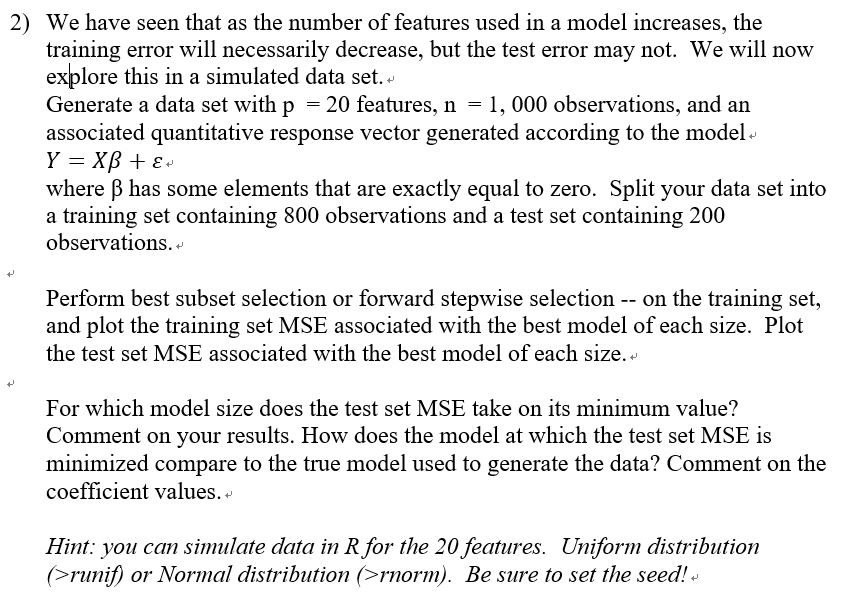
Question 2.

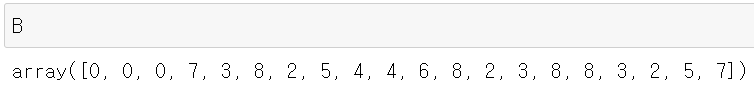


Process:

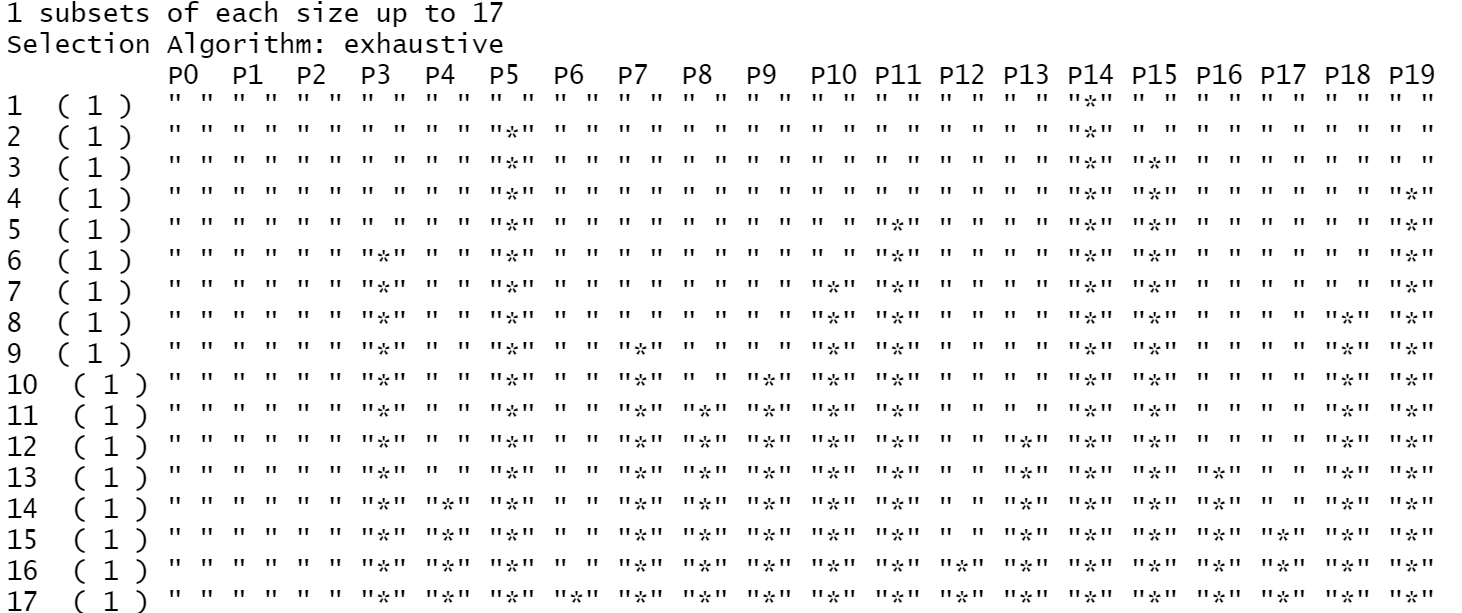
1. **Create data**: I made 20,000(1000 x 20) counts **X** variables. I set these variables with a ‘0’ mean value and the standard deviation is 0.1. Also, I made 20 values of **B**. I set these variables with uniform distribution but I set the first 3 values as 0. I set an **e** value as well randomly.
2. **Make Data frames**: I Multiplied **B** values by each **X** data columns and put them in the Data frame. Add one more column and fill it with the sum of each row data plus **B.**
3. **Split the Train and Test data**: I made an index to separate the train and test data.
4. **Fitting the exhaustive subset selection model with the data**: By using exhaustive subset selection, find out what some of the important variables among all the variables.
5. **Calculate the MSE of the Linear model**: By the number of variables, I put the test and train data into the model and calculated the predicted data for test and train data. Also, I made each MSE list for train and test data by the number of variables as well.

