**1. (Ex. 4 Pg. 168)**

a) To predict 10% of x, x must be an element of [0.05, 0.95], which can be rewritten as [x-0.05, x+0.05]. However if x < 0.05, then we use observations in interval [0, x+0.05] which represents (100x+5)% and if x > 0.95, then we use observations in interval [x-0.05, 100], which represents (105-100x)%.

b) 0.975 \* 0.975 = 0.00950625% assuming the two predictors are independent

c) Assuming the predictors are independent, the fraction of available observations is 0.975100%.

d) As p predictors increase, the fraction of available observations we use for prediction approaches 0, which means nothing to use to train our data.

e) For p = 1, 2, 100, we have length = 0.1, 0.11/2, 0.11/100

**2. (Ex. 6 Pg. 170)**

**a)** Using the logistic model,

Plug in X1 = 40 hrs. and X2 = 3.75 GPA and you should get the above answer

We use this model, because p(X) = Pr(Y = 1 | X), where Y is getting an A and X = (X1,X2)

b) Given a fixed GPA of 3.5,

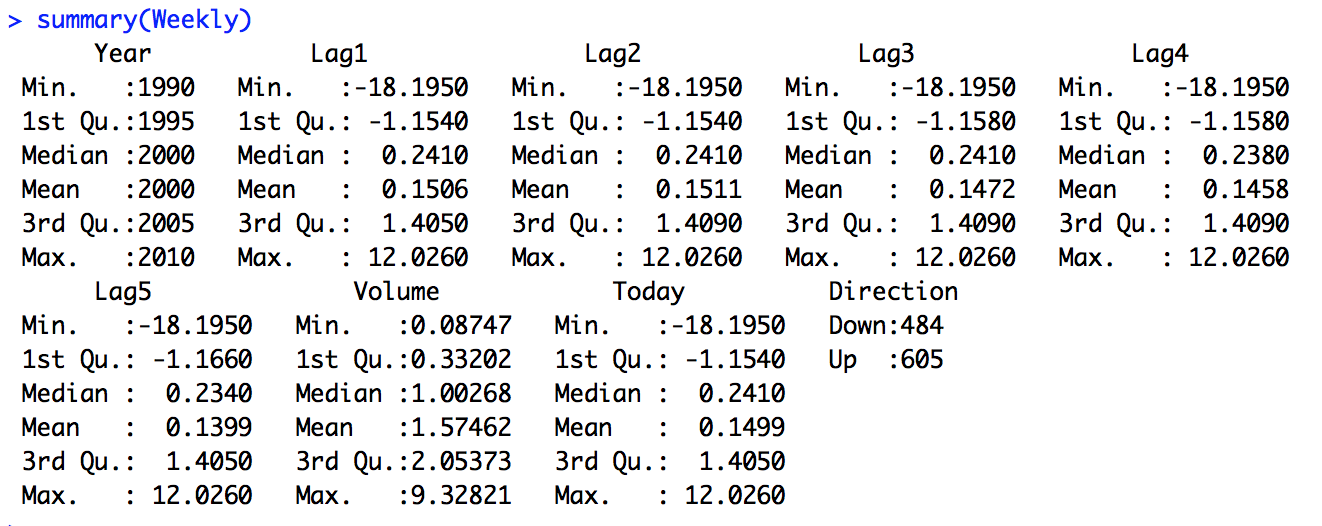
Now we solve for X1 by rewriting the equation as

**3. (Ex. 8 Pg. 170)**

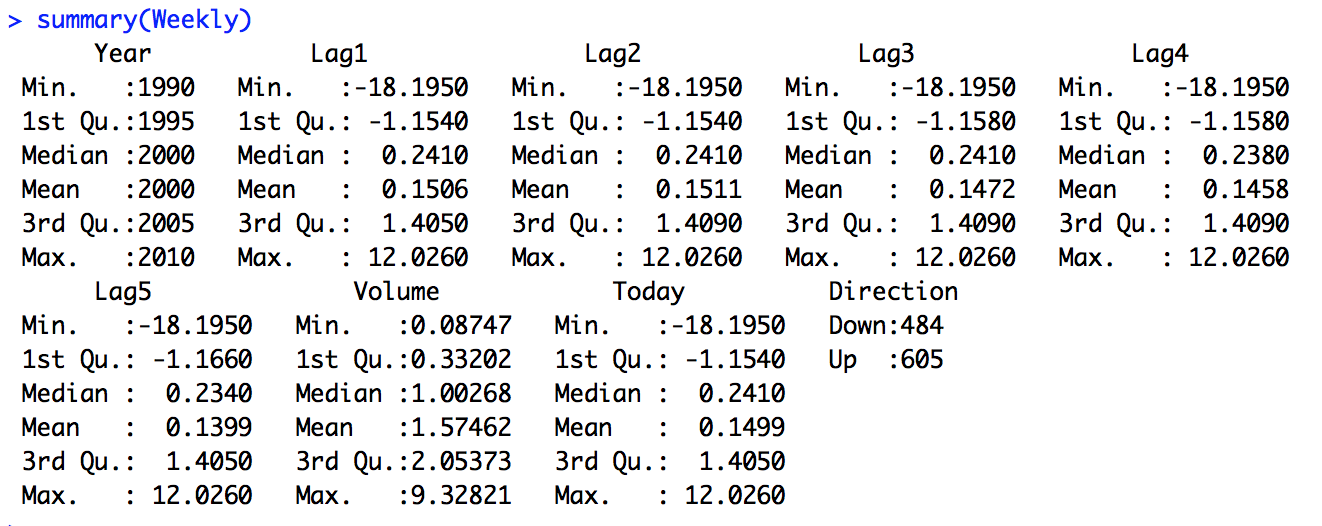
For K-nearest neighbors with K = 1, we have a training error rate of 0%, because P(Y = j | X= xi) = I(yi = j). Recall when yi = j, then I = 0, otherwise I = 1. Since K = 1, the training error rate is 0%, because of flexible classification methods. This implies our test error rate is 36% given the average of the test and training was 18%. The test error was greater than the logistic regression 30% error rate, therefore it is better to choose the logistic regression.

