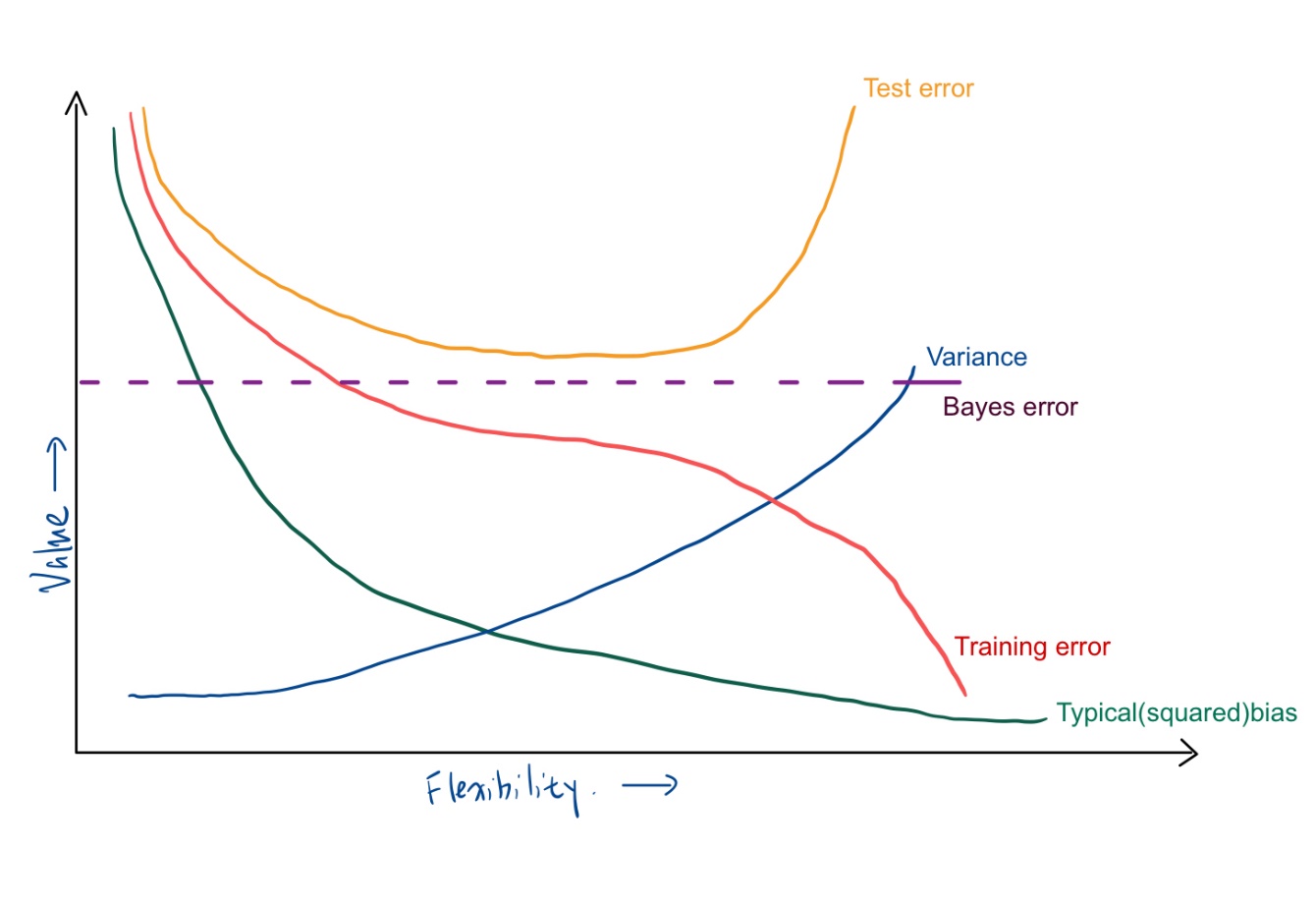
ECE 625 Assignment\_1

**Q1**

**(a).**

**(b)**

Training error: The training error decreases monotonically as the flexibility of the model increases and the function over-fits the data, because as the flexibility increases, the curve fits the observed data more closely, and in the end the error can be treated as a constant.

Typical bias: The bias represents the difference between the predicted and true value, and as the flexibility increases creating a more complex function that better reflects the true problem, the difference between the predicted and true values is smaller.

Variance: The variance reflects the amount of change in the value of the function caused by the use of different training data sets, with the increase in the degree of fit, any point of change may lead to a large change in the value, so the variance will be monotonically increasing.

Test error: The test error gradually decreases as the flexibility increases and at some point will equalize, however, as the fit to the data increases and the model is over-fitted, the test error will begin to increase.

Bayes error: The bayes error is calculated from the training data and is a constant that does not change due to the flexibility of the model.

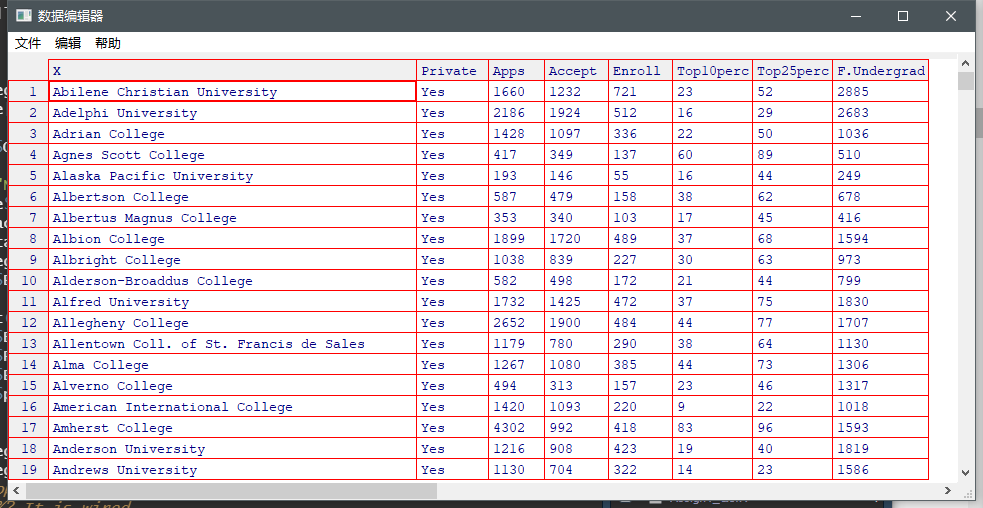
**Q2**

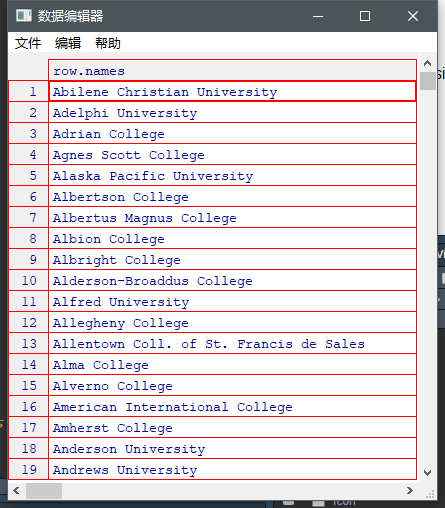
For a very flexible approach, the advantages of it is the bias is small, and it can fit the data better for a non-linear model. The disadvantages of it is it need a large number of parameters, it is easy to overfit and the variance is large.

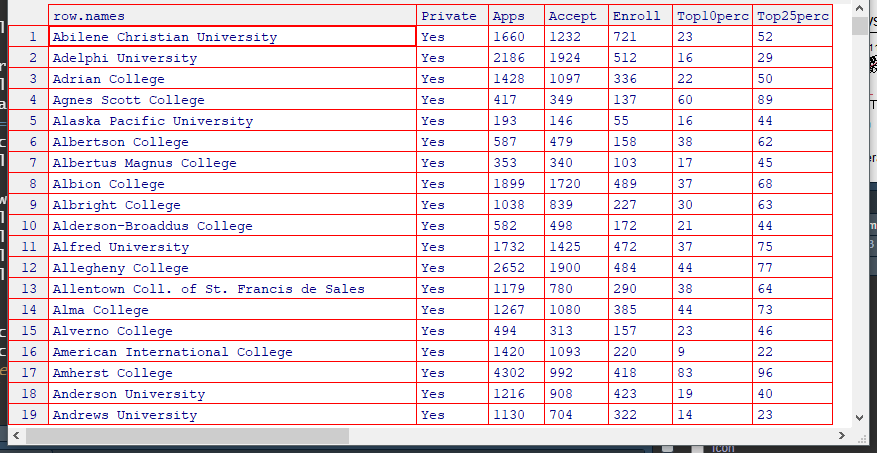
When we have high requirements for the interpretability of the model, less flexible approach is better than more flexible approach.