Outputs:

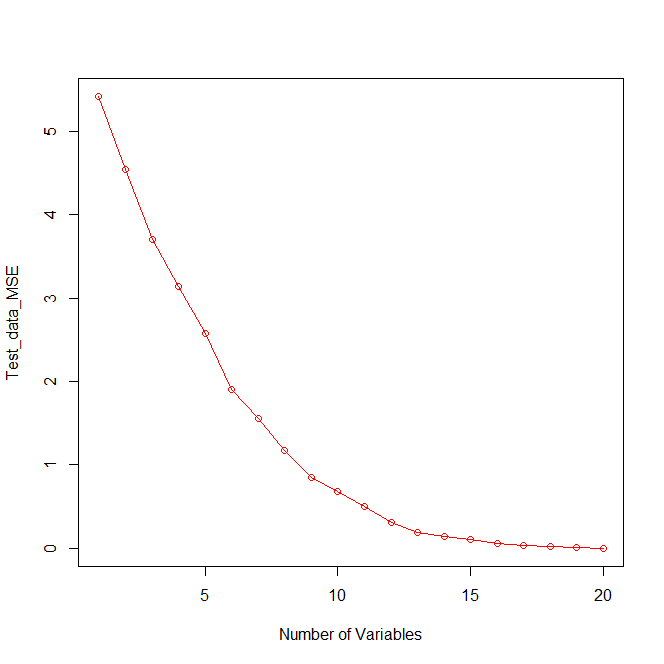
List of **B**



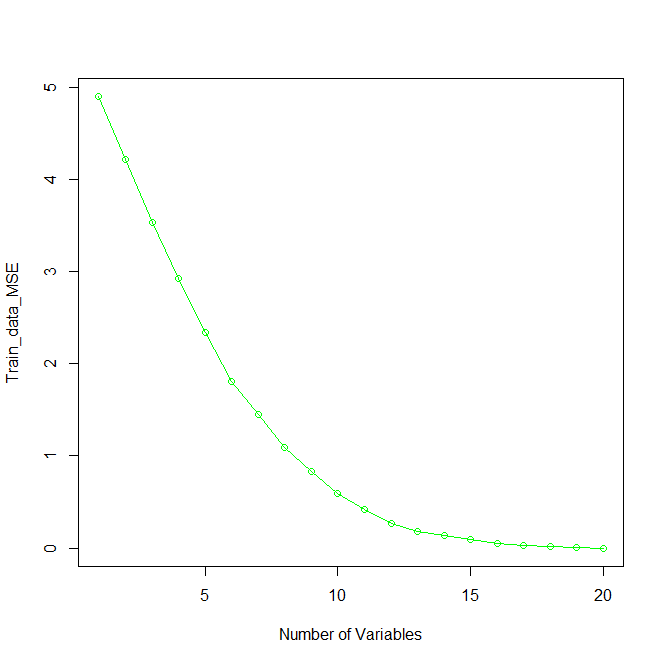
Summary of Exhaustive subset selection



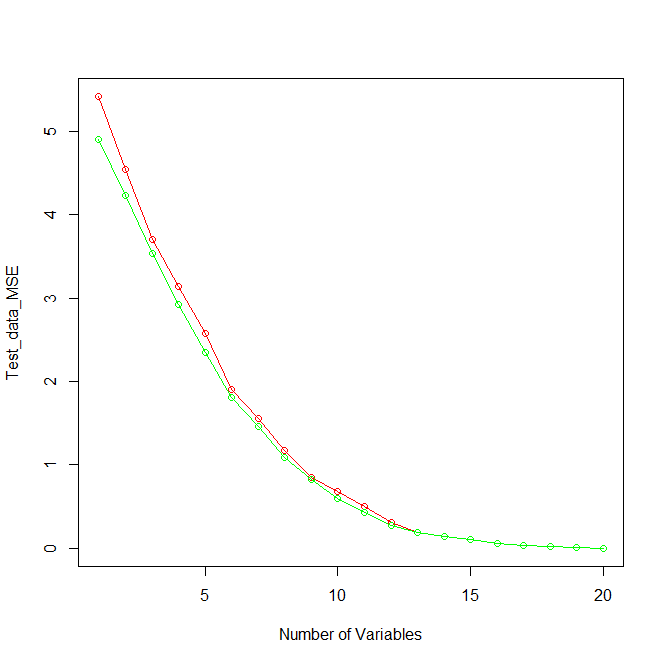
Test data MSE by variables



Train data MSE by variables



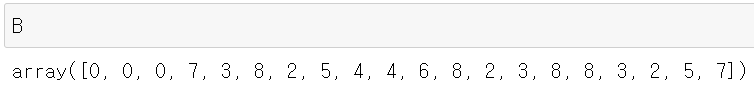
Test data and traind data MSE by variables



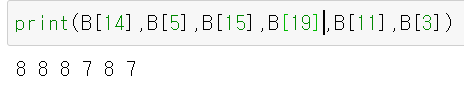
Discussion:

**What Variables have a huge impact on model prediction:**

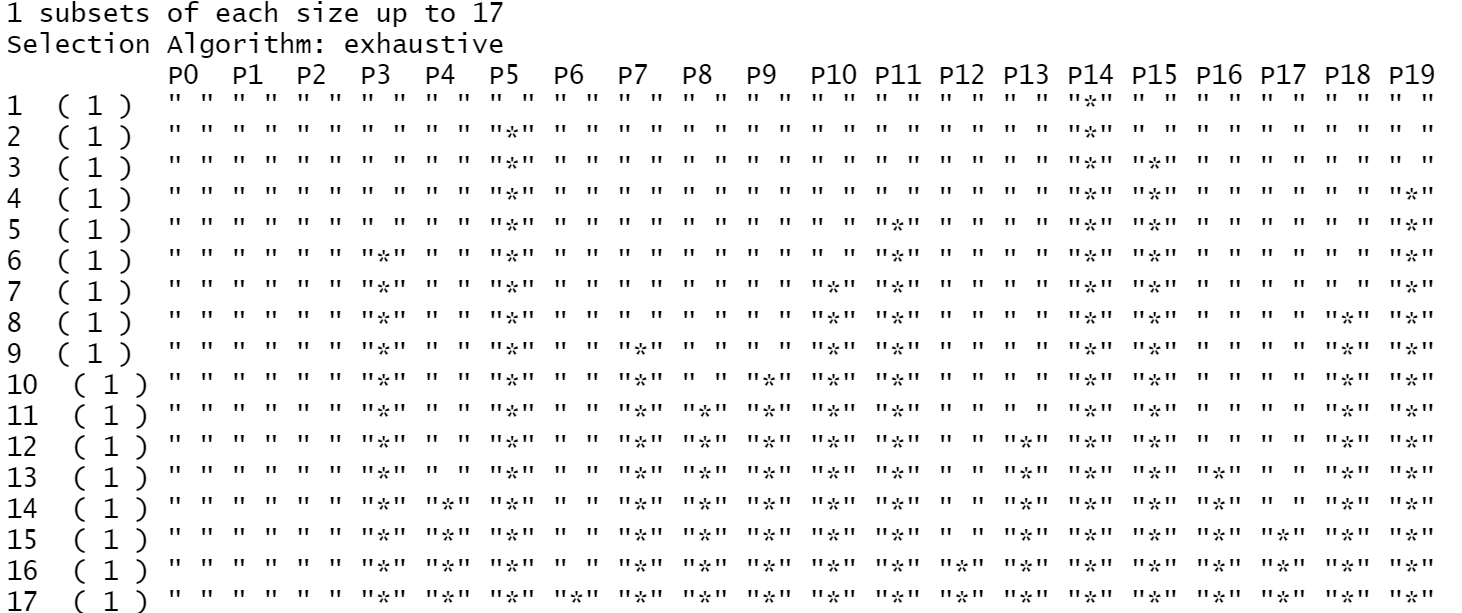
Before I run the model, I checked **B** that I set. Our data is made by 20 linear equations and each column is independent. This implies that the values of **B** directly apply in the Y as its’ value size. Since there are three ‘0’ values in **B**, I guessed that the data applied these three 0 value **B** are not important variables in my model. For this reason, the best subset model that I made with this data will show that the first 3 columns (value 0) are not important as other columns. With the similar hypothesis, the importance column for the model is about the size of **B.**



