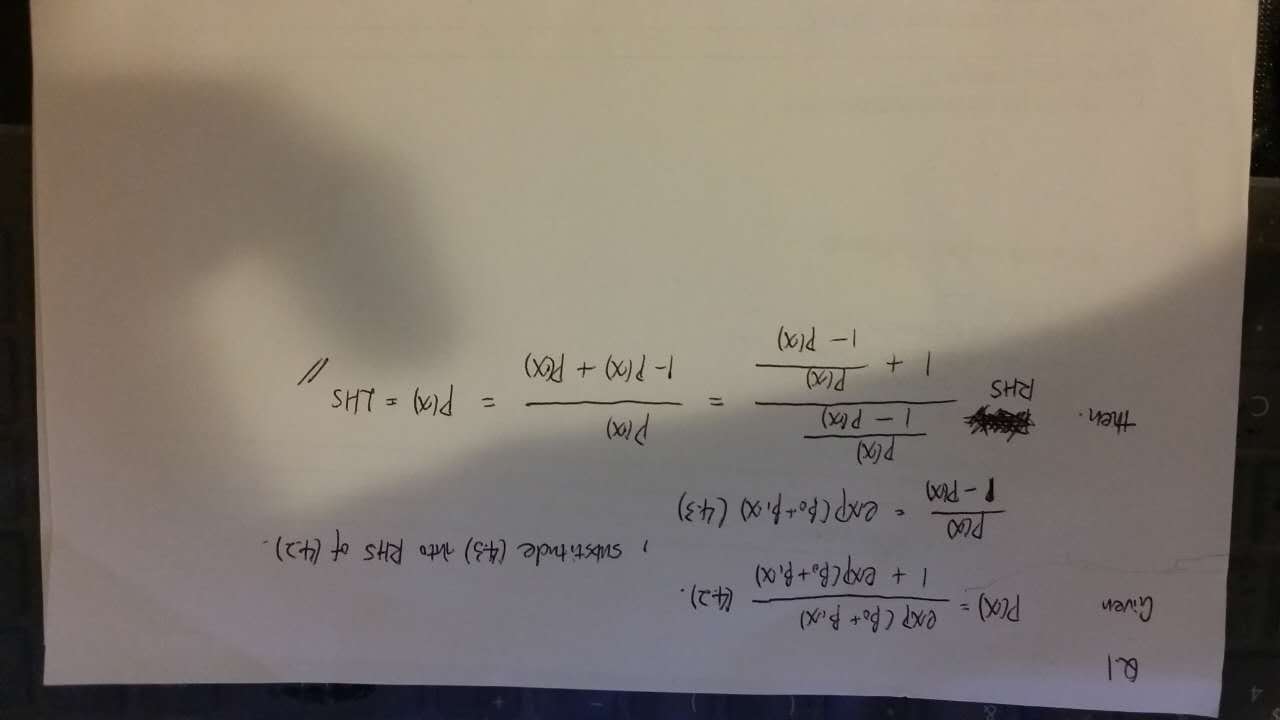
**Homework 2**

Zili Ma, Xiao-Qun Wang

**Q1. Using a little bit of algebra, prove that (4.2) is equivalent to (4.3). In other words, the logistic function representation and logit representation for the logistic regression model are equivalent.**

**A1**.

**Q2. We now examine the differences between LDA and QDA**

1. If the Bayes decision boundary is linear, do we expect LDA or QDA to perform better on the training set? On the test set?

**A2a:** LDA is better in the test set, QDA is better in the training set but less good in the test data. For the linear boundary, LDA can provide an effective classification with both low variance and bias of model, while QDA may introduce more variance and overfit the training set.

1. If the Bayes decision boundary is non-linear, do we expect LDA or QDA to perform better on the training set? On the test set?

**A2b:** QDA is better on both training and test sets than LDA.

1. In general, as the sample size *n* increases, do we expect the test prediction accuracy of QDA relative to LDA to improve, decline, or be unchanged? Why?

**A2c:** Generally, QDA will be better when the training set is very large, then the variance of the classifier will not be a major concern.

1. True or False: Even if the Bayes decision boundary for a given problem is linear, we will probably achieve a superior test error rate using QDA rather than LDA because QDA is flexible enough to model a linear decision boundary. Justify your answer.

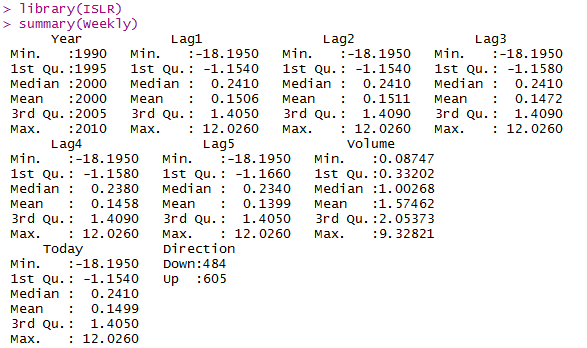
**A2d:** False. QDA will introduce a larger variance of the model for the test data and increase the test error, if overfitting occurs.