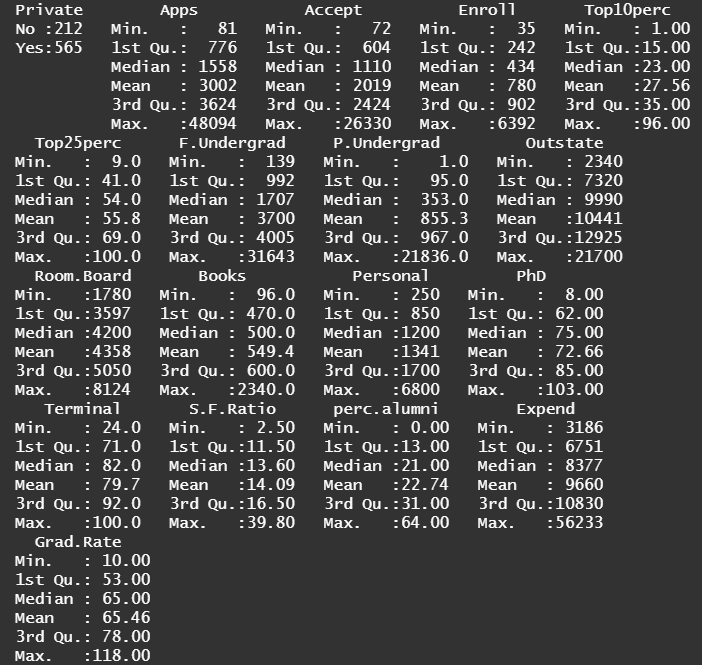
When we are more interested in the accuracy of the prediction, more flexible approach is preferred.

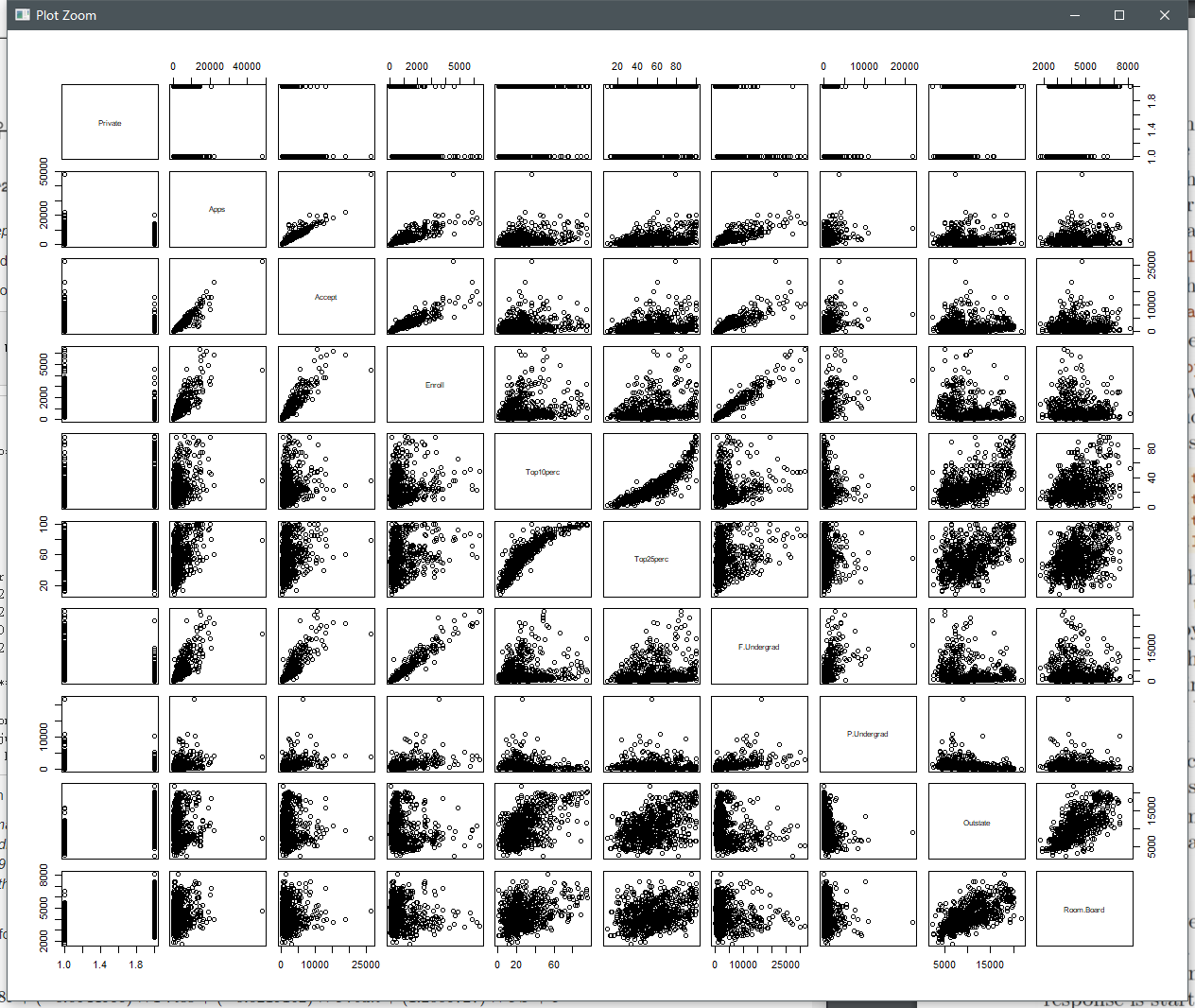
**Q3**

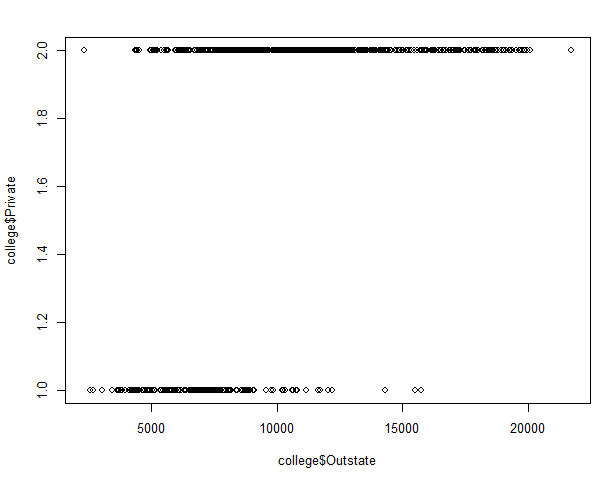




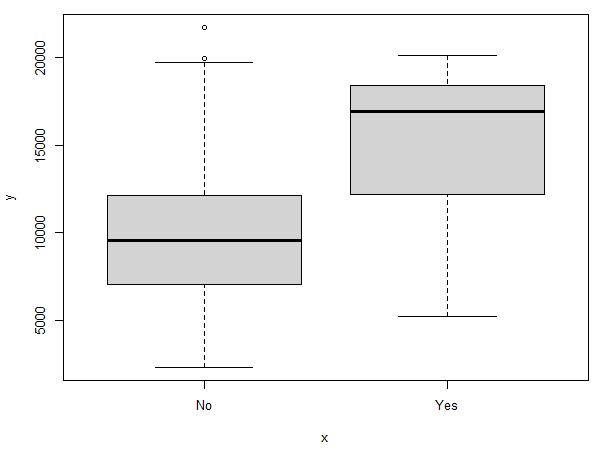


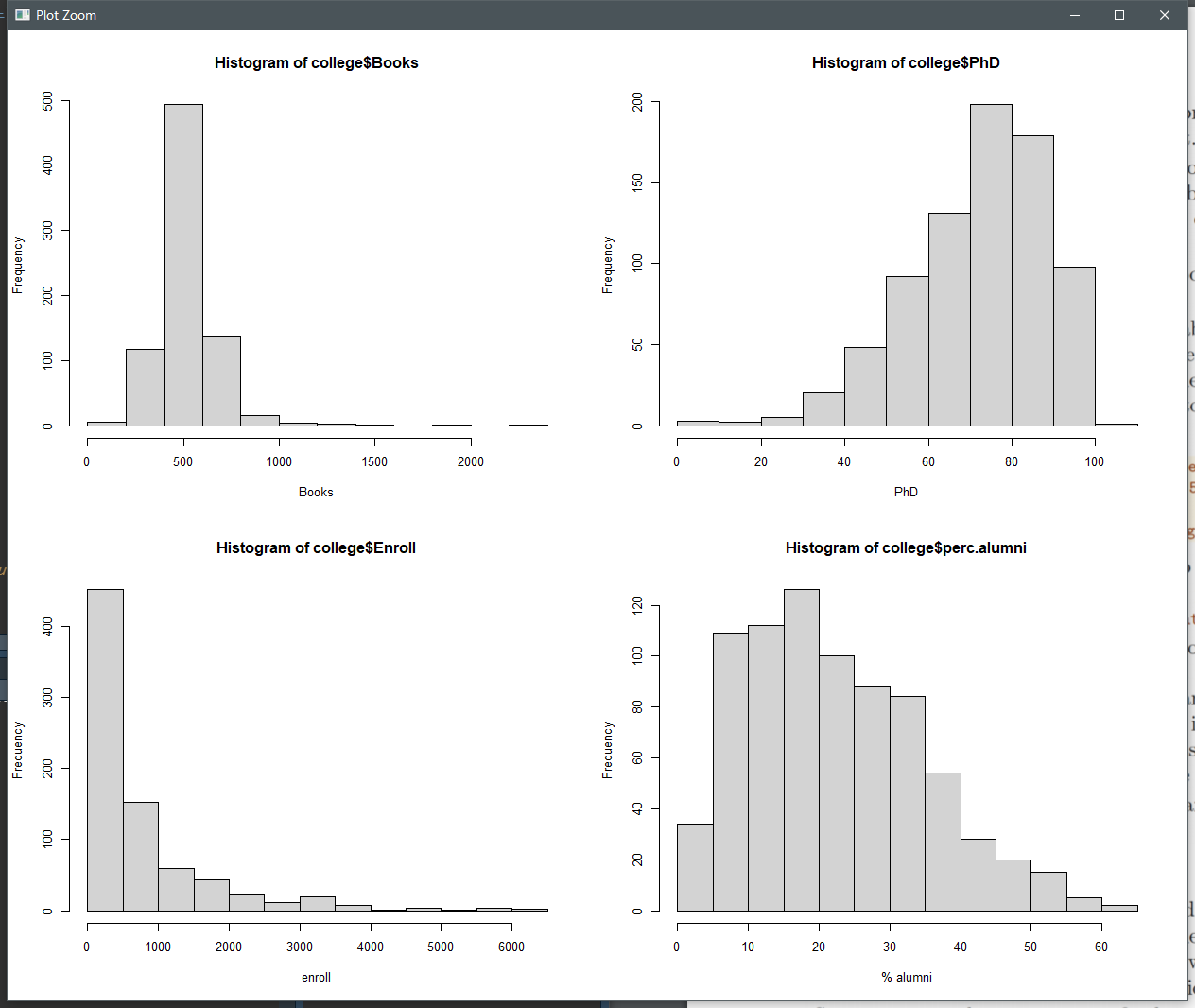












**Q4**

**(a)**

iii. is correct.