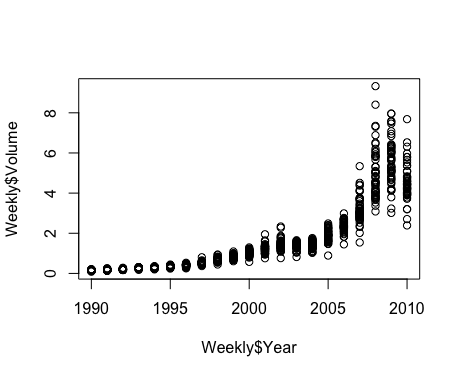
**4. (Ex. 10 Pg. 171 a, b, c, d)**

a)

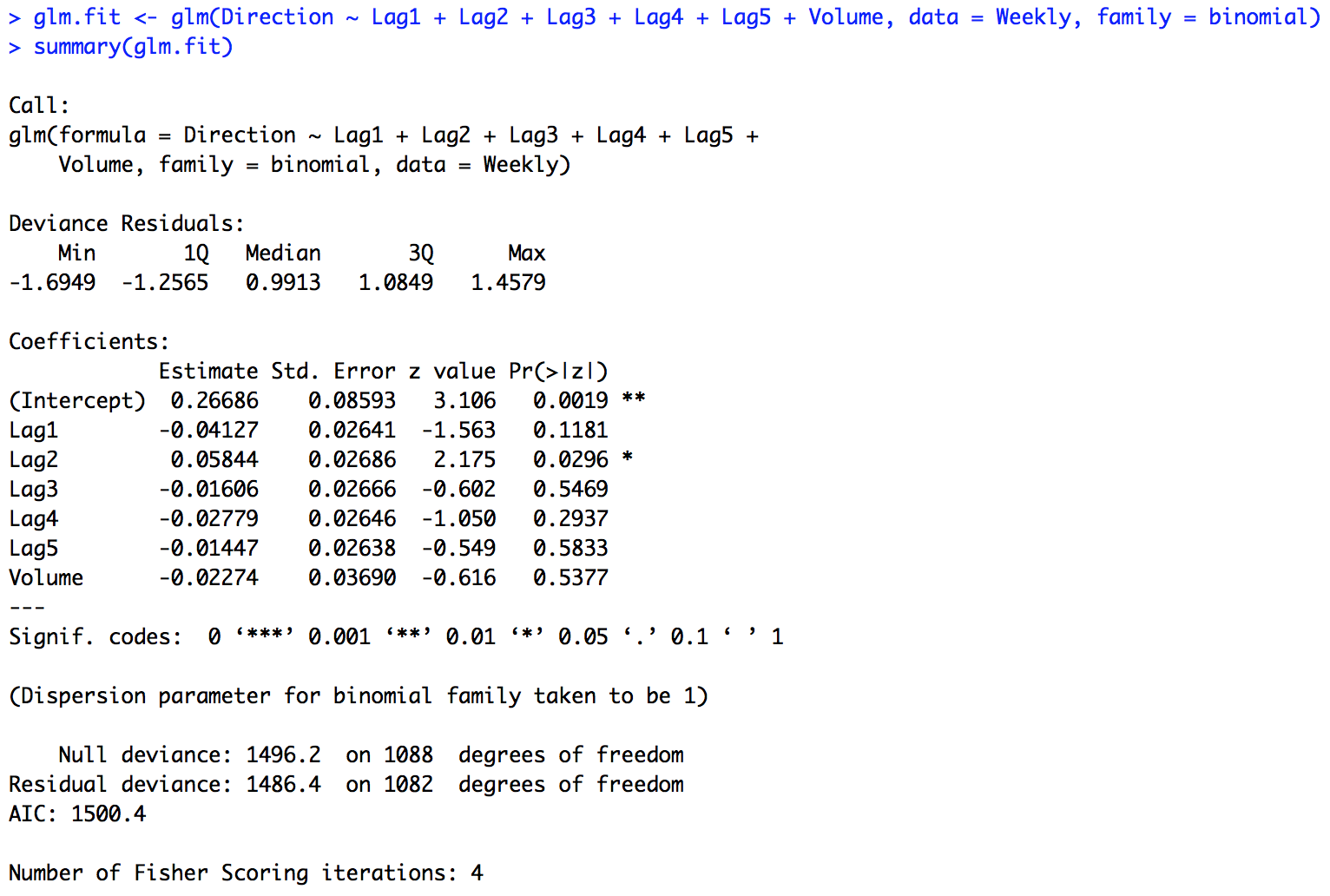


From this correlation function, we can tell year and volume have the biggest correlation, while correlations between Lags are near zero.

