**HW 3 Chapter 5**

**Luyao Zhang (NetID: lzhang94)**

**Ex2**

1. p = 1 – 1/n

The probability that the first bootstrap observation **is** the *j*th observation from the original sample is 1/n.

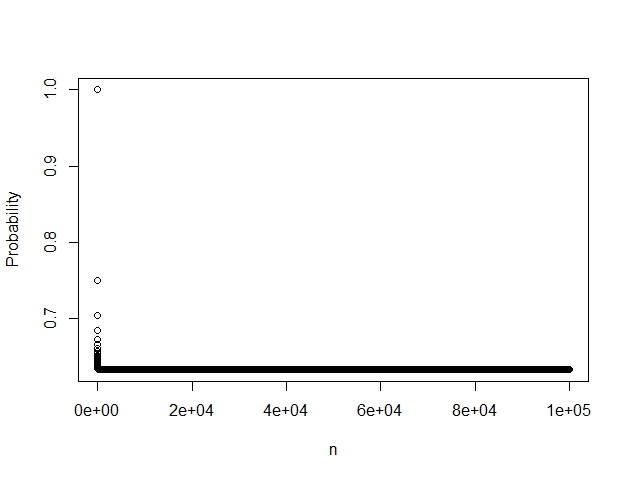
1. P = 1 - 1/n

The selection of the bootstrap sample is independent with replacement, so the probability is the same as in (a).

1. We bootstrap with replacement, and the probabilities of selecting each observation is independent. Therefore, the probability of a certain observation not being in the bootstrap sample is the product of the probability that this observation is not selected each time, which gives us:

p = (1 - 1/n) \* (1 - 1/n) \* (1 - 1/n) \* … \* (1 -1/n) = (1 - 1/n) ^n

1. p = 1- (1 - 1/5) ^ 5 = 0.672
2. p = 1- (1 – 1/100) ^ 100 = 0.634
3. p = (1 – 1/1000) ^ 1000 = 0.632
4. The plot is as below. According to the plot, it looks like that as n goes up, the probability that the *j*th observation is in the bootstrap sample drops quickly when n is small, but remains approximately the same after reaching 0.63. After 0.63, as n goes up, the probability does not change much.

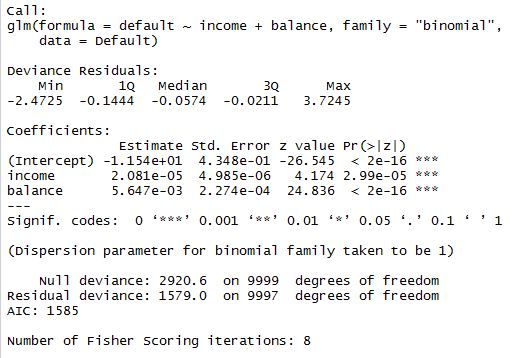


1. The result is as below. Because , when x = 1, 1 - = 1 – 1/e = 0.632.

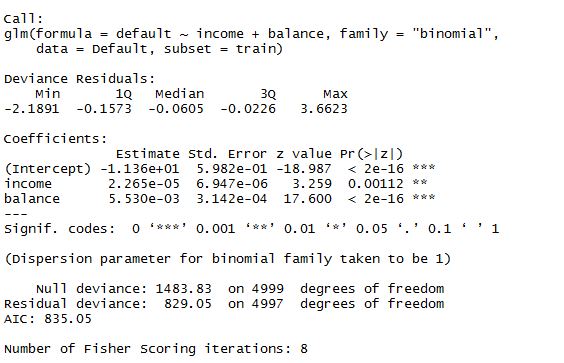


**Ex 5**

1. The summary of the logistic regression model is as below:



1. The summary of fitting a multiple logistic regression model using only the training sample is shown below:



After making the classification using the regression model obtained with the training data, and the predictions made for the other half of the dataset, the validation set error is equal to 0.0236.



1. Three different splits have given the results below. It looks like that validation estimate of the test error rate depends on the split of the data (i.e., which observations are included for training and validation, respectively).

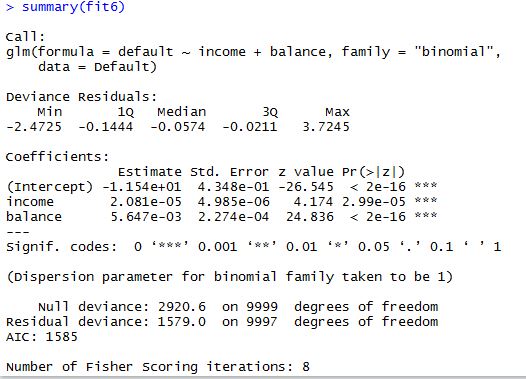


1. The results below have shown that the validation set estimate of the test error rate didn’t drop much after adding the dummy variable “student”.