Based on the given condition we can derive:

*Y = 50 + 20GPA + 0.07IQ + 35Gender + 0.01GPA x IQ – 10GPA x Gender*

For female we can get:

*= 85 + 10GPA + 0.07IQ + 0.01GPA x IQ*

For male we can get:

*= 50 + 20GPA + 0.07IQ + 0.01GPA x IQ*

For fixed value of GPA and IQ without specific value, we can not get result that female or male is better. So i and ii are incorrect. For answer iii, 50 + 20GPA > 85 + 10 GPA we can get solution that GPA > 3.5, so if GPA is high enough, males can earn more than females. So answer iii is correct.

**(b)**

85 + 10 x 4 + 0.07 x 110 + 0.01 x 4 x 110 = 137.1 thousands of dollar = 137100$

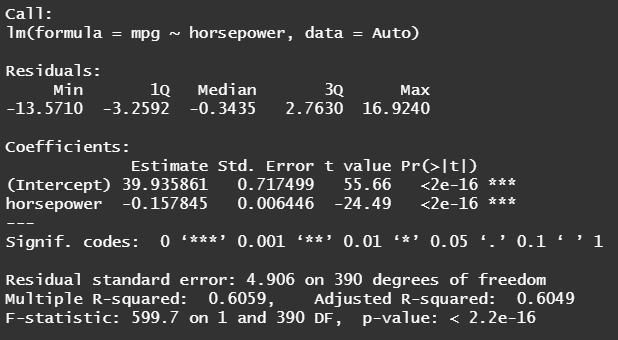
**(c)**

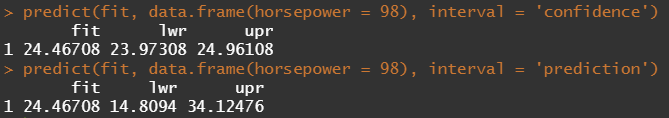
False

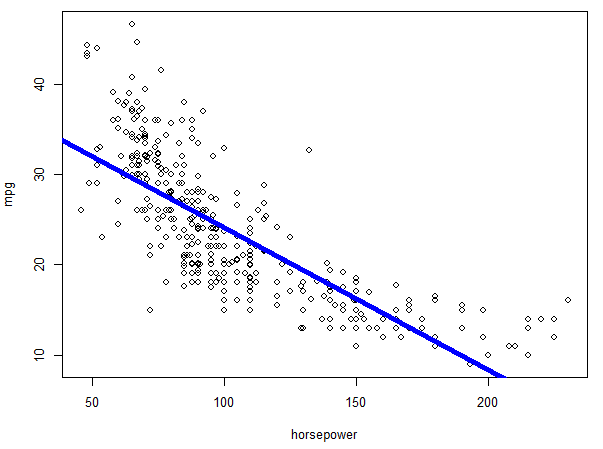
The small coefficient does not indicate the less infect, the degree of influence of the coefficients on the results needs to be judged according to the p-value, so we need to check the p-value of the interaction.

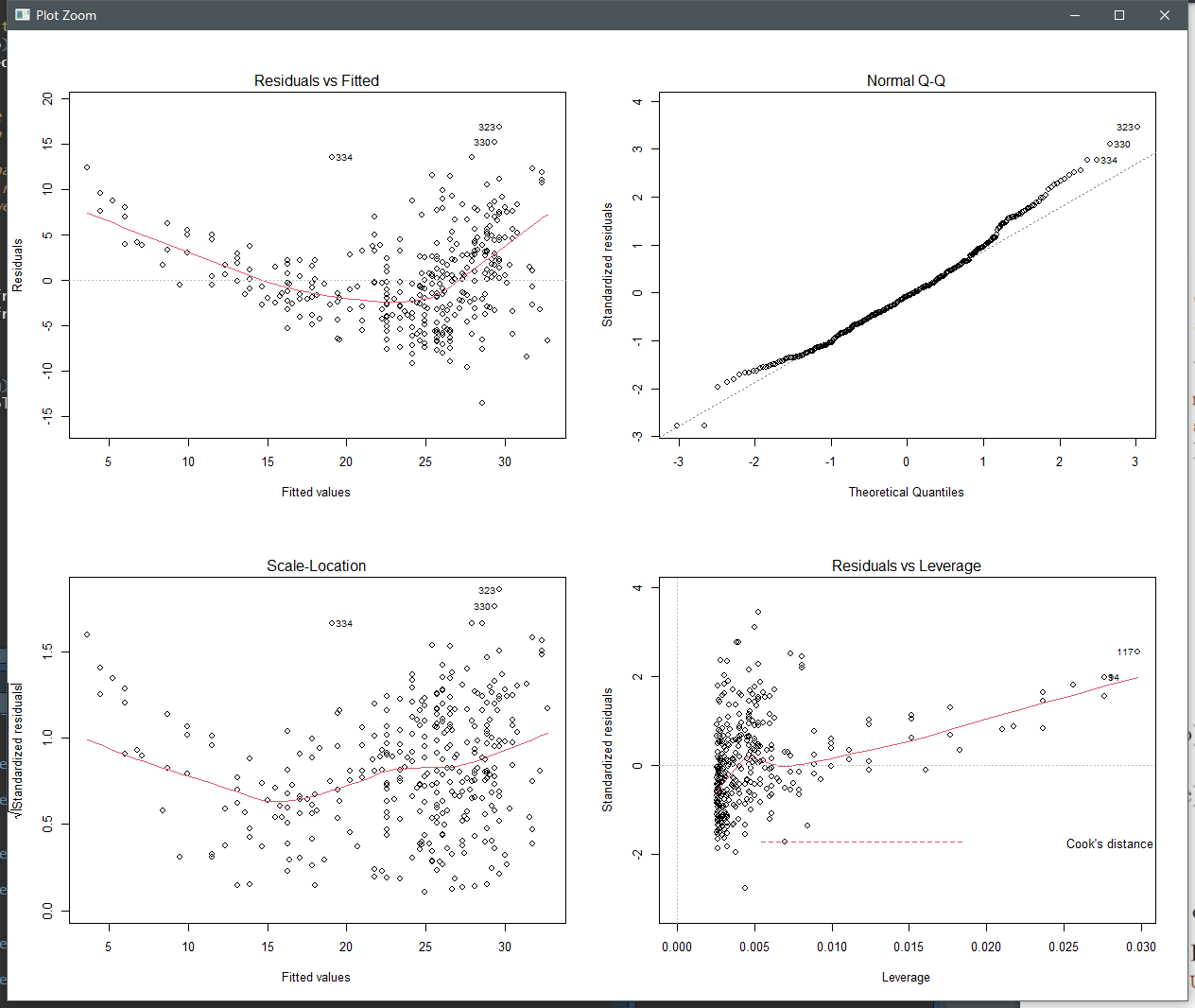
**Q5**

**(a)**



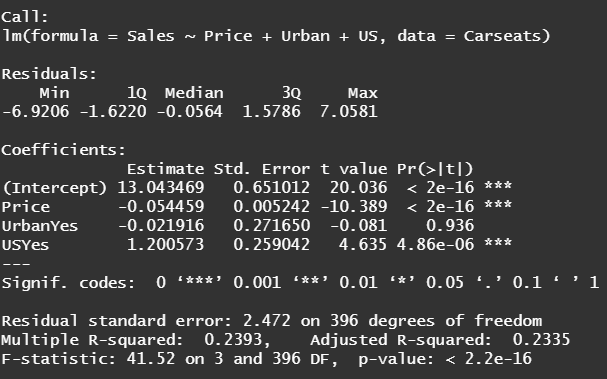






**Q6**

**(a)**



**(b)**

Price: when price increase 1$, sales will decrease 0.05449 unit sales, all other predictors are fixed.

Urban: The average unit sales in urban area was 21.9161 units less than that in rural areas, all other predictors are fixed.

