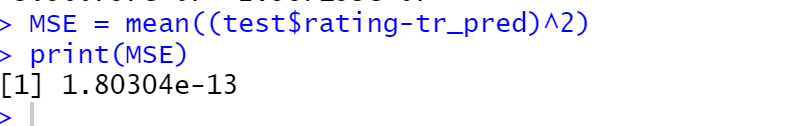
To calculate the MSE, I have to find ‘the lowest MSE’ in the Forward subset selection model and Exhaustive subset selection model. Each model is ordered by the best effective factors. Using all the factor seems to make the model accurate. However, too many factors in the model interfere with each other to predict the values. So, I will calculate the MSE in the model by appending the factors sequentially.

Process:

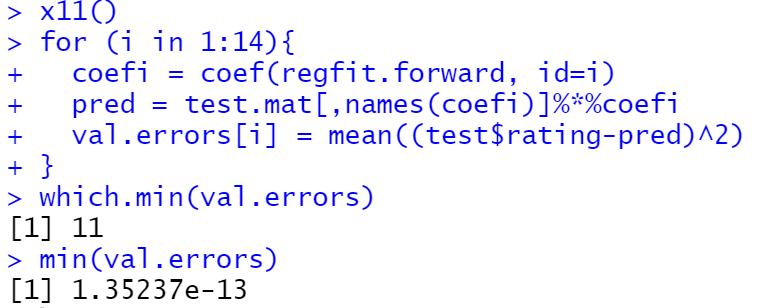
1. **Made a ‘For’ Function**: It starts from ‘1’ which indicates the model includes one the most effective factor, to ‘14’ which indicates the model includes all the factors.
2. **Get coefficient value**: With this coefficient value, it is available for me to calculate the predicted value
3. **Get predict value:** Multiply each test Factor value and coefficient value.
4. **Get MES and Compare the values with a plot:** With the method mean((predicted value – test value)^2) calculate the MSE values and compare it by the number of factors.
5. **Compare three model’s MSE:** Apply 1 – 4 twice (forward subset selection model and exhaustive subset selection model) and compare the MES values to find out the best model.

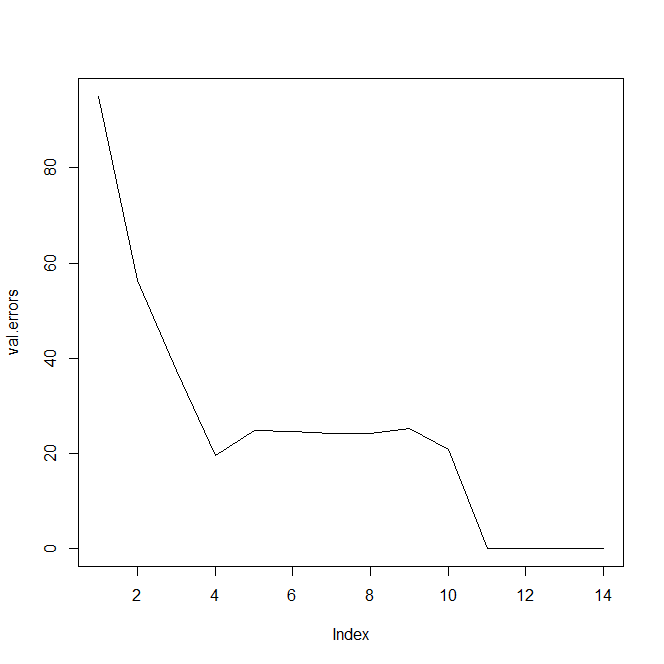
Out puts:

1. **Linear model** (1.80304e-13 MSE)