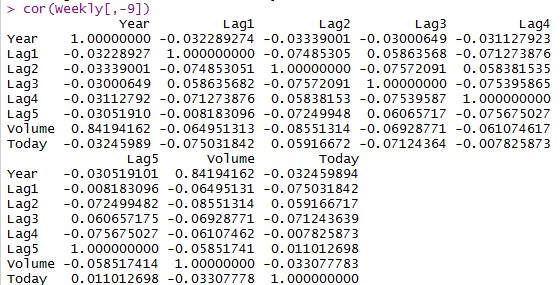
**Q3.** Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1-nearest neighbors (i.e. *K* = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

**A3.** In case of 1-nearest neighbor classifier, the train error should equal to 0, because of the training observation has been classified already and posteriorly.Therefore, the test error rate generated by the 1NN method is actually equal to 36%, which is higher than the former method. So, we should choose logistic regression instead of 1NN.

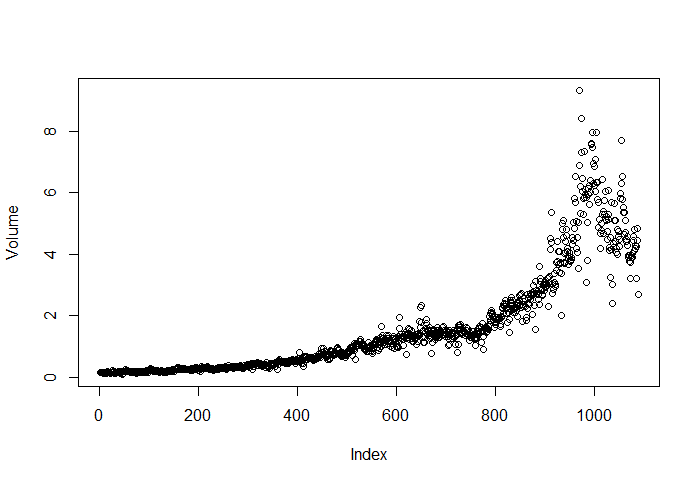
**Q4**. This question should be answered using the *Weekly* data set, which is part of the *ISLR* package. This data is similar in nature to the *Smarket* data from this chapter’s lab, except that it contains 1*,* 089 weekly returns for 21 years, from the beginning of 1990 to the end of 2010.

1. Produce some numerical and graphical summaries of the Weekly data. Do there appear to be any patterns?



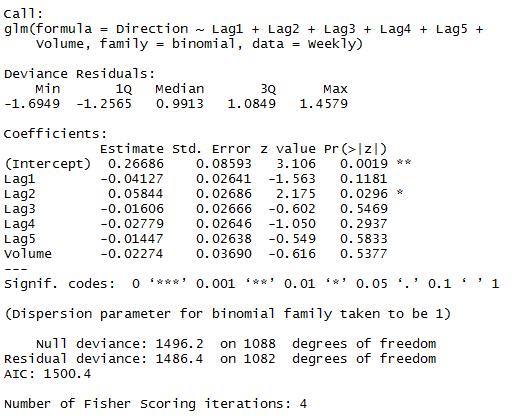






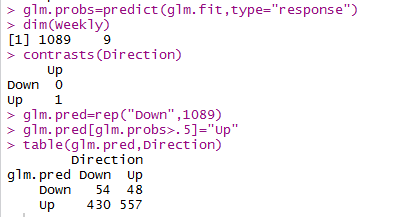
1. Use the full data set to perform a logistic regression with Direction as the response and the five lag variables plus Volume as predictors. Use the summary function to print the results. Do any of the predictors appear to be statistically significant? If so, which ones?





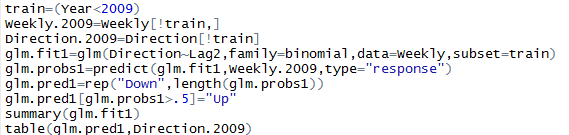
**A4b:** As shown in above result, only ***Lag2*** is shown as a significant predictor with p-value smaller than 0.05.

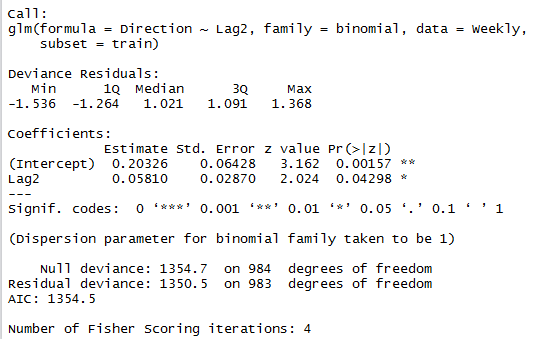
1. Compute the confusion matrix and overall fraction of correct predictions. Explain what the confusion matrix is telling you about the types of mistakes made by logistic regression.