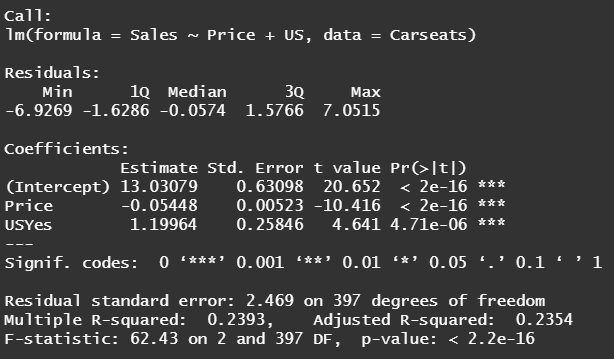
US: Average unit sales in a United State store are 1200.573 units more than in a non-US store, all other predictors are fixed.

**(c)**

**(d)**

Price and US

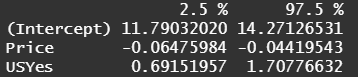
**(e)**



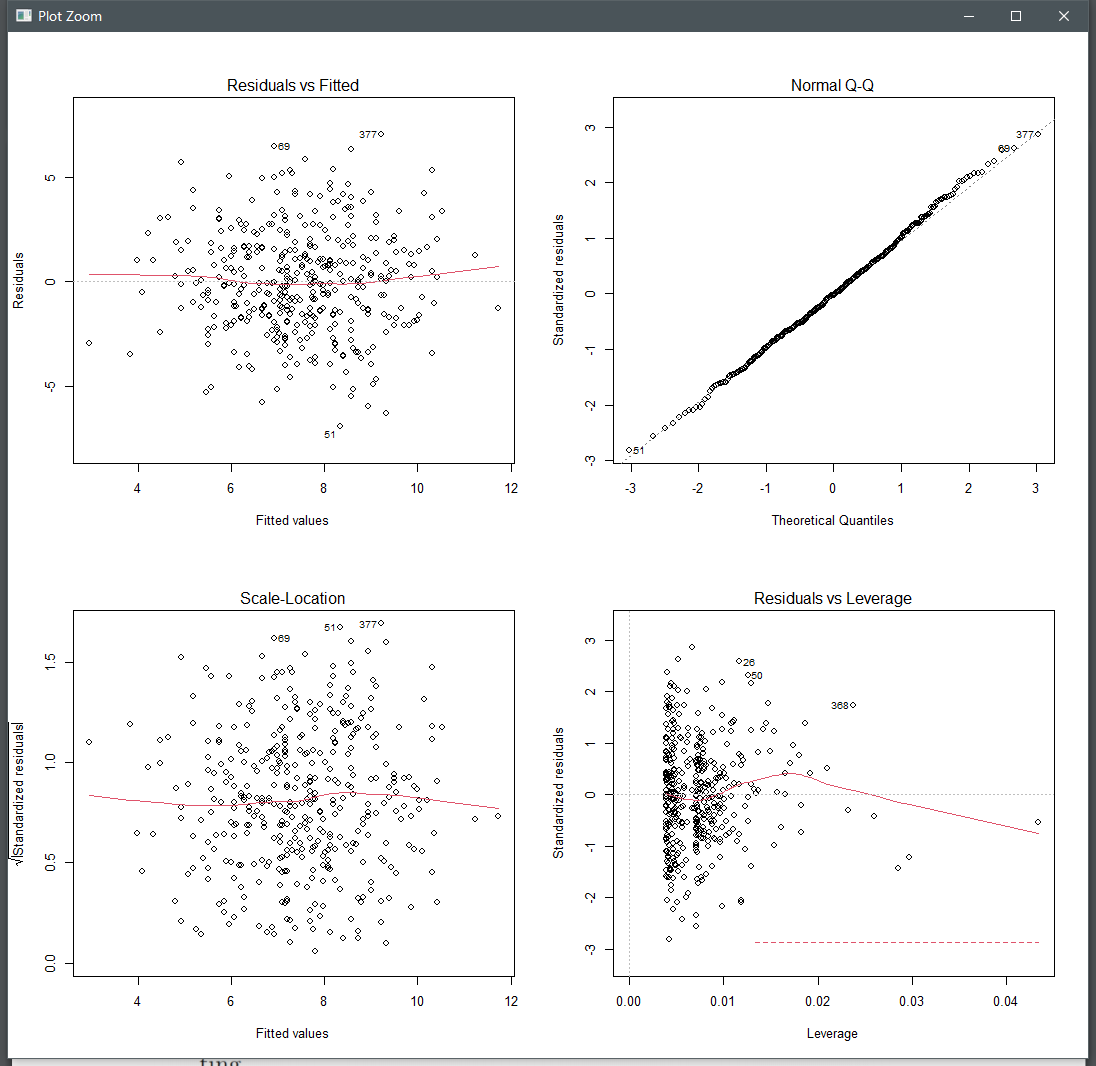
**(f)**

We can see the R-square values, the performance of both models was modest, only 23.93% change in response explained.

**(g)**



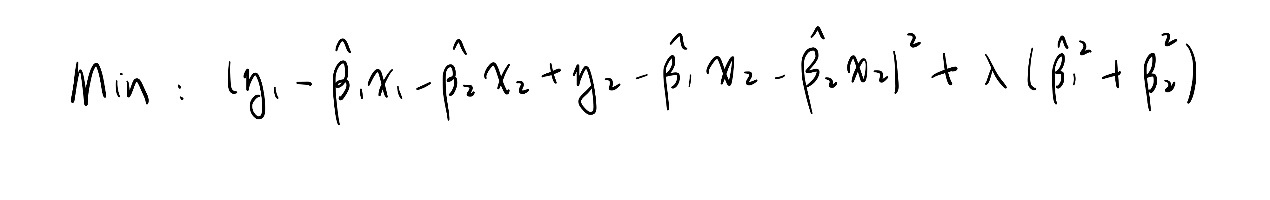
**(f)**



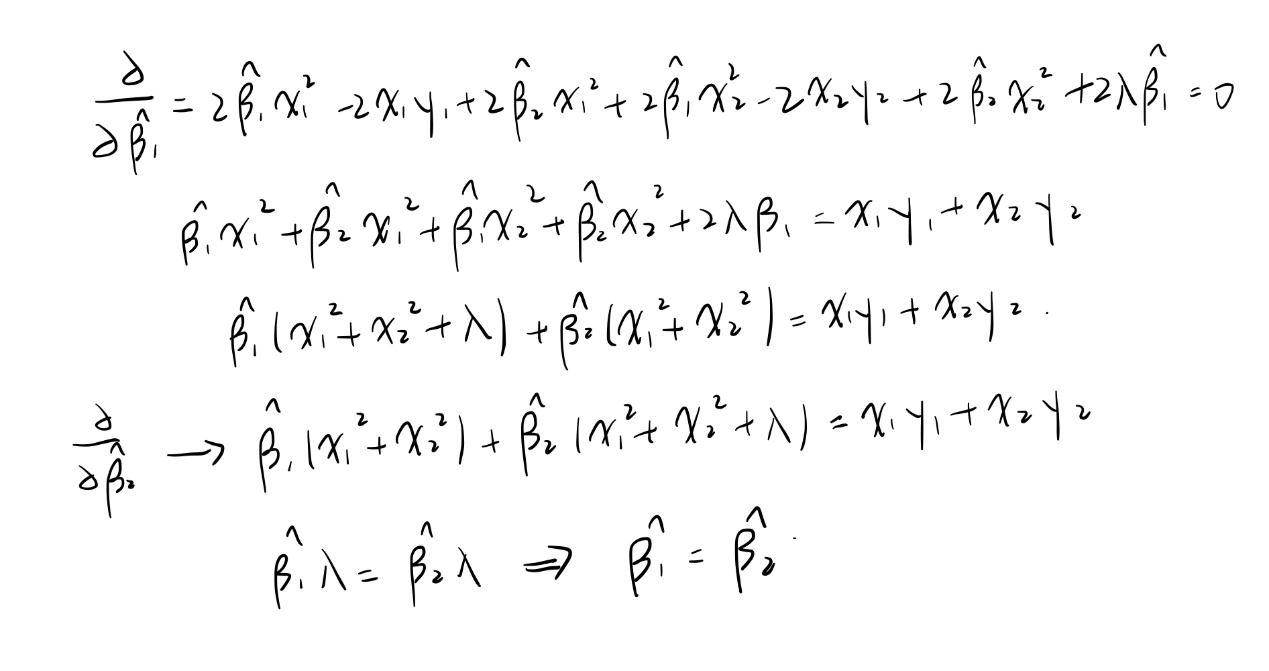
We can find few outliers in the Normal Q-Q figure, and some leverage points in Residuals vs Leverage figure.