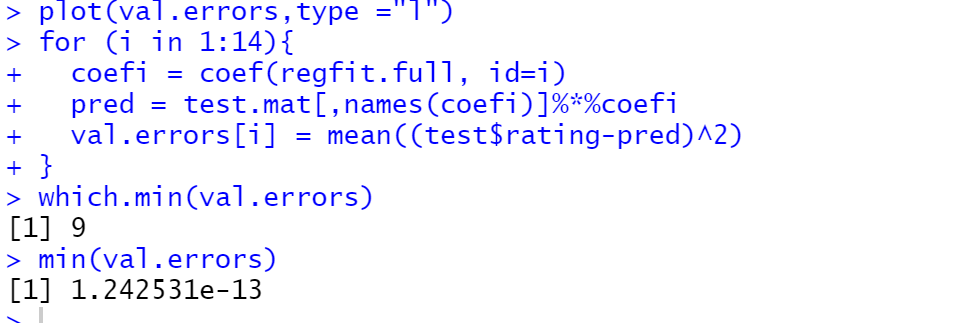


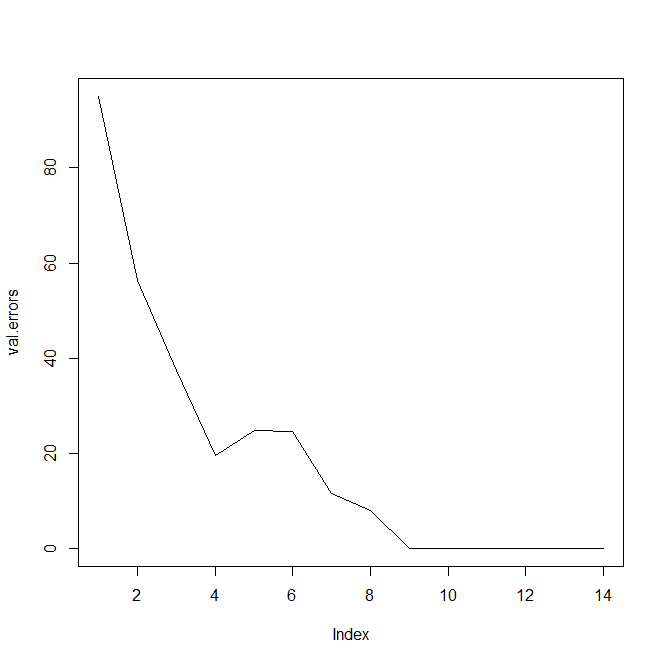
1. **Perform forward subset selection** (1.35237e-13)





1. **Perform exhaustive subset selection**(1.242531e-13)