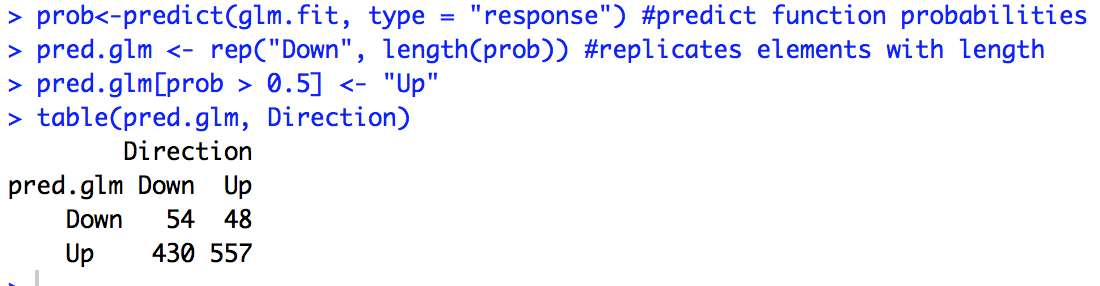


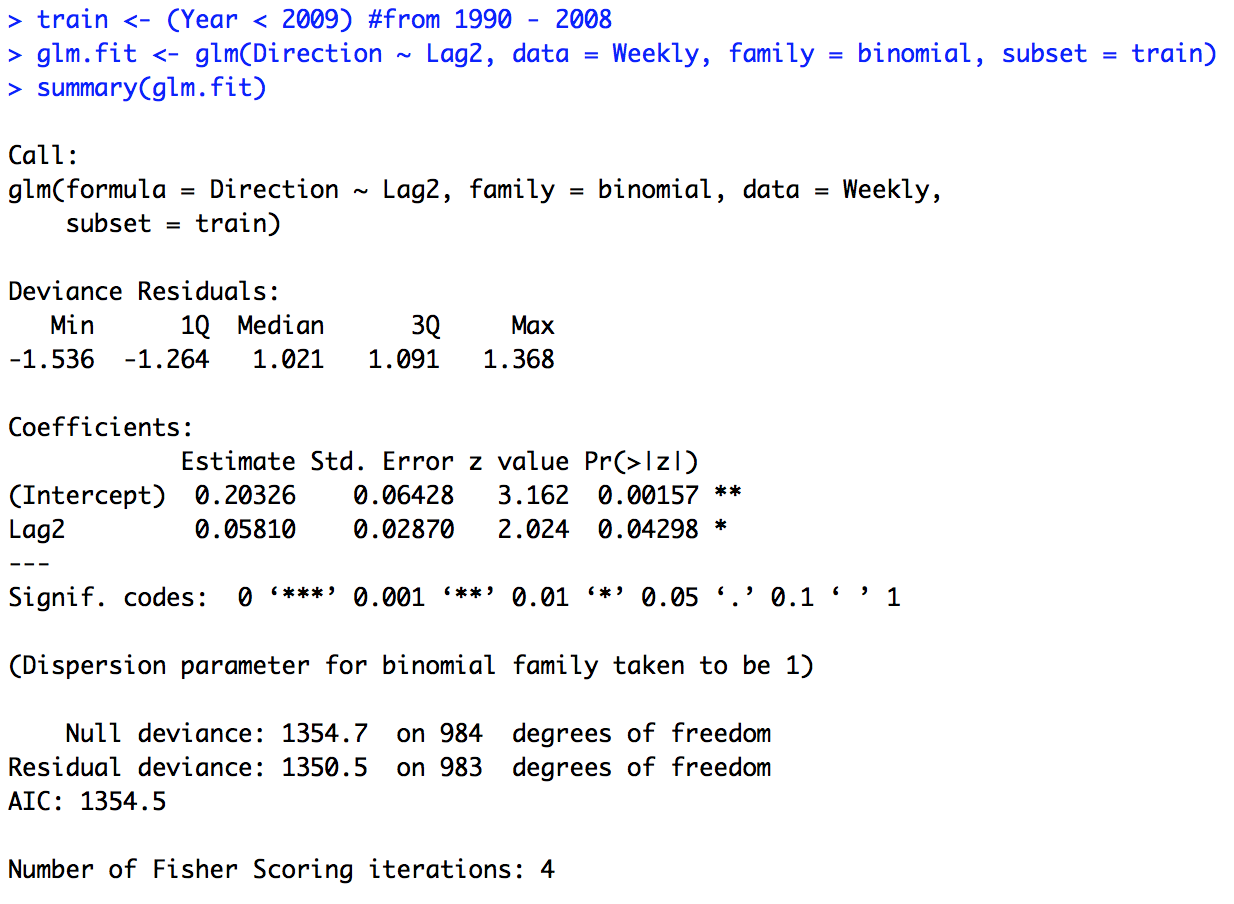


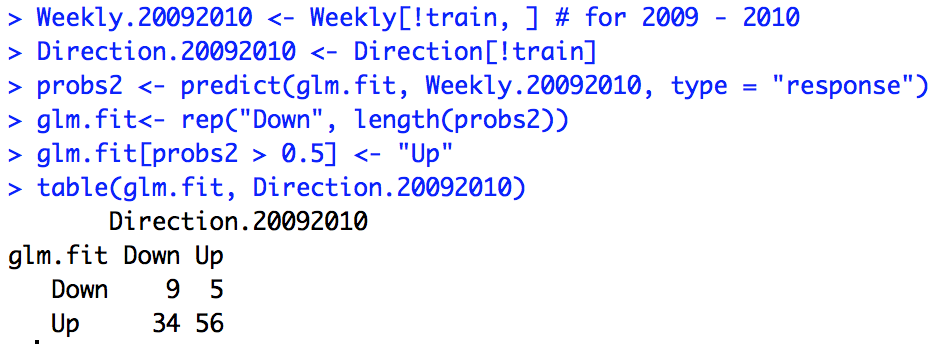
b) Must pass argument family = binomial to run a logistic regression



c) Matrix gives (54+557)/1089 = 0.5651. 1-0.5651 = 0.4389 or 43.89% training error rate. When the market goes up, the model is right 557/(557+48) = 0.921 or 92.1% of the time. When the market goes down, the model is right 54/(54+430) = 0.11157 or 11.16% of the time.