**Q7**

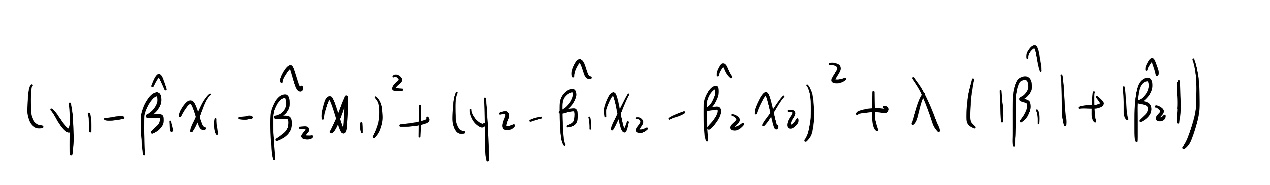
**(a)**

****

**(b)**

****

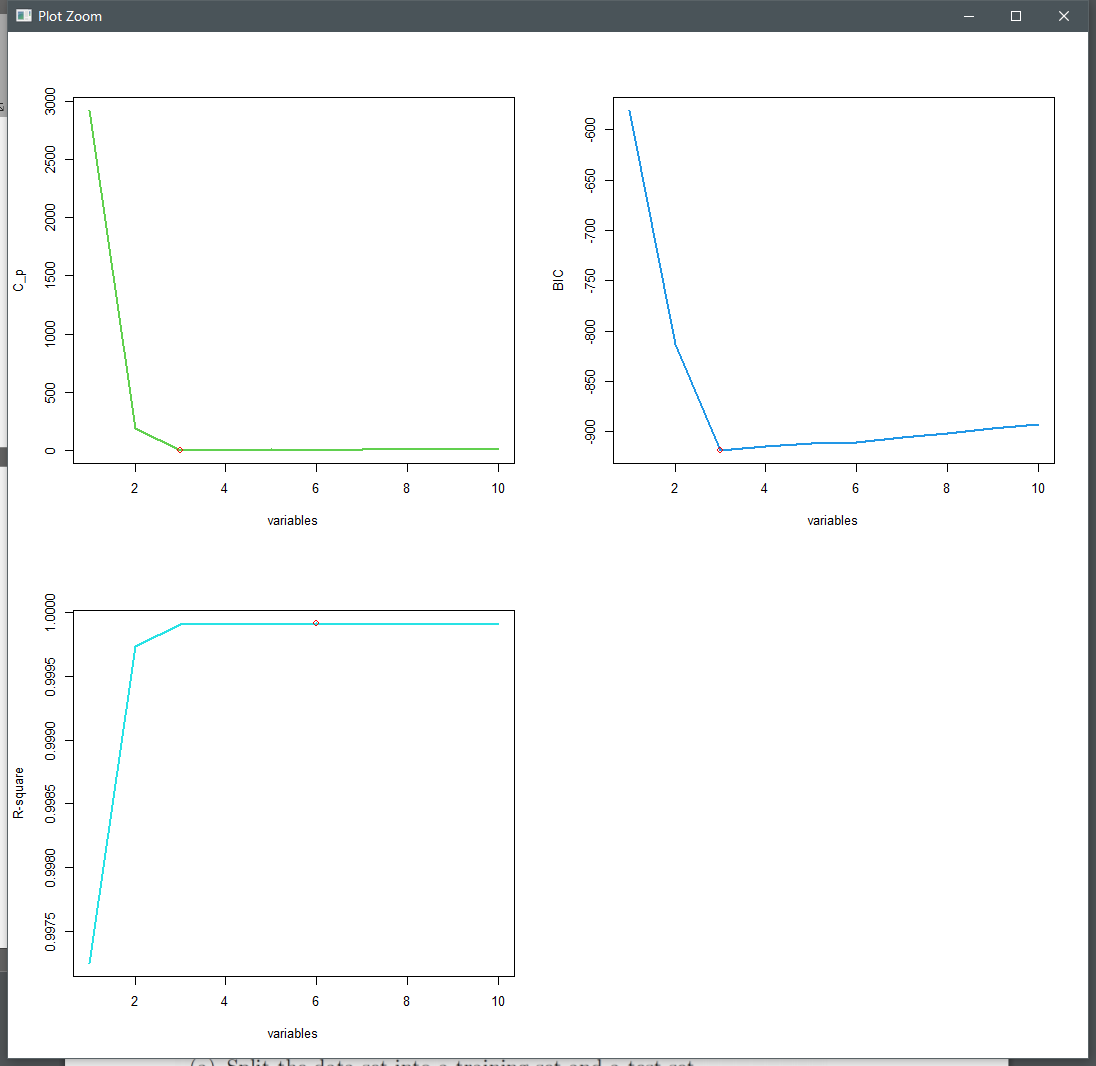
**(c)**

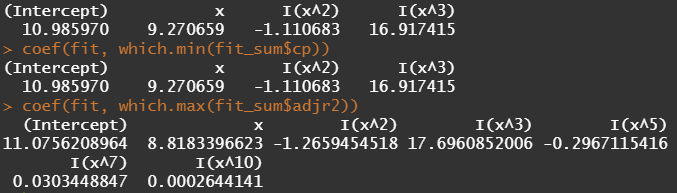
****

**(d)**

**Q8**

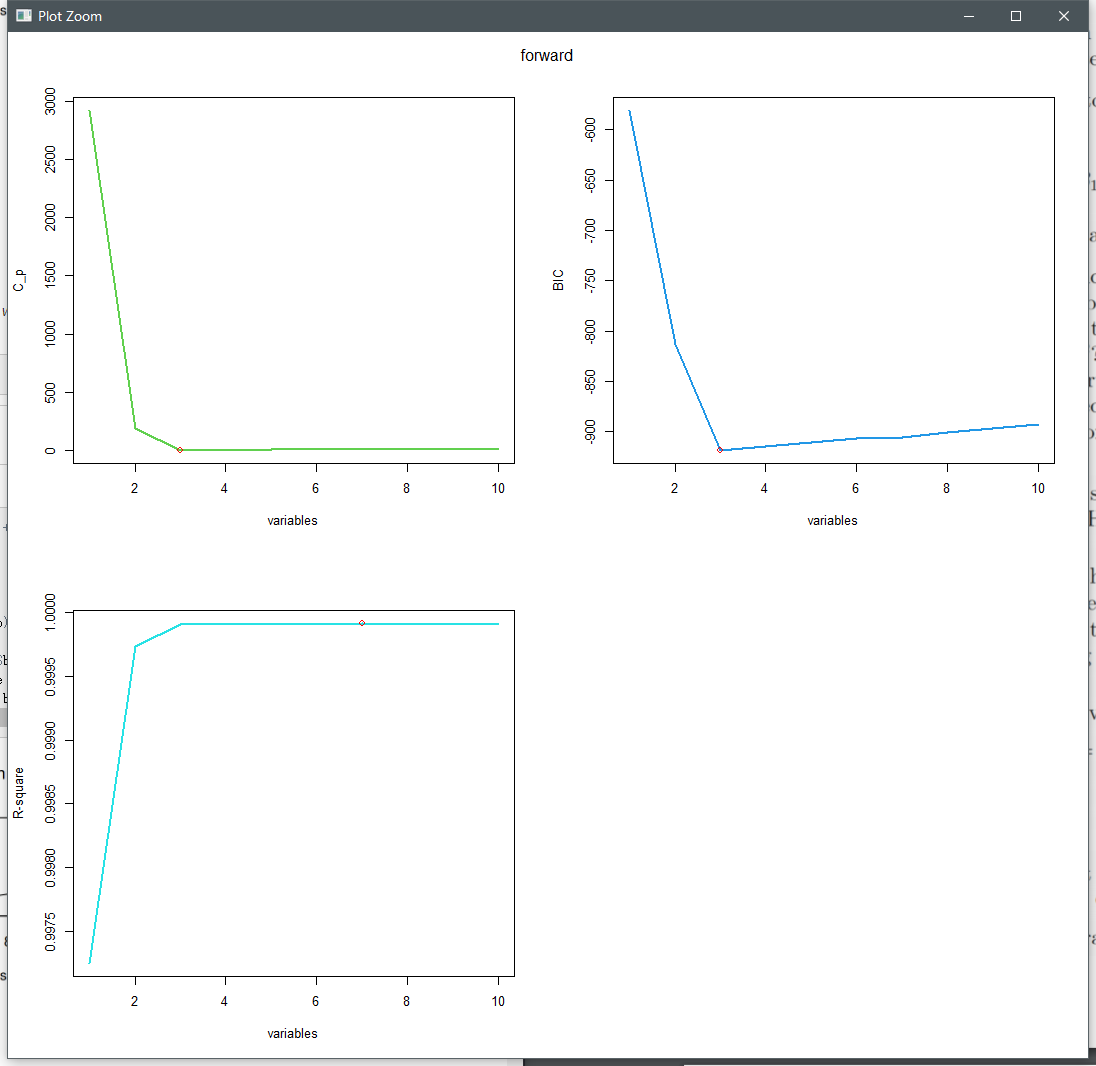
**(abc)**

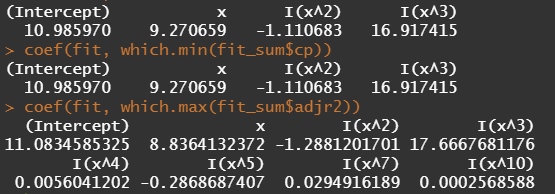