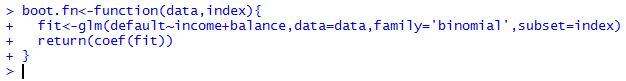


**Ex 6**

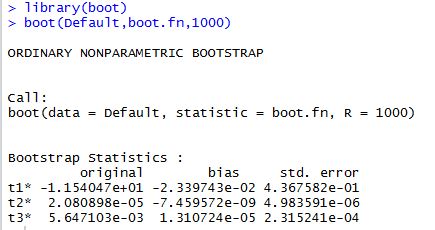
1. The summary of the model is as below. The estimated s.e. for the coefficients are respectively 0.434, 4.985\*10^(-6), and 2.274\*10^(-4).



1. The function is as below:



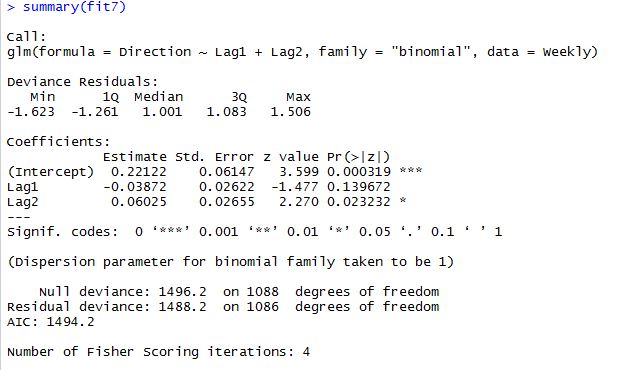
1. The results are as below. The s.e. estimates are respectively 0.437, 4.984\*10^(-6), and 2.315\*10^(-4).



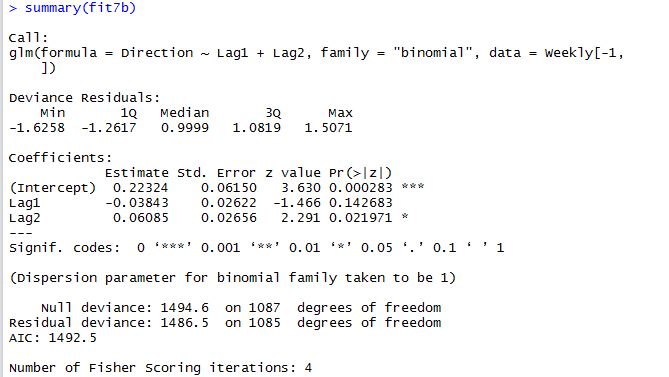
1. According to results in (c), it seems that estimates for coefficients standard errors obtained using logistic regression do not differ much from those obtained using bootstrap function.

**Ex 7**

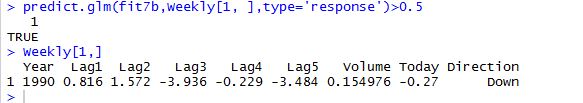
1. The summary of the model is as below:



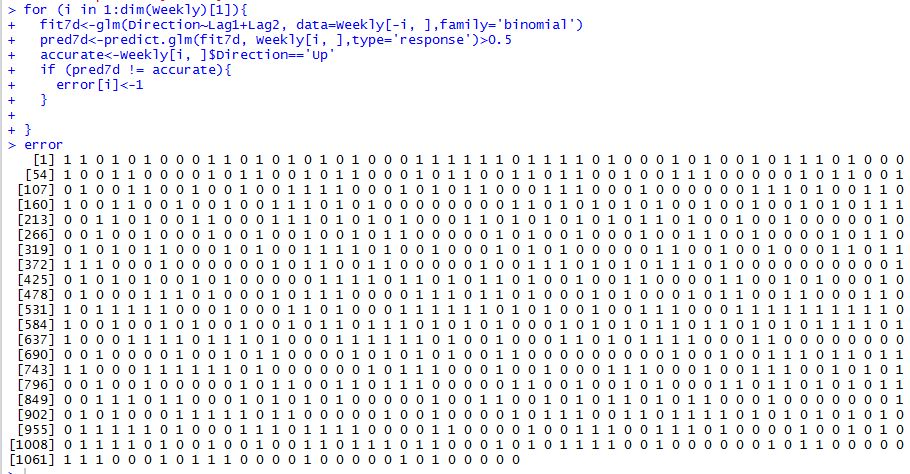
1. The summary of the model is as below:



1. Our prediction is that the first observation is “Up”, but in fact the first observation is “Down”. This one was not predicted correctly.