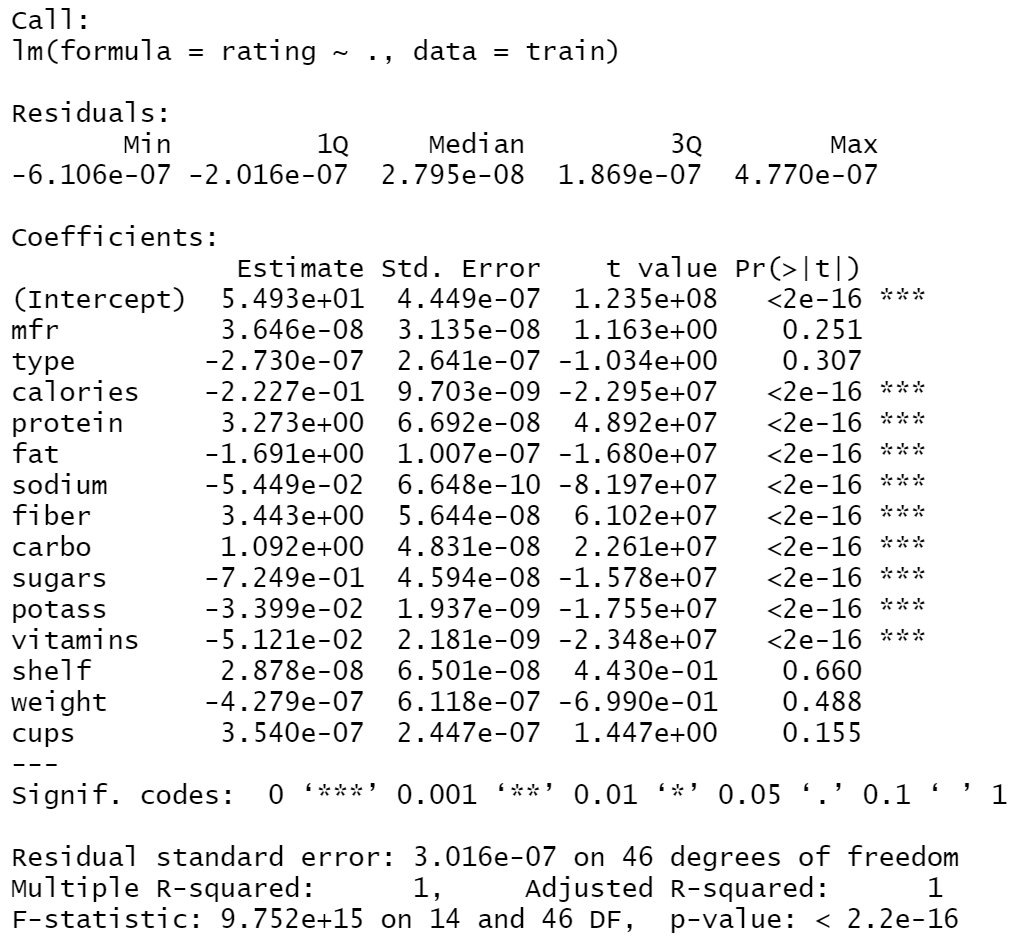


Explanation and Conclusion:

* **Linear model** (1.80304e-13 MSE)
* **Forward subset section** (1.35237e-13 MSE)
* **Exhaustive subset selection**(1.242531e-13 MSE)

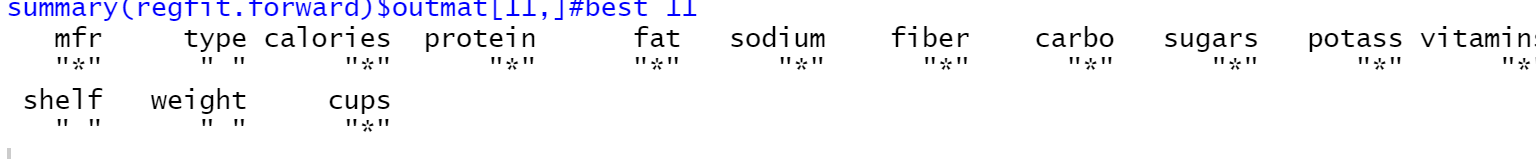
When we compare the three model’s MSE, the best model for predicting the test data is ‘**Exhaustive subset selection**’ This is because MSE shows how different between real data and predict data. The lower MSE means better prediction.

Let’s look at the Linear model summery.



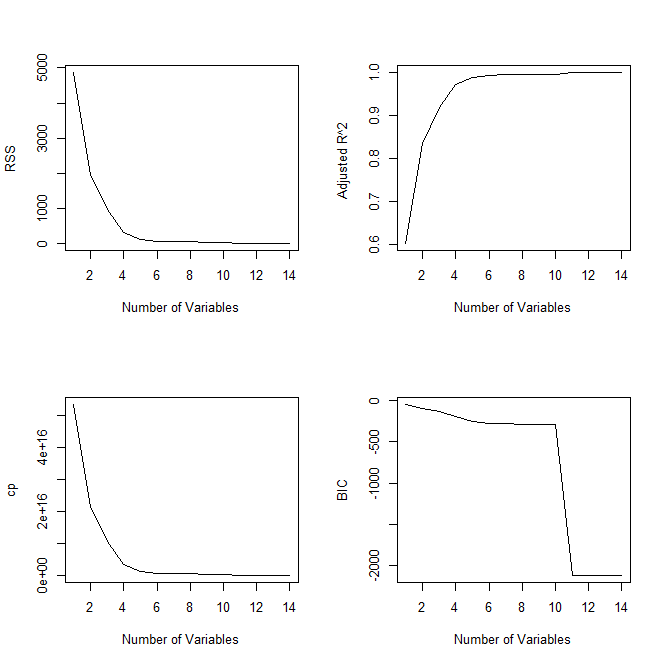
If we see the coefficients cells, there are significant 9 factors for this linear model. This is based on the P-values. Calories, Protein, Fat, Sodium, Fiber, Carbo, Sugars, Potass, Vitamins; these 9 factors have very low P-value.