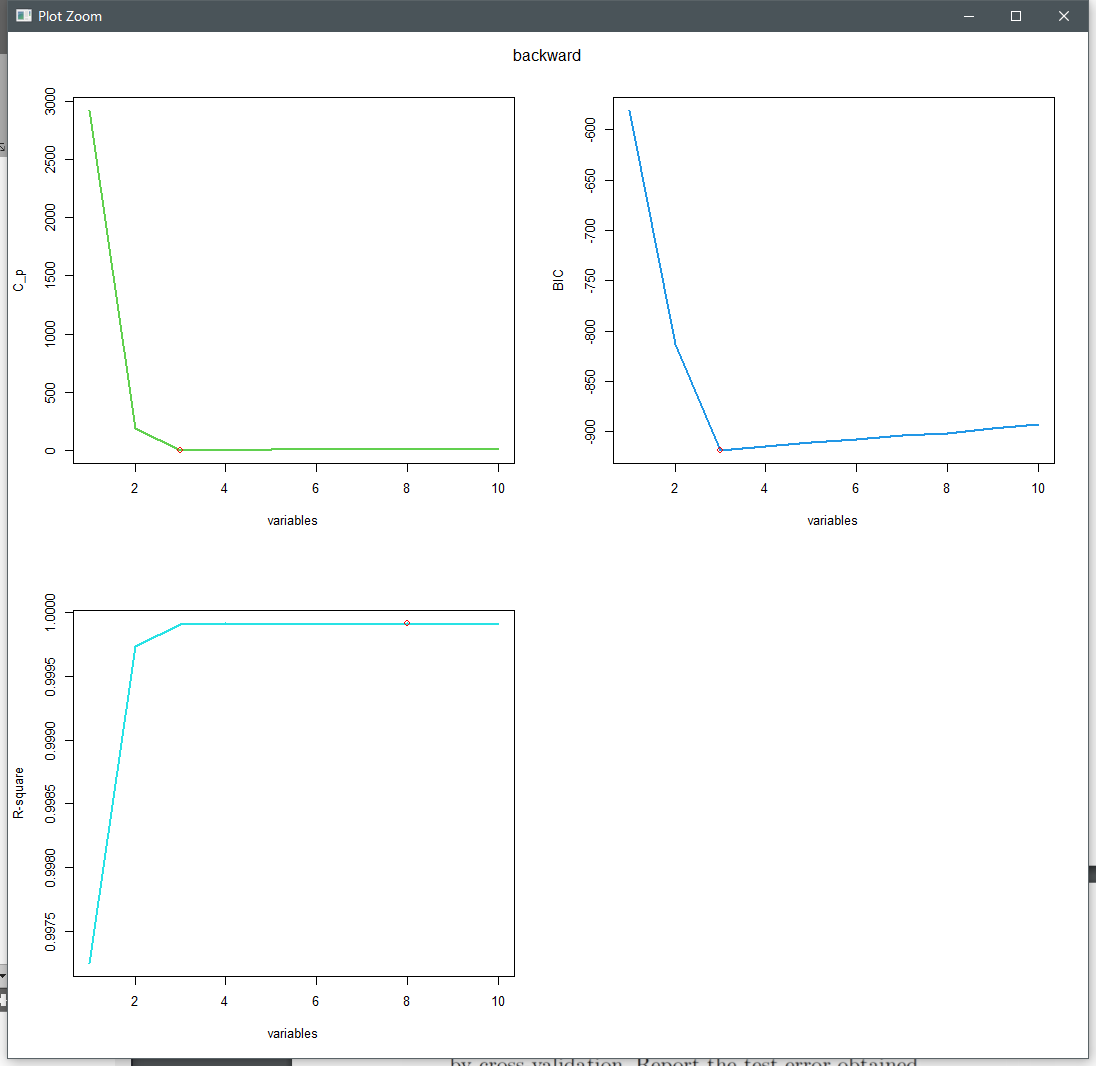
Cp and BIC choose three variables, R-square choose six variables

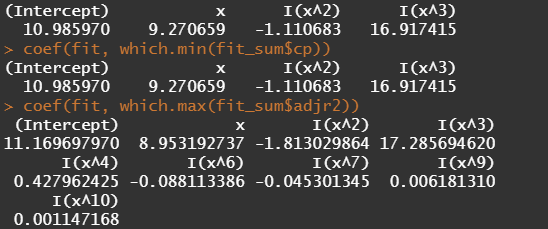
**(d)**





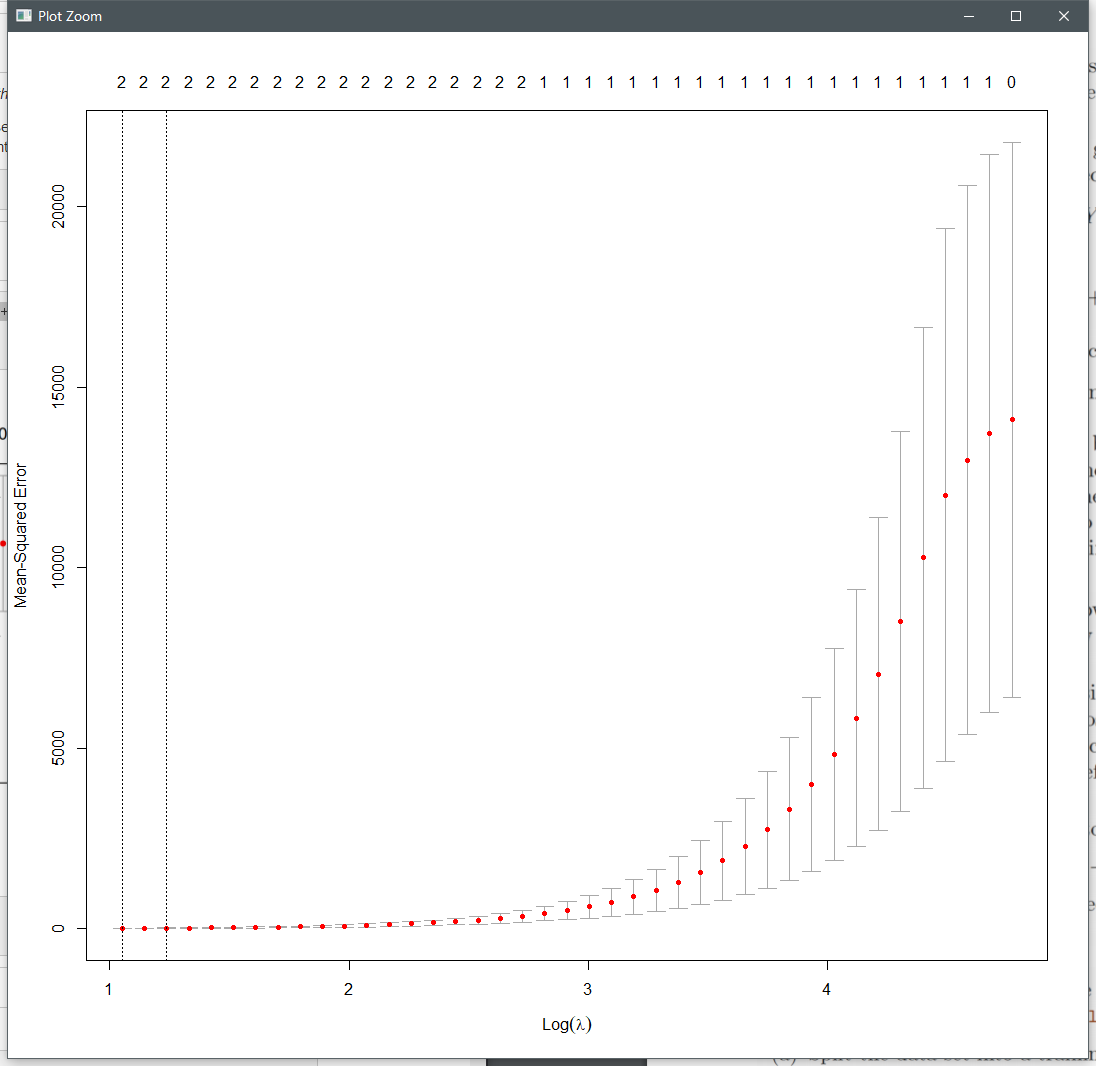
Forward: Cp and BIC choose three variables, R-square chooses seven variables