1. The code for the loop and the results are as below:

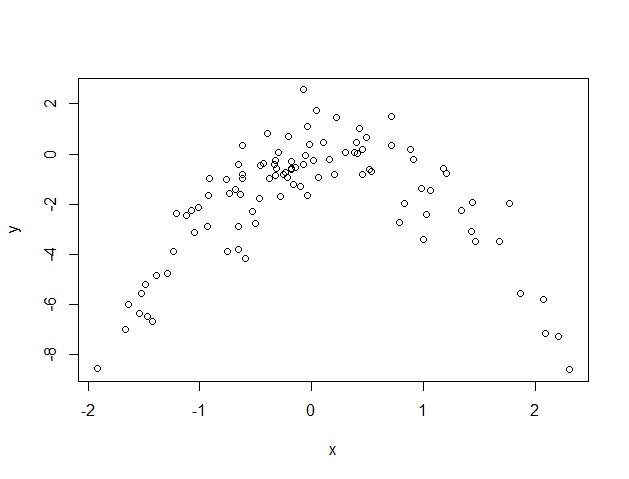


1. The LOOCV estimate for the test error is 0.45.

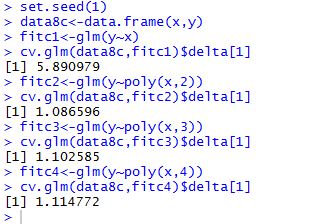


**Ex 8**

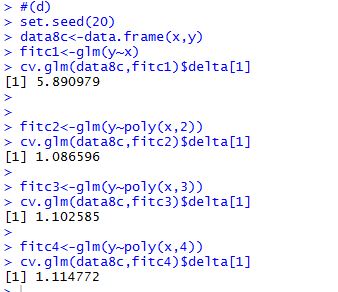
1. In this model, *n* = 100, and *p* = 2. The model used is :.
2. The plot is as below. We can observe a curvilinear relationship between x and y.



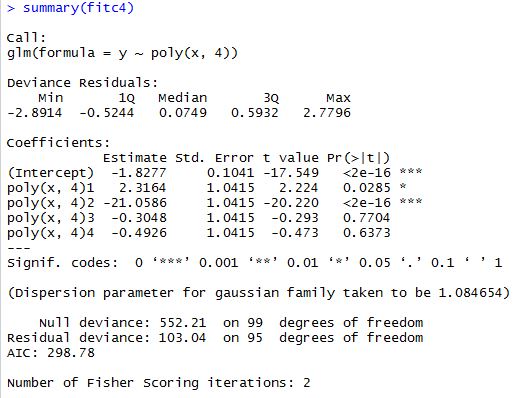
1. The LOOCV errors for the four models respectively, are 5.89, 1.09, 1.10, and 1.11.



1. The LOOCV errors for the four models respectively, are 5.89, 1.09, 1.10, and 1.11. These results are exactly the same as obtained from (c), and this is because LOOCV looks at n folds of every single observation in the dataset.



1. Model 2 in (c) has the smallest LOOCV error (1.09). This is what I expected, because in (b) we have already realized that the relationship between y and x is quadratic.
2. Based on the summary of model 4 in (c) below, we can find that x and x^2 are the only significant predictors. This is consistent with the cross-validation result that the quadratic model is the best model.



**Ex 9**

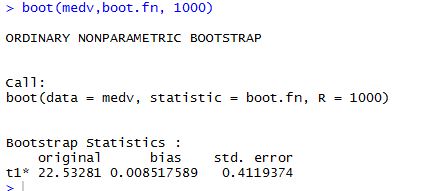
1. The estimate for the population mean equals 22.53.



1. The estimate of the standard error of the sample mean equals 0.409.



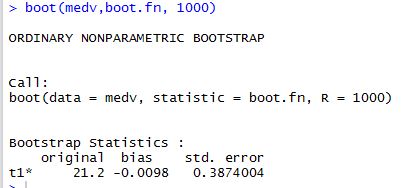
1. The estimate of the standard error of the sample mean using the bootstrap is 0.412. This is pretty similar to what we found in (b).



1. The CI is as below. It seems that the CI obtained with bootstrap is pretty similar to that obtained with t-test.
2. The estimate for the median value is 21.2.



1. The standard error of the median using bootstrap is 0.387.



1. The estimate for the tenth percentile of medv in Boston suburbs is 12.75.



1. The estimate using bootstrap is 0.511. The tenth percentile value obtained here using bootstrap is the same as what we obtained in (g), and the standard error, compared with the tenth percentile value, is relatively small.