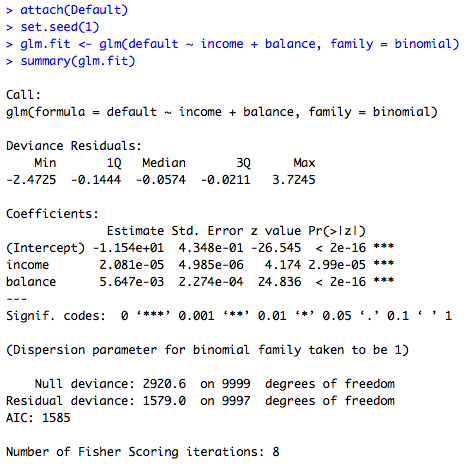
d) 



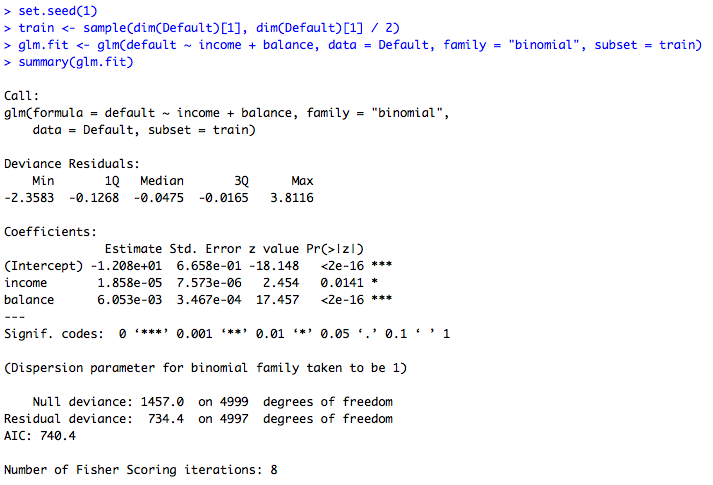
(9+56)/104 = 62.5%, then 1 – 0.625 = 37.5% training error rate. When the market goes up, the model is right 56/(56+5) = 91.8% of the time. When the market goes down, the model is right 9/(34+9) = 20.93% of the time.

**5. (Ex. 5 Pg. 198)**

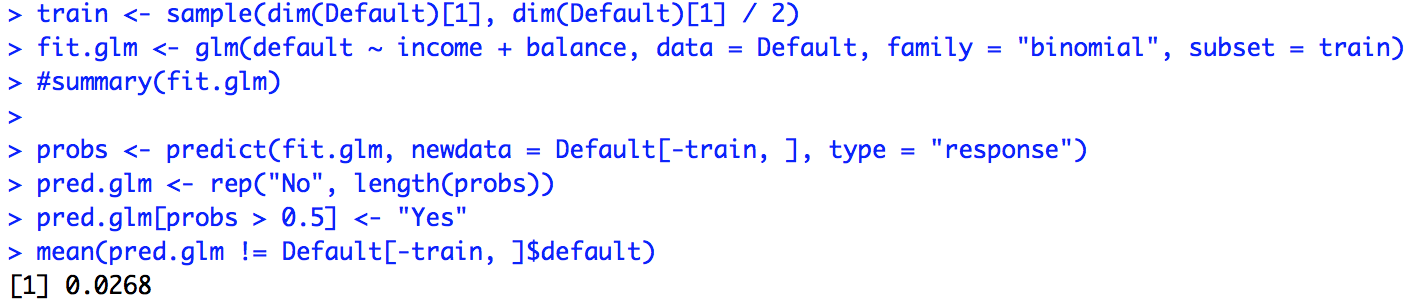
a)