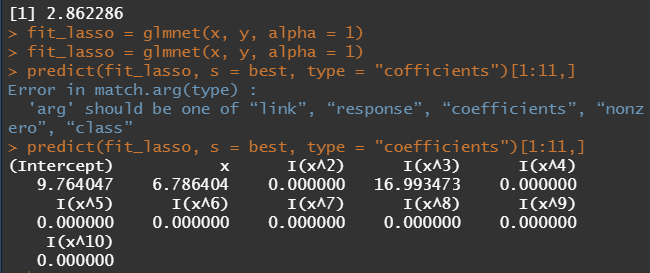




Backward: Cp and BIC choose three variables, R-square chooses eight variables

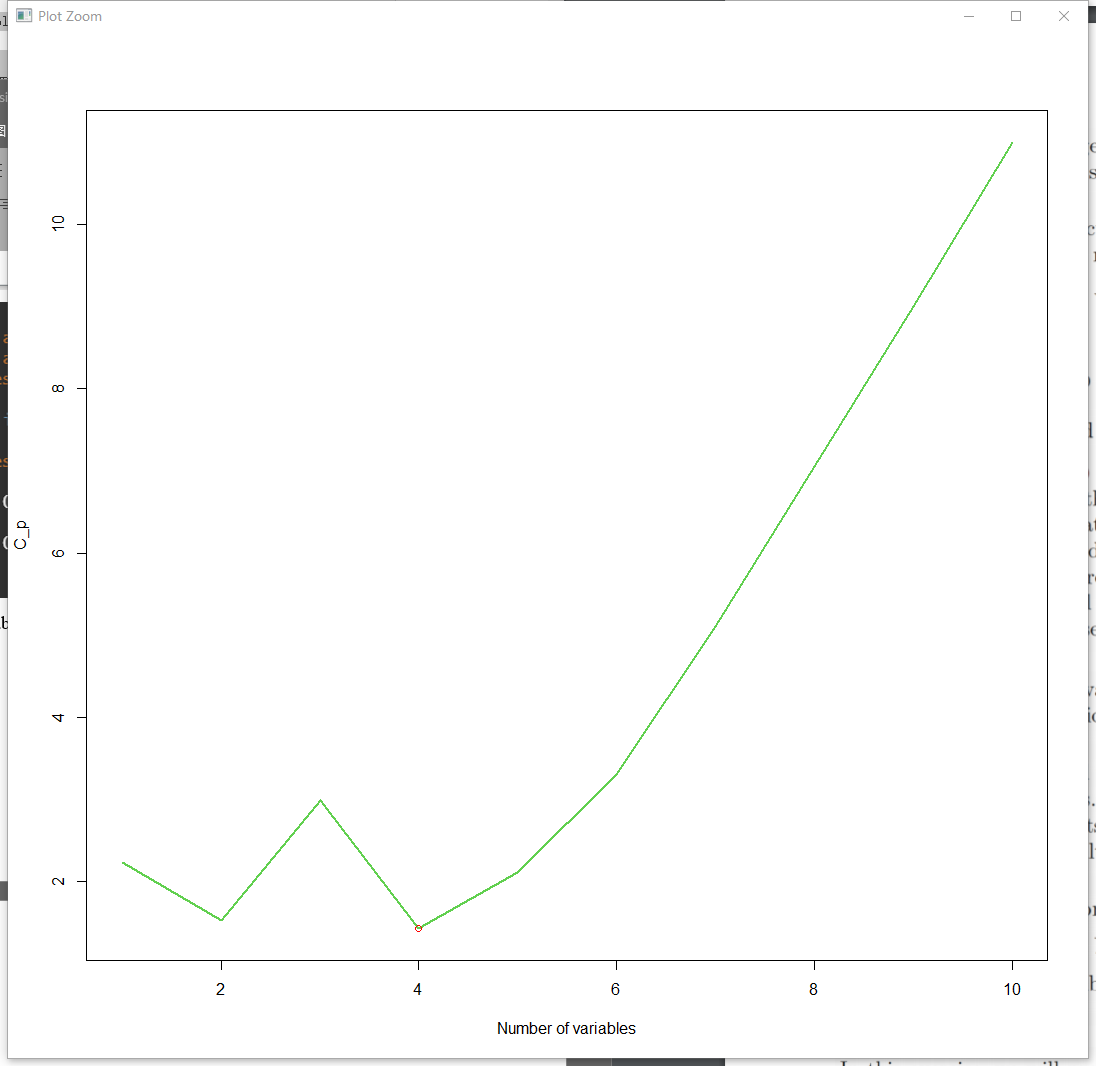
**(e)**

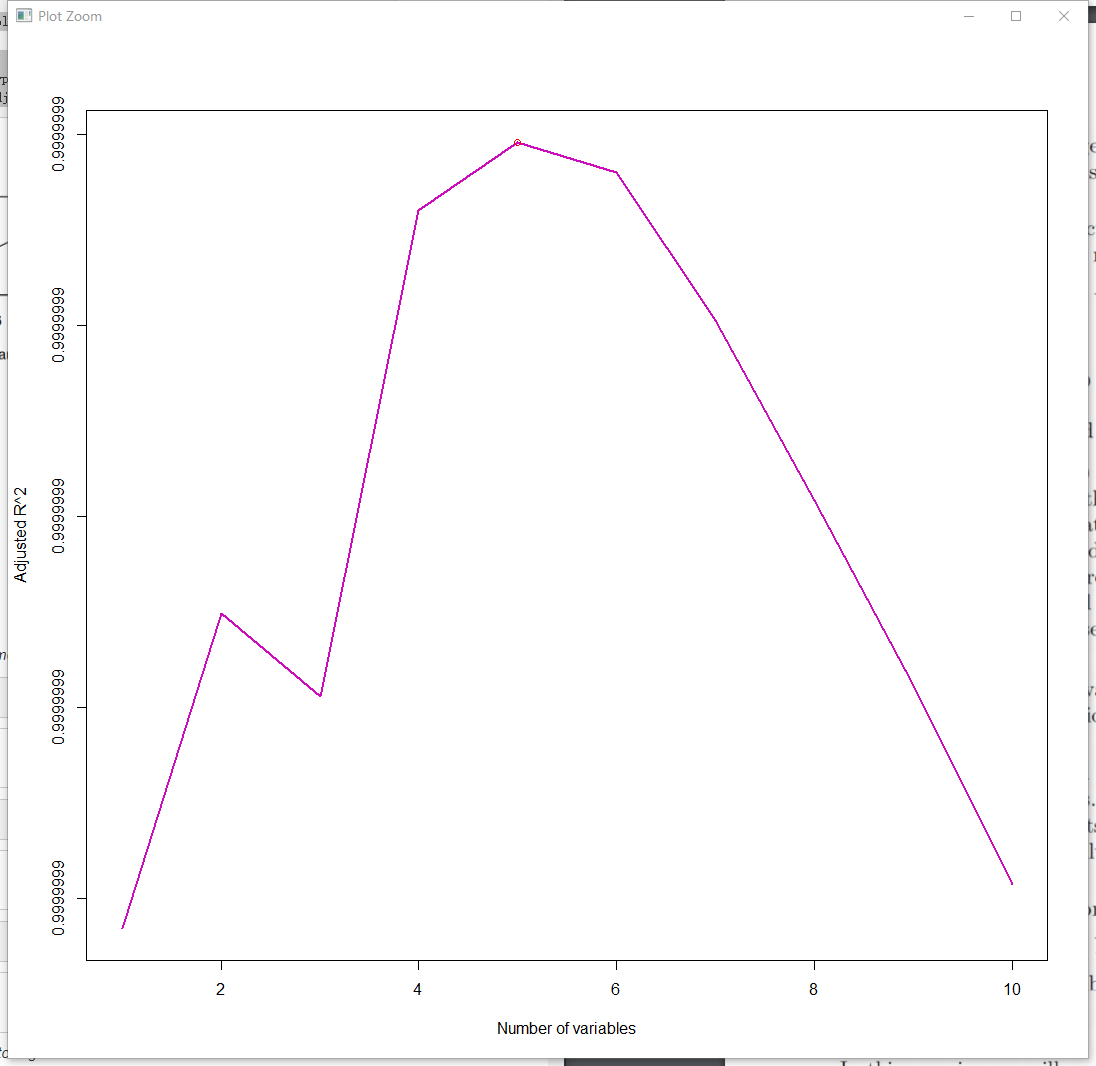
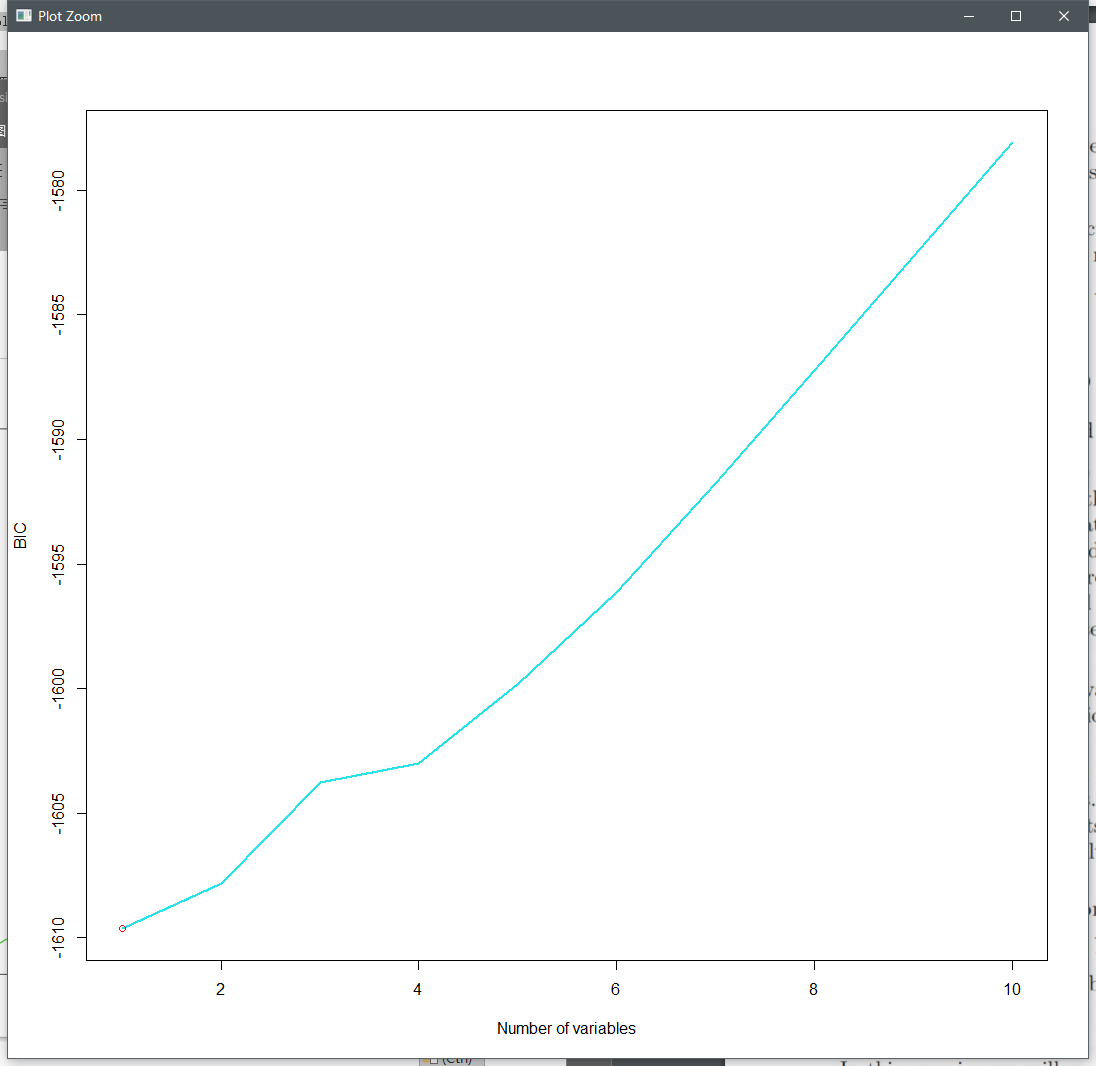