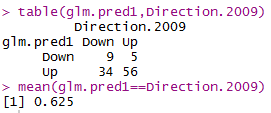


**A4c:** As shown in above result, the error rate of the training data is (430+48)/1089\*100% which equals to 43.9%, The percentage of correct prediction is 1-43.9% equaling to 56.1%. For weeks when the market goes up, the probability for the model predicting the right direction is 557/(557+48) equaling to 92%. For weeks when the market goes down, the probability for the model predicting the right direction is 55/(54+430) equaling to 11.6%.

1. Now fit the logistic regression model using a training data period from 1990 to 2008, with Lag2 as the only predictor. Compute the confusion matrix and the overall fraction of correct predictions for the held out data (that is, the data from 2009 and 2010).



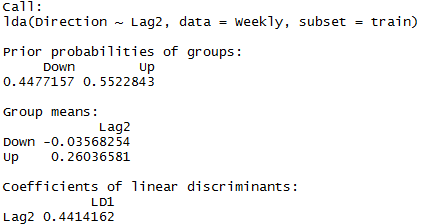




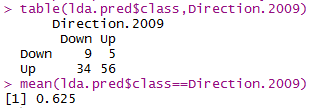
**A4d:** The overall fraction of correct predictions for the held out data is (9+56)/(9+5+34+56) equaling to 62.5%.

1. Repeat (d) using LDA.





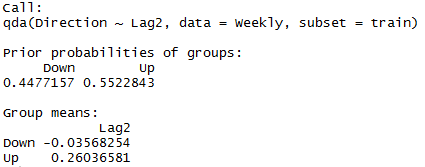




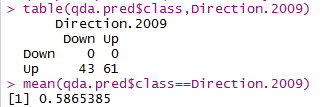
**A4e:** The overall fraction of correct predictions for the held out data is (9+56)/(9+5+34+56) equaling to 62.5%.

1. Repeat (d) using QDA.



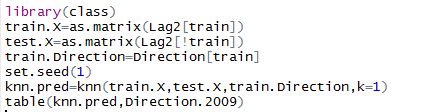


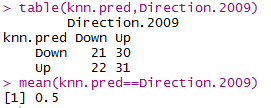




**A4f:** The overall fraction of correct predictions for the held out data is equaling to 58.6%. We may note, that QDA achieves a correctness of 58.7% even though the model chooses “Up” the whole time.

1. Repeat (d) using KNN with *K* = 1.





**A4g:** In this case, we may conclude that the percentage of correct predictions on the test data is 50%. In other words 50% is the test error rate. We could also say that for weeks when the market goes up, the model is right 50.8196721% of the time. For weeks when the market goes down, the model is right only 48.8372093% of the time.

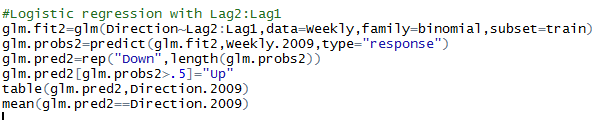
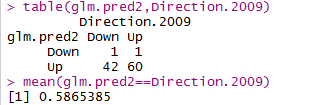
1. Which of these methods appears to provide the best results on this data?

**A4h:** If we compare the test error rates, we see that logistic regression and LDA have the minimum error rates, followed by QDA and KNN.

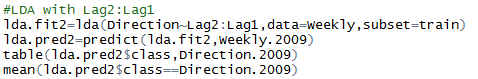
1. Experiment with different combinations of predictors, including possible transformations and interactions, for each of the methods. Report the variables, method, and associated confusion matrix that appears to provide the best results on the held out data. Note that you should also experiment with values for *K* in the KNN classifier.

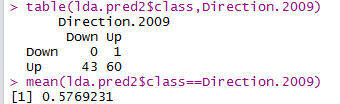
**A4i:**

1. Logistic regression with Lag2 interaction with Lag1.

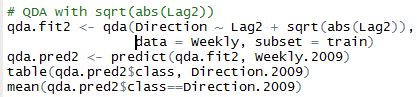
 

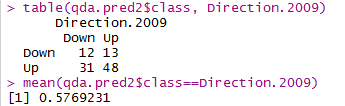
1. LDA with Lag2 interaction with Lag1



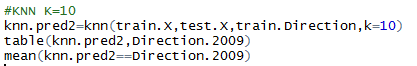


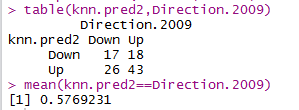
1. QDA with square root of the absolute value of Lag2



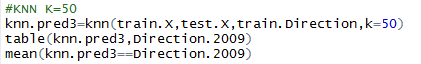


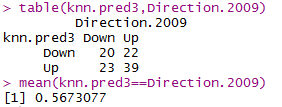
1. KNN with K=10





1. KNN with K=50

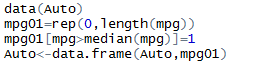




From above results of combinations, the original logistic regression and LDA have the best performance in terms of test error rates.