In the Forward subset selection, using 11 best factors in the model made a lowest MSE (1.341005e-13 MSE)



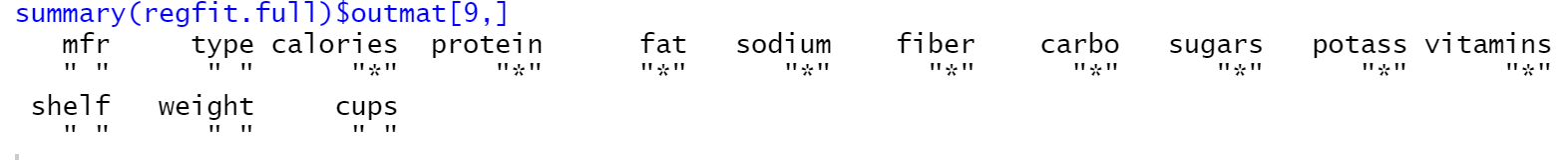
These 11 factors are used in this Forward subset selection model.

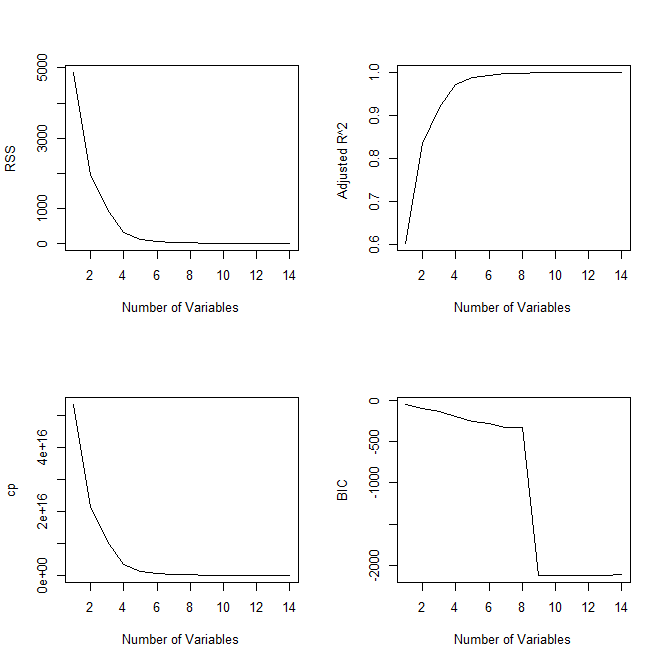
We can also find out optimum number of variables in the model by ‘RSS’, ‘Adjusted R2’, ‘CP’ and ‘BIC’.



In RSS, CP and BIC as the value going lower, it indicates the model is going to be accurate. Otherwise, as adjusted R2 values are going up, the model is going to be accurate. In these 4 graphs, we can guess the best number of the value we get with the MSE is similar to the value from R2, BIC Adjusted R2 and RSS

In the Forward subset selection, using 9 best factors in the model made the lowest MSE (1.242531e-13)



We can also find out optimum number of variables in the model by ‘RSS’, ‘Adjusted R2’,‘CP’and‘BIC’. 

It was a good chance for me to see the fact that using many factors is not the way to make an accurate model.