b) i. & ii. Split data and fit a model with only training data

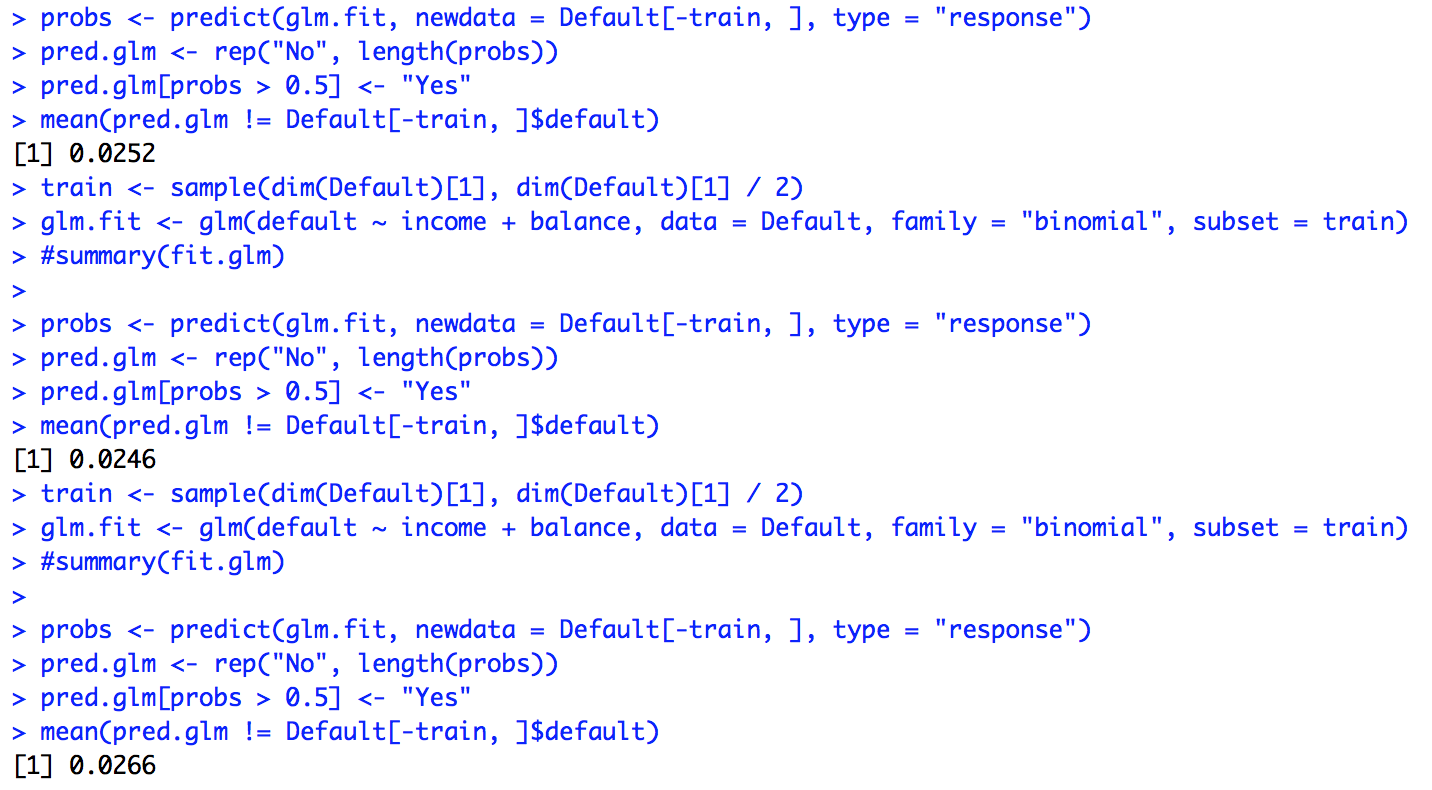


iii. & iv.



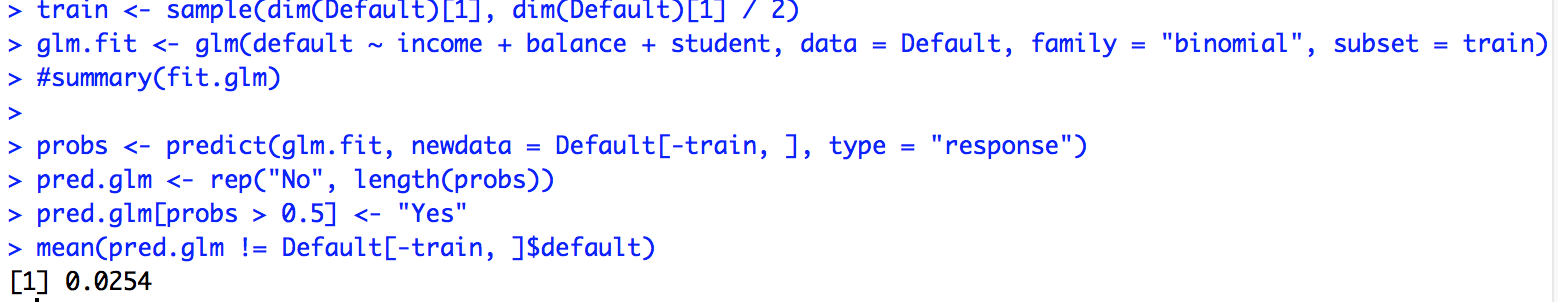
2.68% test error rate with validation set

c)