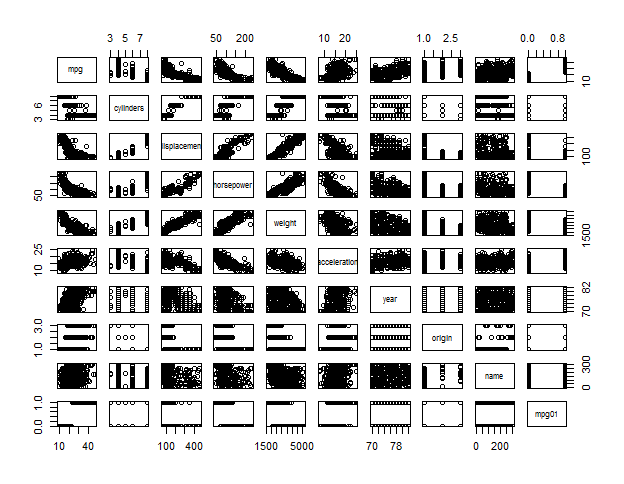
**Q5.** In this problem, you will develop a model to predict whether a given car gets high or low gas mileage based on the Auto data set.

1. Create a binary variable, mpg01, that contains a 1 if mpg contains a value above its median, and a 0 if mpg contains a value below its median. You can compute the median using the median() function. Note you may find it helpful to use the data.frame() function to create a single data set containing both mpg01 and the other Auto variables.

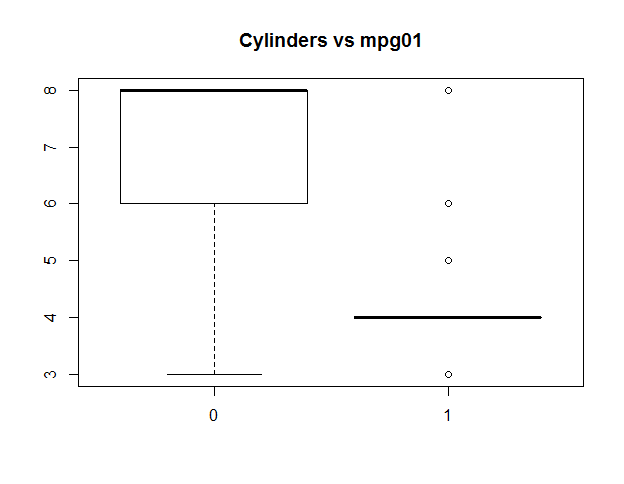


1. Explore the data graphically in order to investigate the association between mpg01 and the other features. Which of the other features seem most likely to be useful in predicting mpg01? Scatterplots and boxplots may be useful tools to answer this question. Describe your findings.