Let’s see the list of the above the picture. The largest number in the list is ‘8’, and there are four ‘8’ in the list. Likewise, the second largest number is ‘7’ and there are two ‘7’ on the list.



Above the picture, we can find out what column in the data indicate the values ‘8’ and ‘7’(column 3,5,11,14,15,19). Next, we will compare with the summary of our exhaustive model selection.

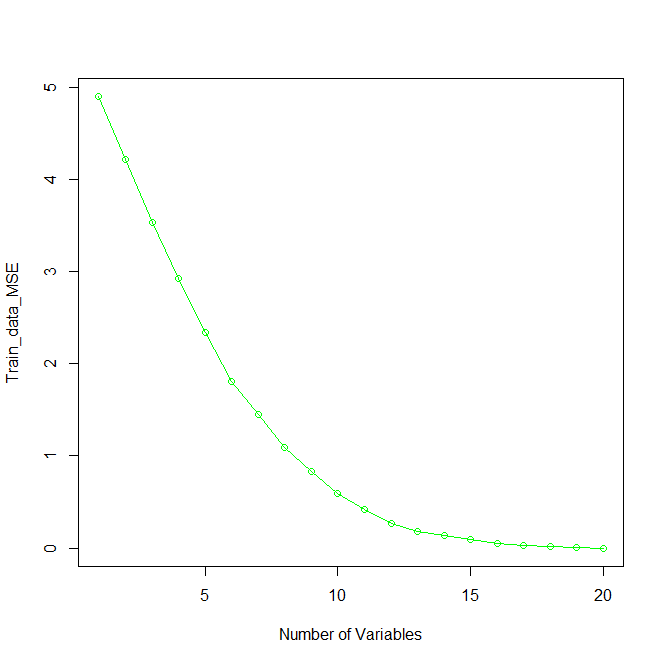


With the ‘summary’ function with the model, we can find out what column is important for the model. The top 6 variables are matched with columns that B value for ‘7’ or ‘8’.

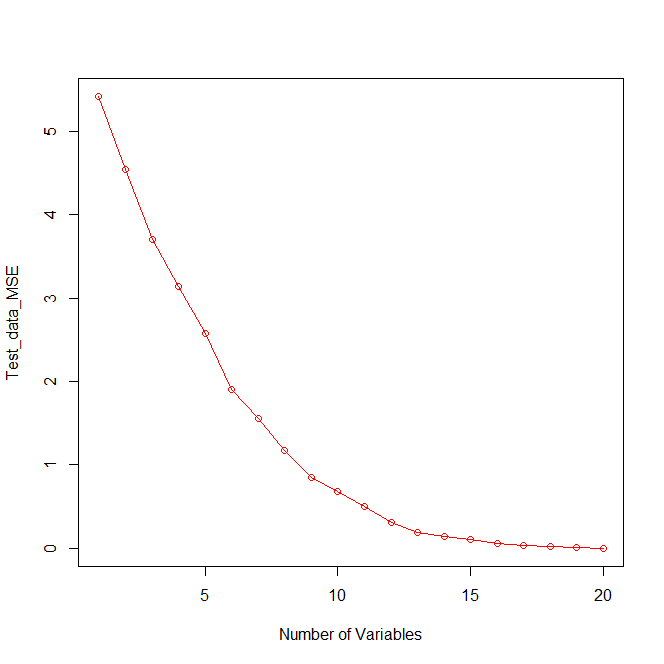
With this comparison, it is to say that the column has large values of ‘**B**’ is important in this model.