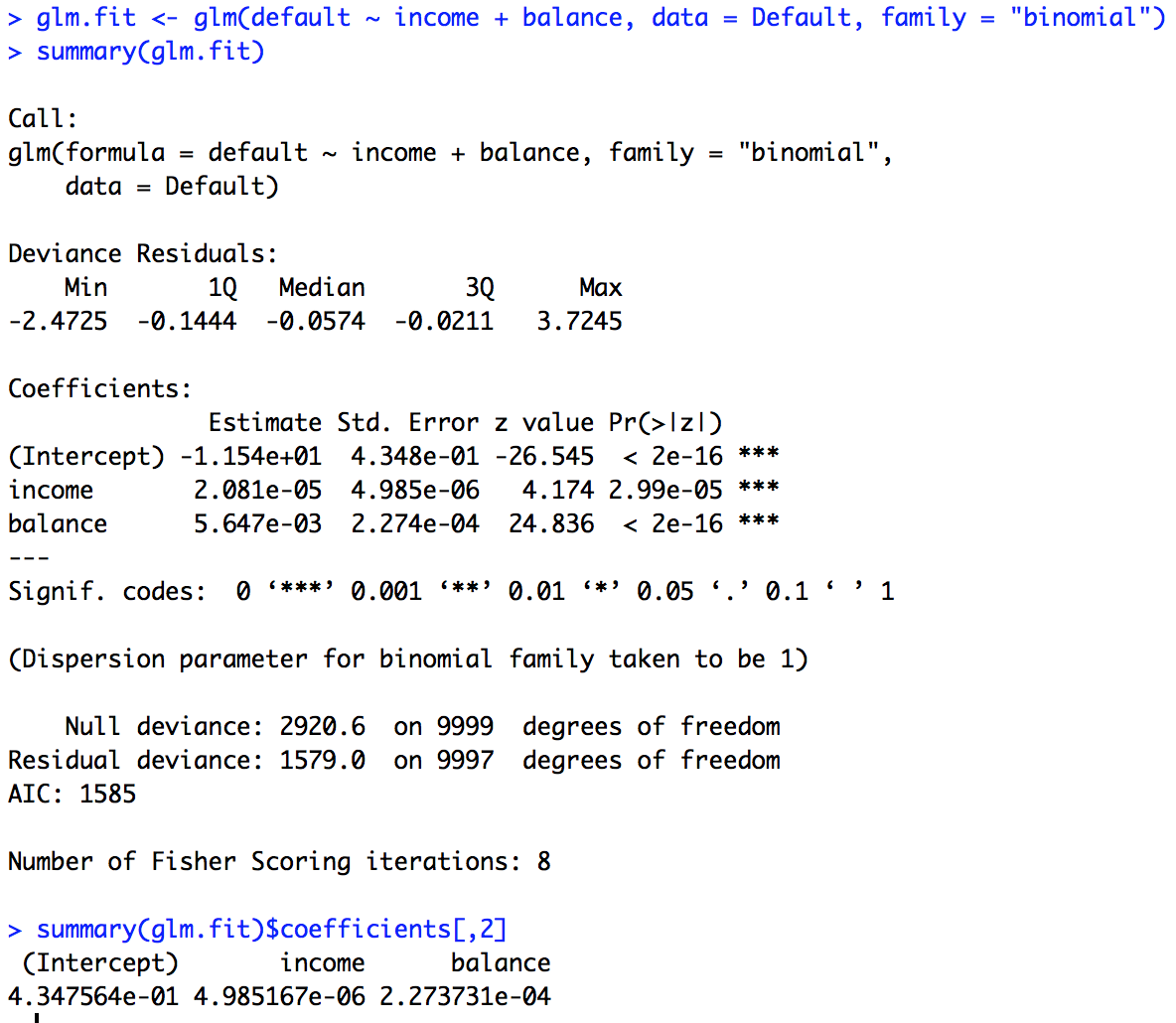
Validation set’s test error rate varies depending on what observations are in the training set and what observations in the validation set.

d) Dummy variable student does not seem to affect the reduction of test error rate on the validation set.



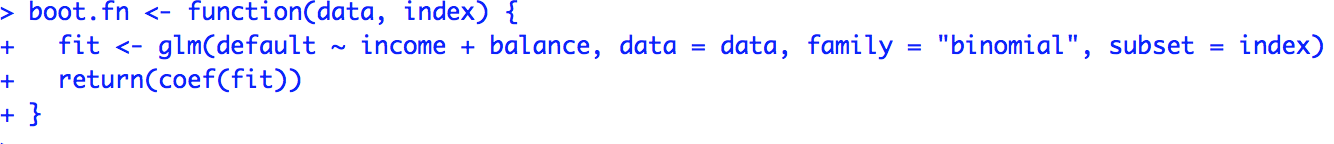
**6. (Ex. 6 Pg. 199)**

a)

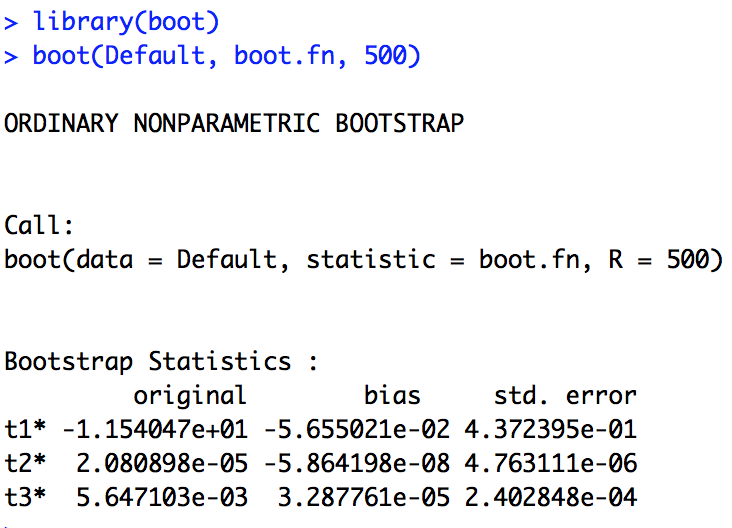


The coefficients for B0, B1, B2 are 0.4347564, 4.985167\*10-6, and 2.2733731\*10-4

b)



c)

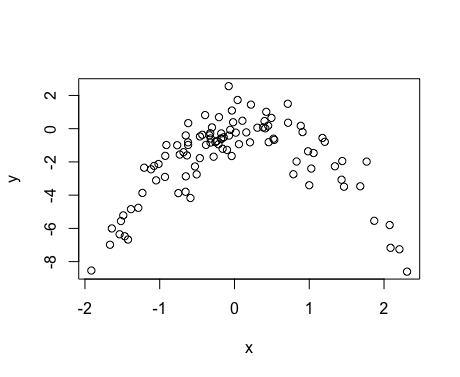


d) Notice from c) that the standard error of for B0, B1, B2 are 0.437239, 4.763111\*10-6 and 2.40284\*10-4 respectively, which is close to the standard error from b) where the coefficients 0.4347564, 4.985167\*10-6, and 2.2733731\*10-4. Therefore we can conclude both estimation methods are pretty close.