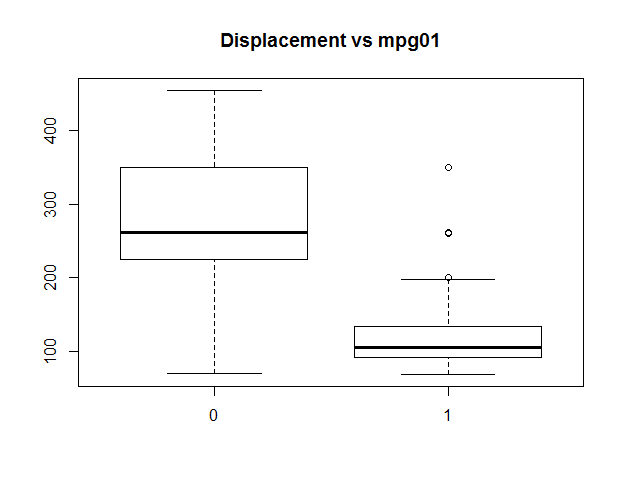




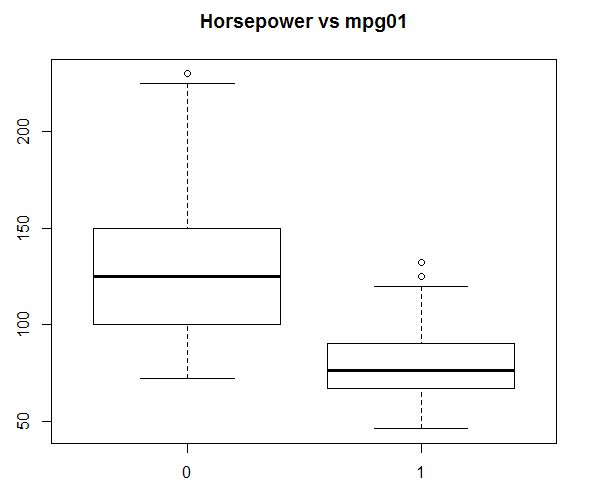




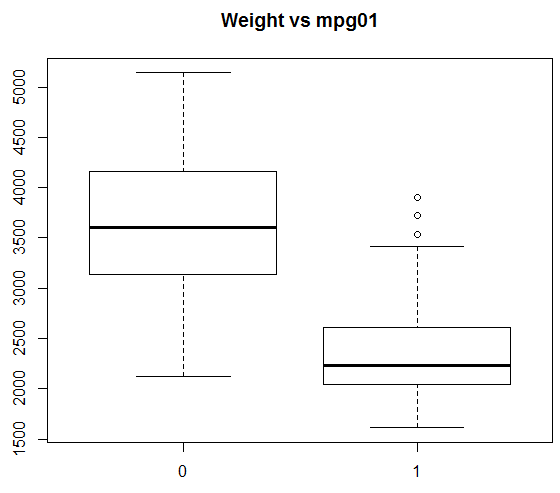




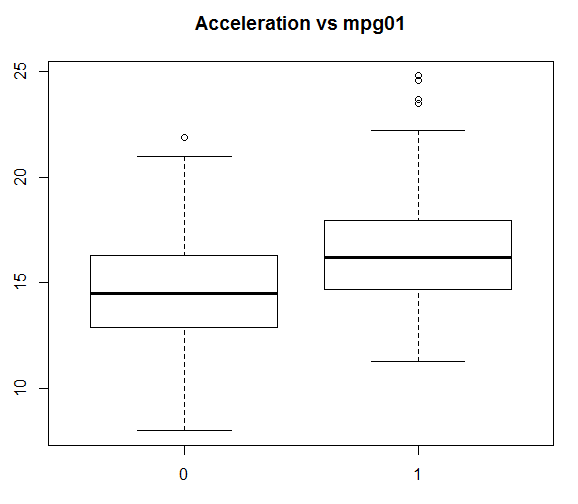




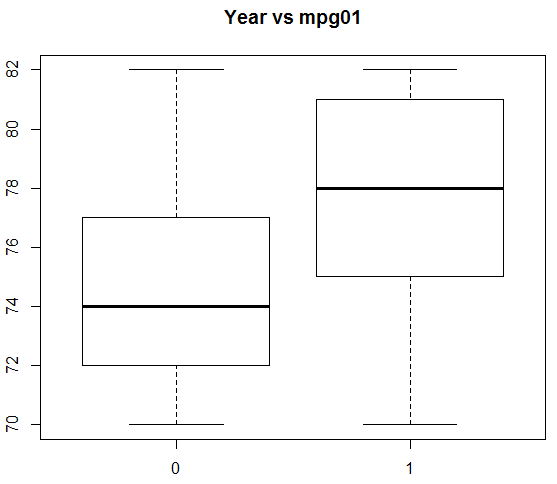


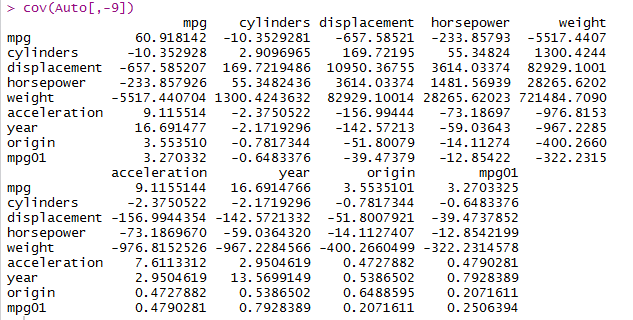












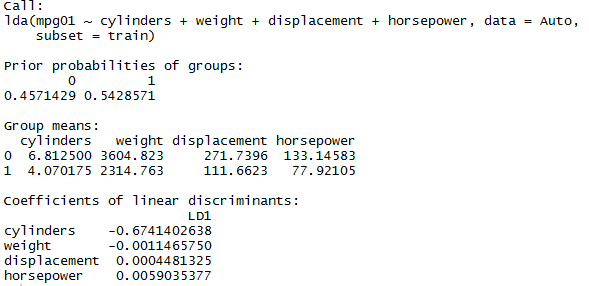
**A5b:** There exists some association between “mpg01” and “cylinders”, “weight”, “displacement” and “horsepower”.