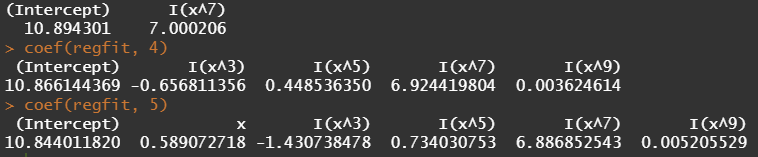


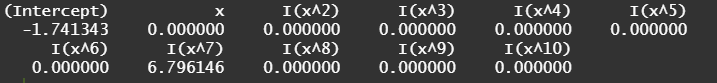
the lasso method choose X, X^3 as variables

**(f)**









BIC chooses 1 variable model and lasso also chooses 1-variable model.

**Q9**

**(b)**

Test Error

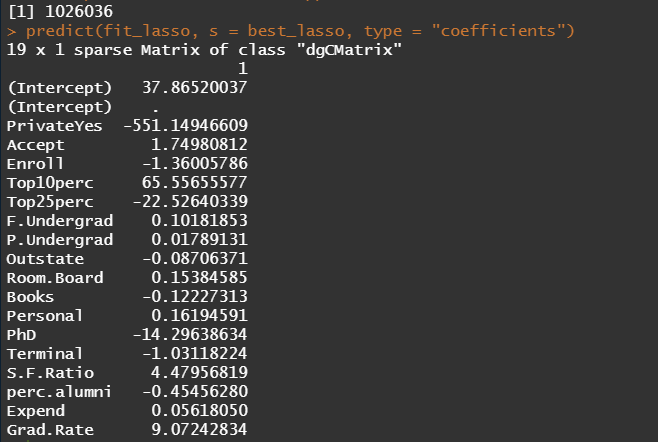


**(c)**

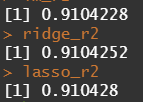
Test error



**(d)**



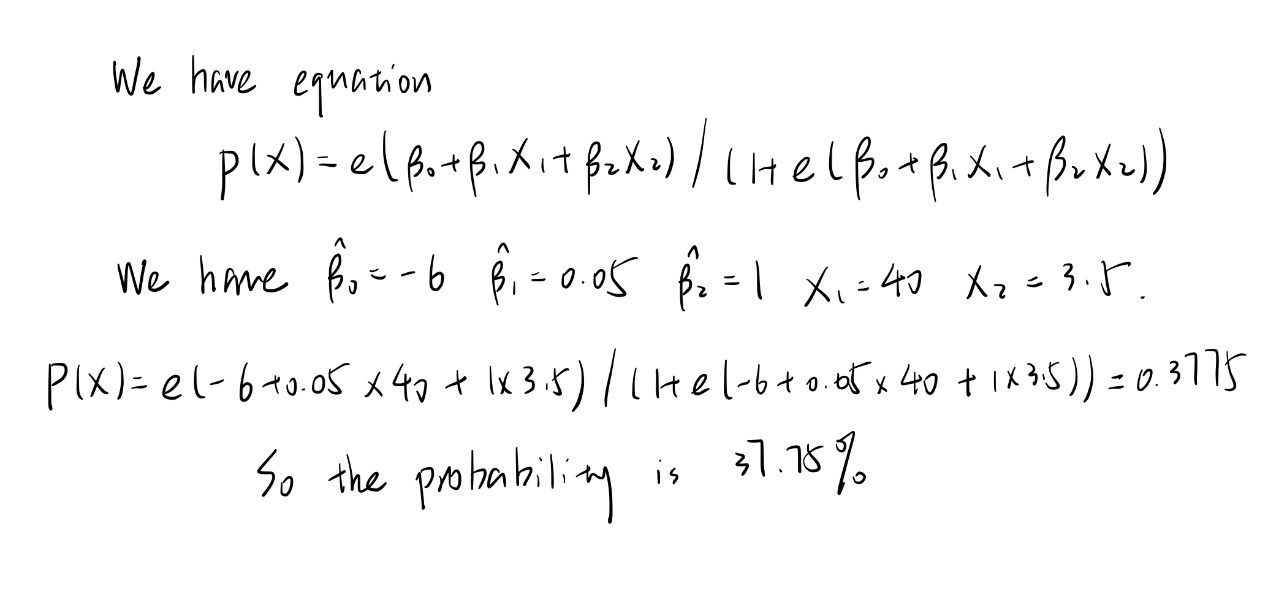
**(e)**



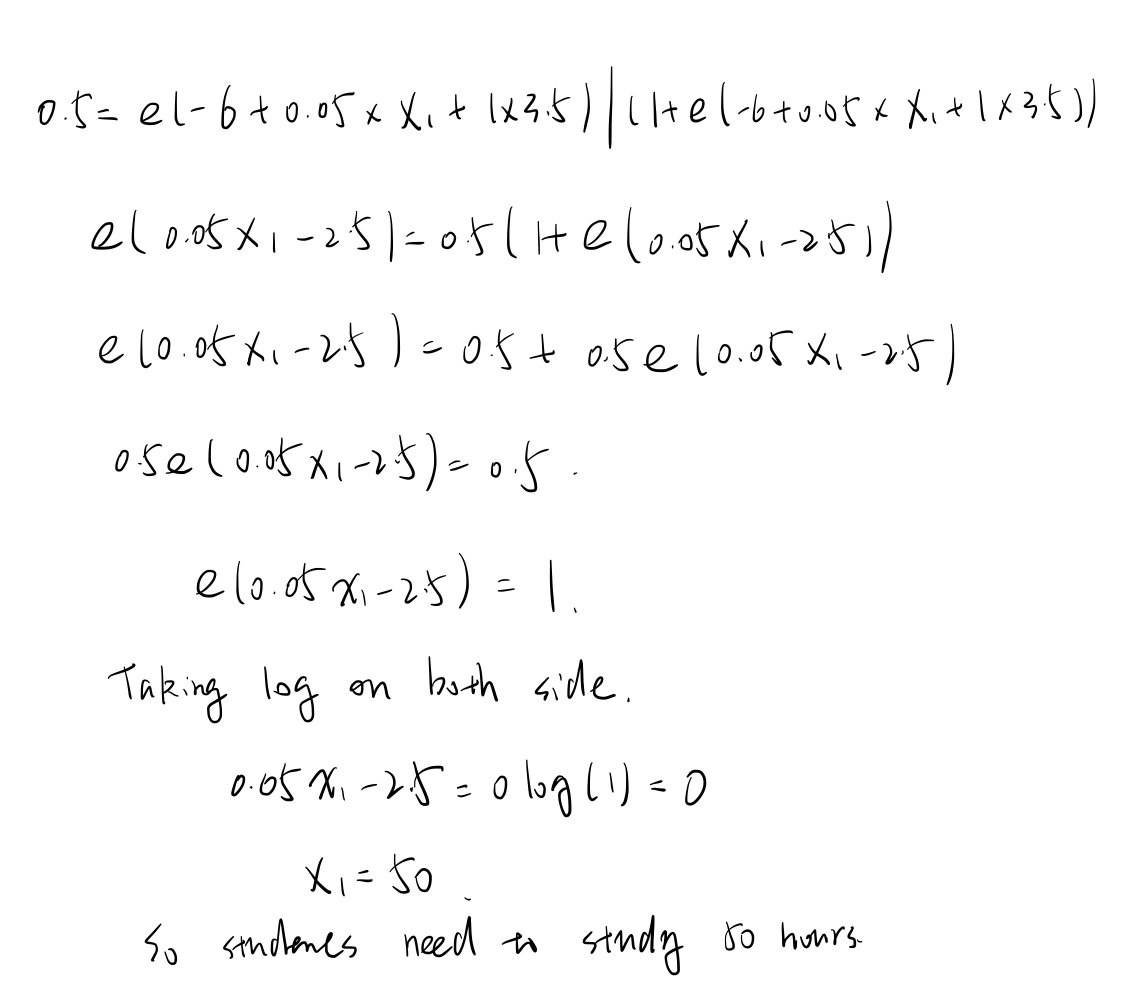
R\_square for these models are similar, all these three models can predict with high accuracy.

**Q10**

**(a)**

****

**(b)**

****