**7. (Ex. 8 Pg. 200)**

a) In the data set, n = 100 observations, and p = 2 predictors

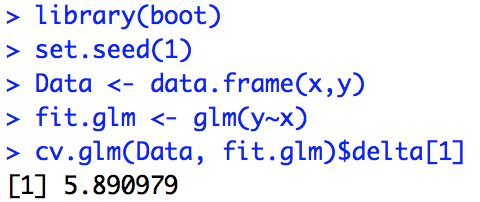
b)



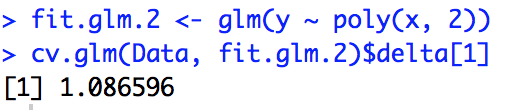
There is a curved relationship between x and y.

c)

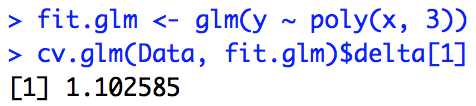
i.



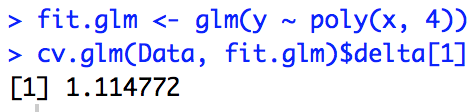
ii.



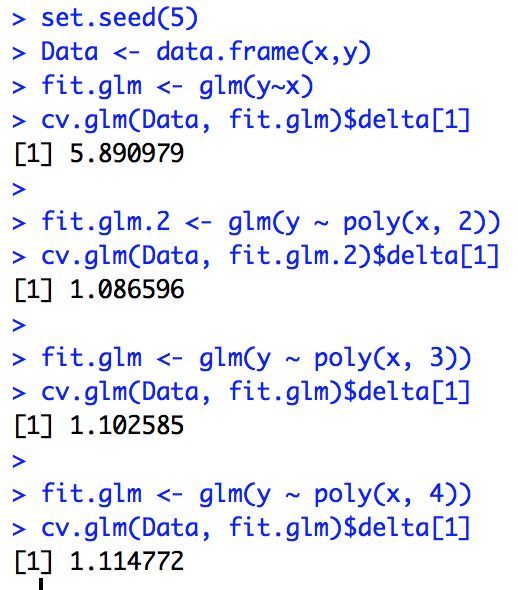
iii.



iv.

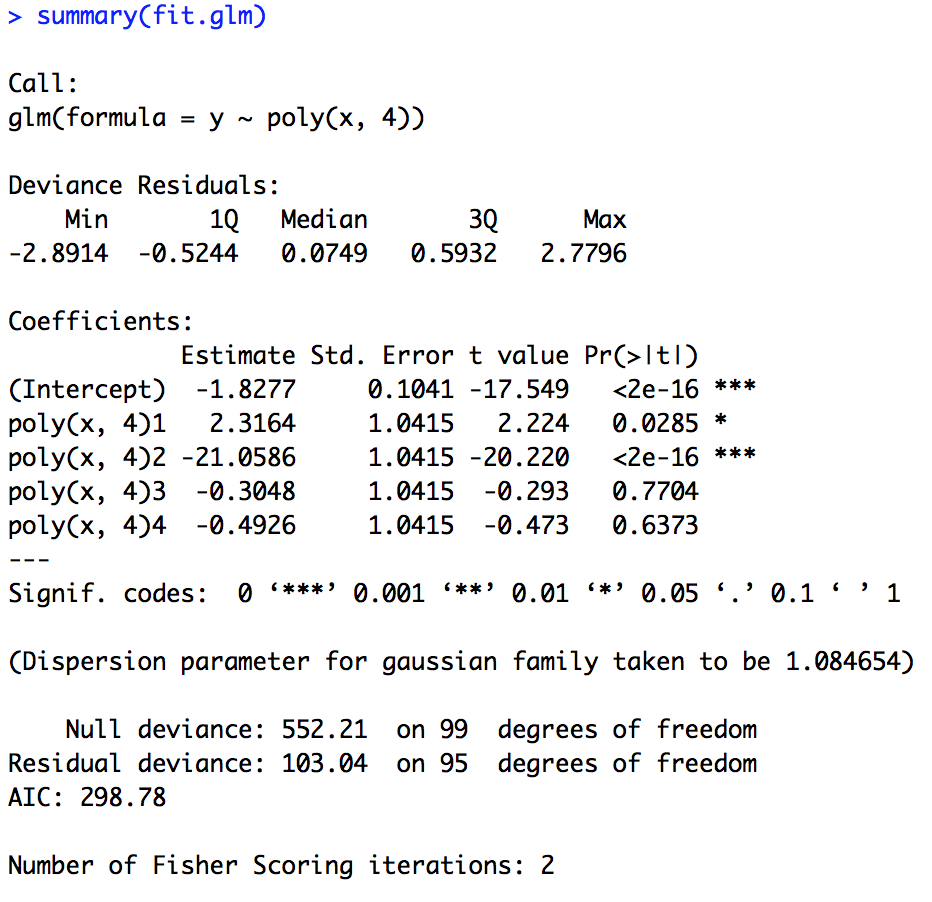


d) Yes, the results are identical, because leave one out cross validation evaluates n folds of a single observation



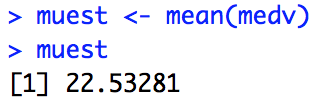
e) The second fit had the smallest error, because the relationship between x and y is quadratic as we plotted in b.