1. Split the data into a training set and a test set.

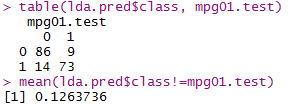


1. Perform LDA on the training data in order to predict mpg01using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?





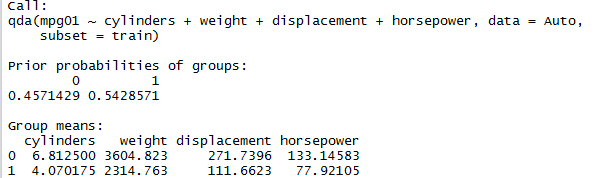




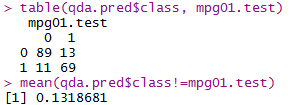
**A5d:** The test error of the model obtained is 12.6%.

1. Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?



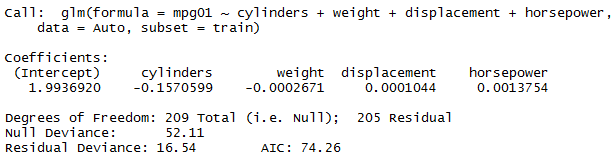


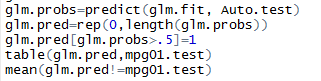


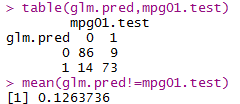
  
**A5e:** The test error of the model obtained is 13.2%.

1. Perform logistic regression on the training data in order to predict mpg01 using the variables that seemed most associated with mpg01 in (b). What is the test error of the model obtained?





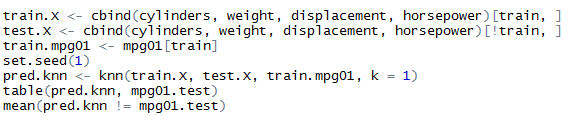


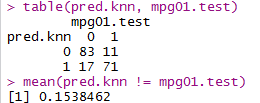


**A5f:** The test error of the model obtained is 12.6%.

1. Perform KNN on the training data, with several values of K, in order to predict mpg01. Use only the variables that seemed most associated with mpg01 in (b). What test errors do you obtain? Which value of K seems to perform the best on this data set?

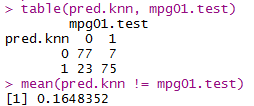
K=1





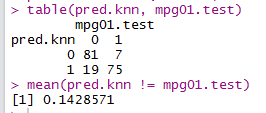
K=10





K=100





**A5g:** We may conclude that we have a test error rate of 14.3% for K=100. So, a K value of 100 seems to perform the best.

**Q6.** **Using basic statistical properties of the variance, as well as single variable calculus, derive (5.6). In other words, prove that *α* given by (5.6) does indeed minimize Var(*αX*+(1−*α*)*Y*).**

**Ans:**

First,

Var(αX+(1−α)Y)=++2α(1−α).

We now take the fist derivative of Var(αX+(1−α)Y) relative to α:

=2α−2+2α+2−4α.

We then set this equation to zero in order to pursue a minimized Var(αX+(1−α)Y). So we can get:

2α−2+2α+2−4α = 0

Then we get:

α =

Also we can check whether the second derivative is positive to assure that it is indeed minimized value.

= 2+2−4 = 2 Var(X-Y)

**Q7. We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of *n* observations.**

1. **What is the probability that the first bootstrap observation is not the jth observation from the original sample? Justify your answer.**

Ans: *1−1/n*

1. **What is the probability that the second bootstrap observation is not the jth observation from the original sample ?**

*Ans: 1−1/n*

1. **Argue that the probability that the jth observation is not in the bootstrap sample is**

Ans:   
The probability that the jth observation is not in the bootstrap sample is the product of the probabilities that each bootstrap observation is not the jth observation from the original sample :

(1−1/n)⋯(1−1/n)=

Note that these probabilities are independant.

1. When *n*=5, what is the probability that the jth observation is in the bootstrap sample?

*Ans : The* probability that the jth observation is in the bootstrap sample *is 1−=0.672*

1. When *n*=100, what is the probability that the jth observation is in the bootstrap sample ?

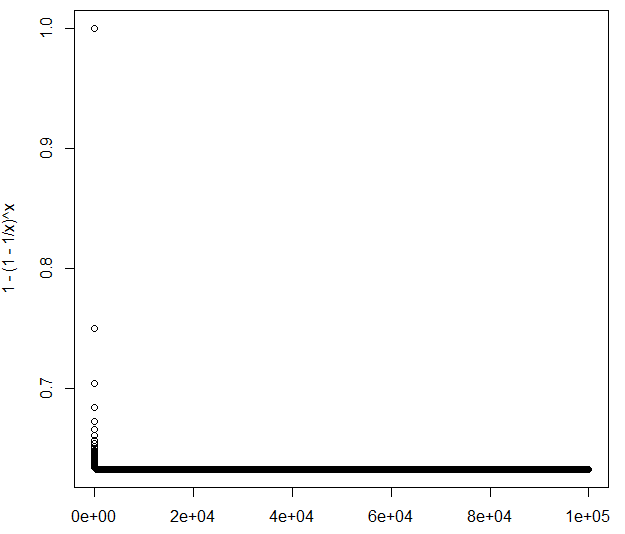
*Ans : The* probability that the jth observation is in the bootstrap sample *is 1−=0.634*