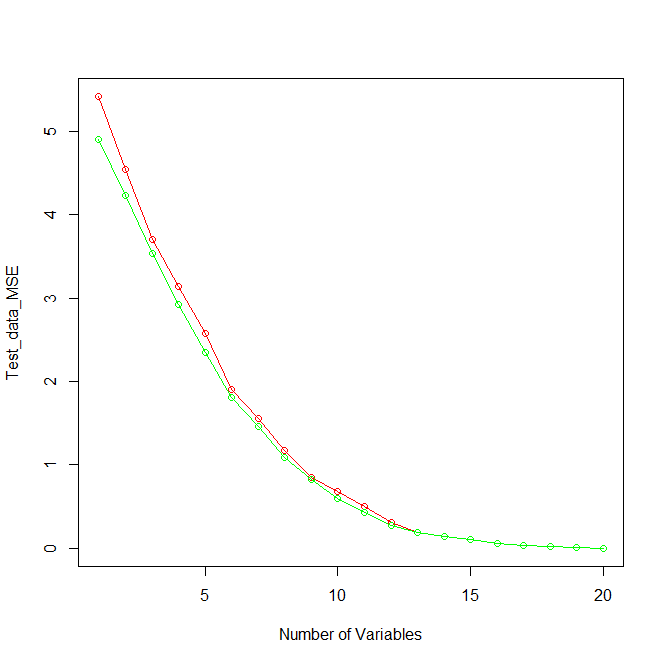
**MSE of Test and Training data by the number of variables in the model.**

As I mentioned above, this model is fitted by the data that is made by 20 linear equations and each column is independent. Each column has information so that using many columns for the data do not cause ‘dimensional cures. So that as the number of columns used in the model larger, the model prediction will be accurate.



Above picture, you can see the large number of variables return less MSE.

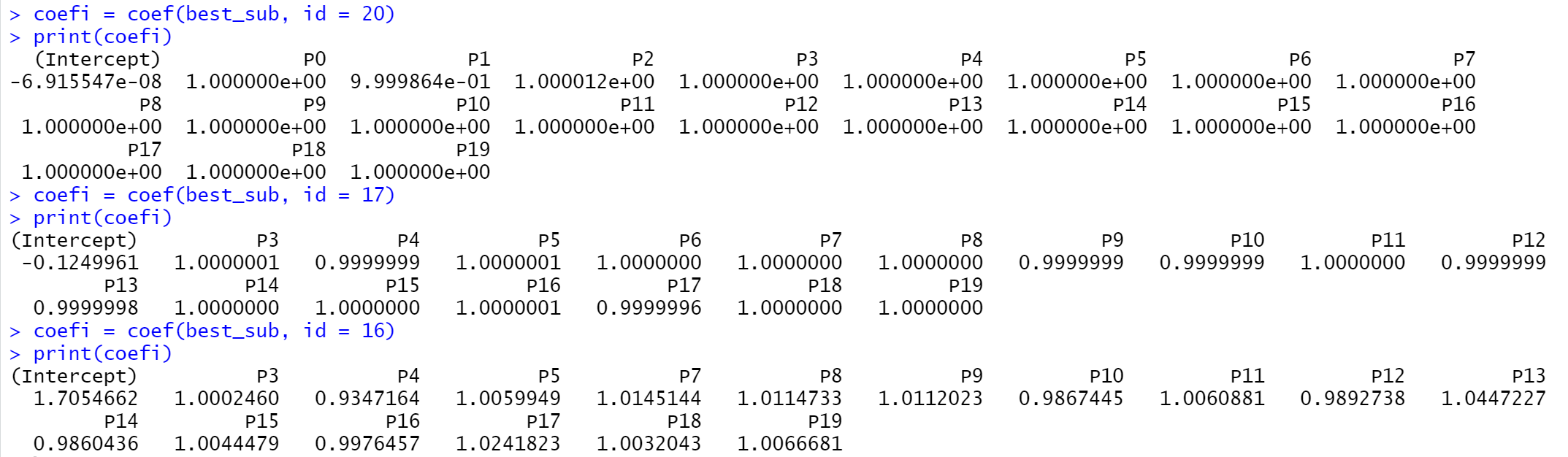




When we see the test MSE by the number of variables, the similar phenomenon can be found. This is because I made the variable with normal distribution so that it is hard to find out any outliers which can deter the fitting model or predicting data.

Since the train and test data include 3 columns of 0 value. These three columns will not affect on the model prediction. For this reason, the model with 18, 19, and 20 columns.

So, In my opinion, the best variable for the model is 17 which is the total variable number – the sum of columns with 0 value of ‘**B**’.



Let’s look at the above picture. (Remind that my data has **three** 0 value of ‘**B’.)** This showsCoefficient values for each column.In the best subset selection model with 20 variables and 17 variables, most of the coefficient values are about exact 1. However, when I see the best subset selection model with 16 variables, the coefficient values changed a lot compared with 17 variables. This indicate that 3 columns in the data with 0 ‘**B**’ value, got less information