f) P-values of linear and quadratic were more significant compared to cubic and quartic, which matches our cross-validation results

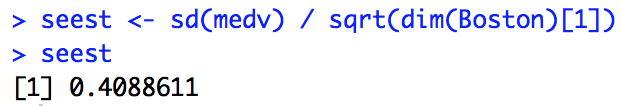


**8. (Ex. 9 Pg. 200)**

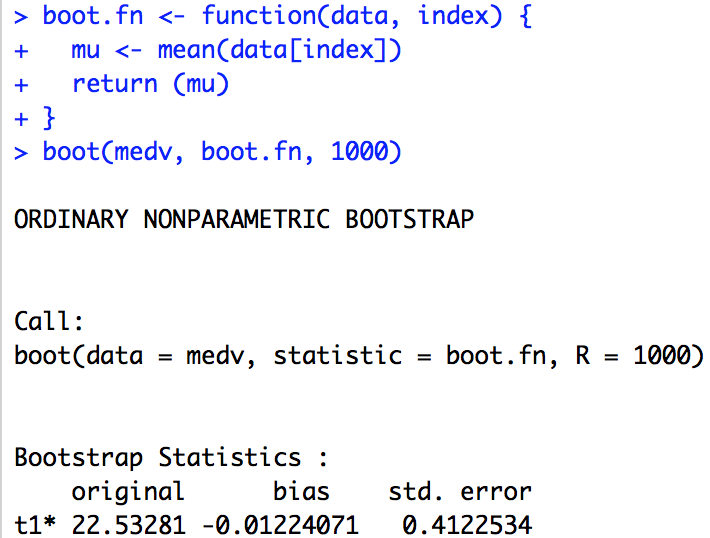
a)



b)

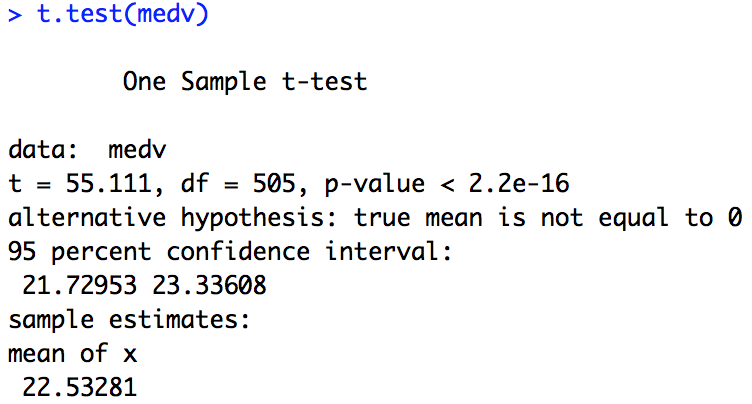


c)

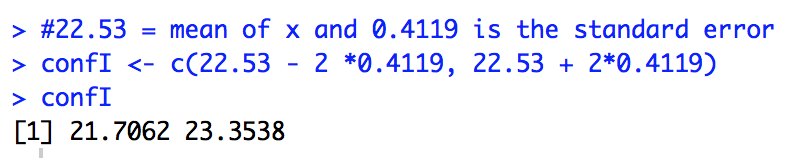


Bootstrap estimated standard error is 0.4122534, which is very close to b

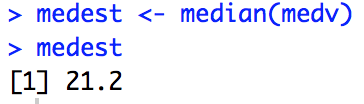
d)



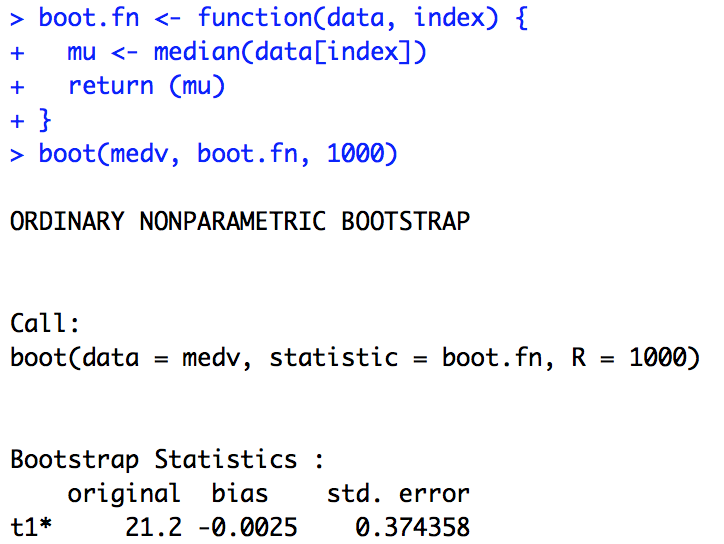
T-test 95% confidence interval: [21.73, 23.33] and bootstrap interval: [21.71, 23.35] are very close.



e)

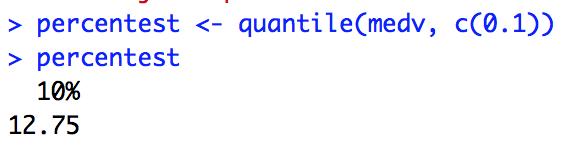


f)



Here we see an estimated 21.2 median value, which is equivalent to the value from e. Also with standard error 0.374 which is small relative to our value.

g)



h) Similar to h, here we see an estimated tenth percentile value of 12.75, which matches our result from g with standard error approximately 0.5, but small enough to not affect our value.