1. When *n*=10000, what is the probability that the jth observation is in the bootstrap sample ?

*Ans : The* probability that the jth observation is in the bootstrap sample *is 1−=0.632*

1. Create a plot that displays, for each integer value of *n* from 1 to 100000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

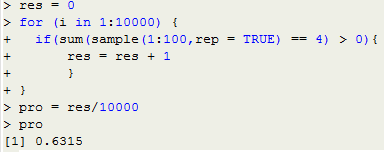
* x <- 1:100000
* plot(x, 1 - (1 - 1/x)^x)



As shown in above, this plot quickly reaches an asymptote at about y = *0.63*.

1. We will now investigate numerically the probability that a bootstrap sample of size *n*=100 contains the jth observation. Here *j*=4. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

Ans:

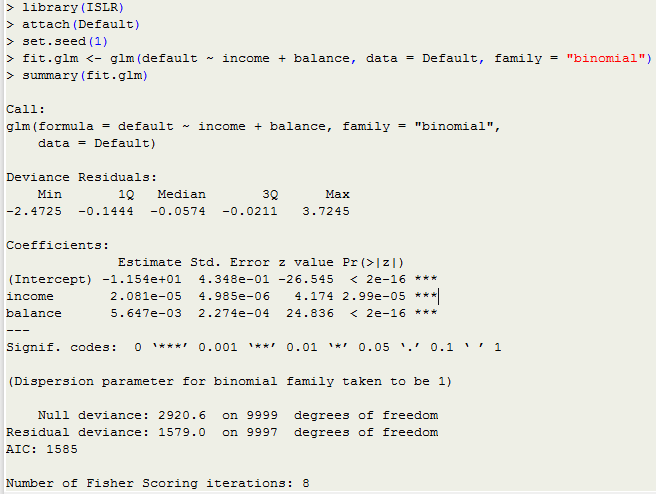


As shown in above, if the times we repeat is large enough, then the probability of a bootstrap sample that contains the jth observation will approach to 0.632. As we known, *=. So the probability will converge to 1−1/e = 0.632 when*

**Q8. In Chapter 4, we used logisitc regression to predict the probability of “default” using “income” and “balance” on the “Default” data set. We will now estimate the test error of this logistic regression model using the validation set approach. Do not forget to set a random seed before beginning your analysis.**

1. **Fit a logistic regression model that uses “income” and “balance” to predict “default”.**

Ans:



1. **Using the validation set approach, estimate the test error of this model. In order to do this, you must perform the following steps:**
2. **Split the sample set into a training set and a validation set.**

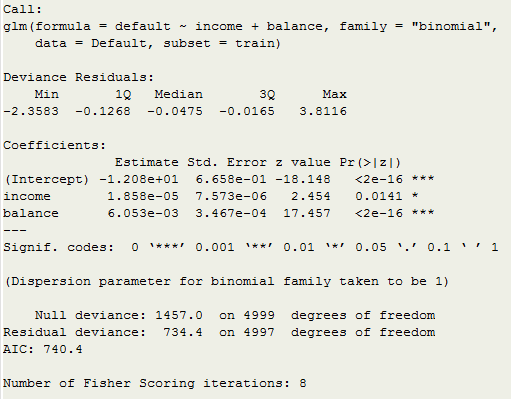
Ans: train <- sample(dim(Default)[1], dim(Default)[1] / 2)

1. **Fit a multiple logistic regression model using only the training observations.**

**Ans:**

fit.glm <- glm(default ~ income + balance, data = Default, family = "binomial", subset = train)

summary(fit.glm)

****

1. **Obtain a prediction of default status for each individual in the validation set by computing the posterior probability of default for that individual, and classifying the individual to the “default” category if the posterior probability is greater than 0.5.**

Ans:

probs <- predict(fit.glm, newdata = Default[-train, ], type = "response")

pred.glm <- rep("No", length(probs))

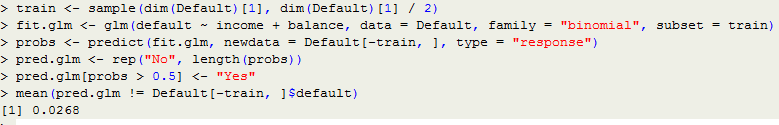
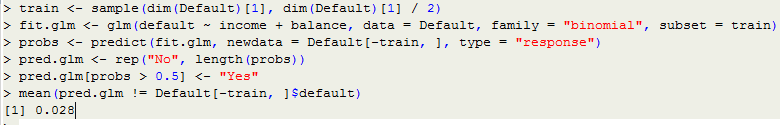
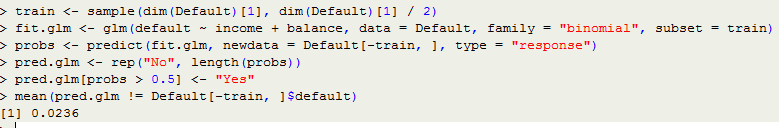
pred.glm[probs > 0.5] <- "Yes"

1. **Compute the validation set error, which is the fraction of the observations in the validation set that are misclassified.**

Ans:  


We have a 2.86% test error rate with the validation set approach.

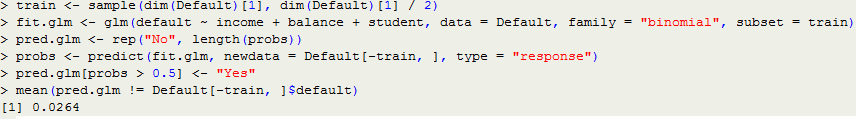
1. **Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.**

****

As shown in above, the validation estimate of the test error rate can be variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

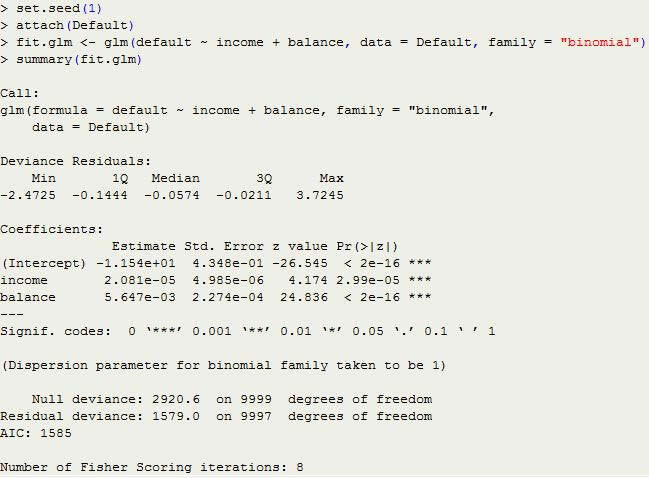
1. **Now consider a logistic regression model that predicts the probability of “default” using “income”, “balance”, and a dummy variable for “student”. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for “student” leads to a reduction in the test error rate.**

Ans:

****As shown in above, it doesn’t seem that adding the “student” dummy variable leads to a reduction in the validation set estimate of the test error rate.

**Q9. We continue to consider the use of a logistic regression model to predict the probability of “default” using “income” and “balance” on the “Default” data set. In particular, we will now computes estimates for the standard errors of the “income” and “balance” logistic regression coefficients in two different ways : (1) using the bootstrap, and (2) using the standard formula for computing the standard errors in the glm() function. Do not forget to set a random seed before beginning your analysis.**

1. **Using the summary() and glm() functions, determine the estimated standard errors for the coefficients associated with “income” and “balance” in a multiple logistic regression model that uses both predictors.**Ans:

****

The glm() estimates of the standard errors for the coefficients *β0*, *β1* and *β2* are respectively 0.4348, 4.985\*and 2.274\*.

1. **Write a function, boot.fn(), that takes as input the “Default” data set as well as an index of the observations, and that outputs the coefficient estimates for “income” and “balance” in the multiple logistic regression model.**Ans:

boot.fn <- function(data, index) {

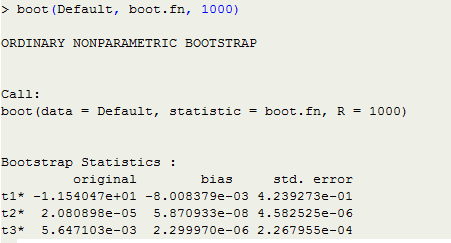
fit <- glm(default ~ income + balance, data = data, family = "binomial", subset = index)

return (coef(fit))

}

1. **Use the boot() function together with your boot.fn() function to estimate the standard errors of the logistic regression coefficients for “income” and “balance”.**Ans:

library(boot)

****

As shown in above, the bootstrap estimates of the standard errors for the coefficients *β0*, *β1* and *β2* are respectively 0.4239273, 4.582525\*and 2.267955\*

1. **Comment on the estimated standard errors obtained using the glm() function and using your bootstrap function.**

Ans:Although the estimated standard errors obtained by standard formula is slightly bigger than by bootstrap, but the two methods are actually very close. The difference for *β0*, *β1* and *β2* between two methods are approximately (0.4348-0.4239273)/0.4348=2.5%, (4.985-4.582525)/4.985=8.07% and (2.274-2.267955)/2.274=0.27%