**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 ?

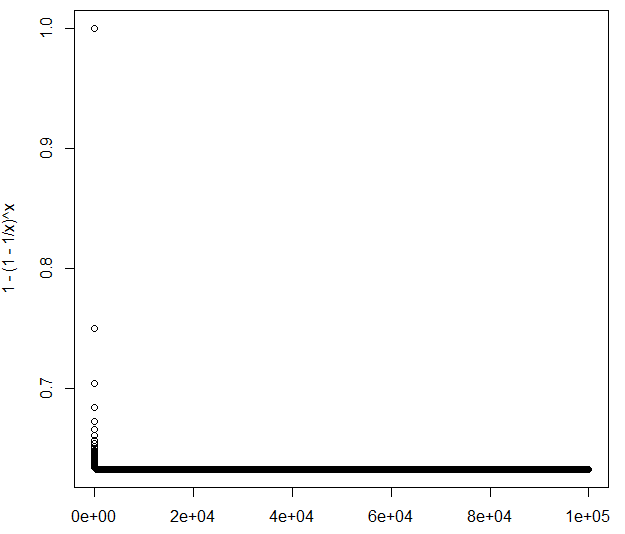
*Ans : The* probability that the jth observation is in the bootstrap sample *is 1−=0.634*

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.

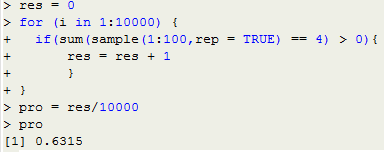
* x <- 1:100000
* plot(x, 1 - (1 - 1/x)^x)



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:

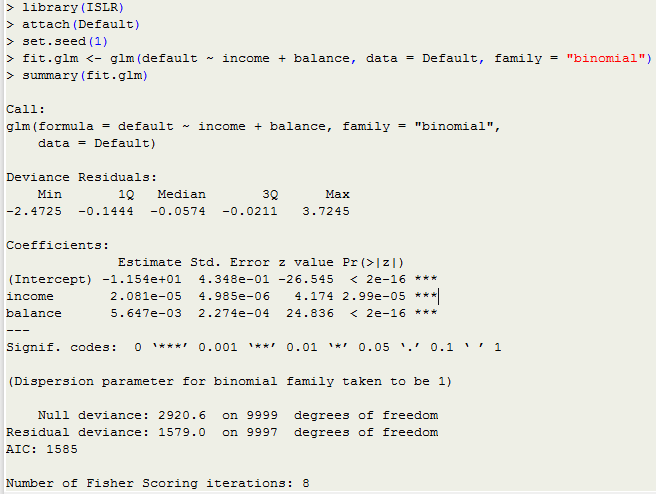


As shown in above, if the times we repeat is large enough, then the probability of a bootstrap sample that contains the jth observation will approach to 0.632. As we known, *=. So the probability will converge to 1−1/e = 0.632 when*

**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.**

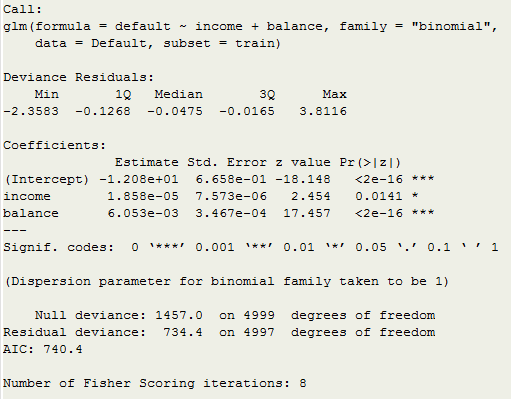
Ans:

* 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:

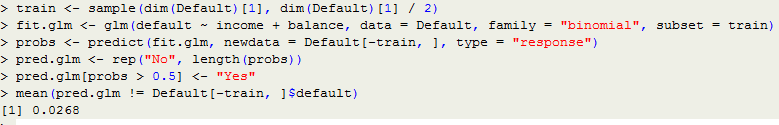
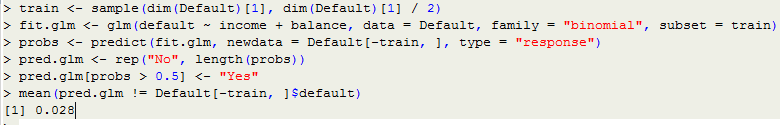
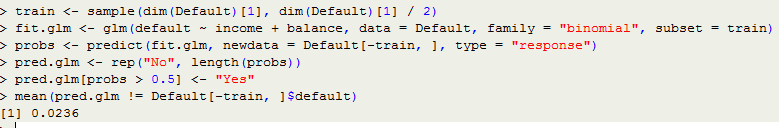
* probs <- predict(fit.glm, newdata = Default[-train, ], type = "response")
* pred.glm <- rep("No", length(probs))
* pred.glm[probs > 0.5] <- "Yes"

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.

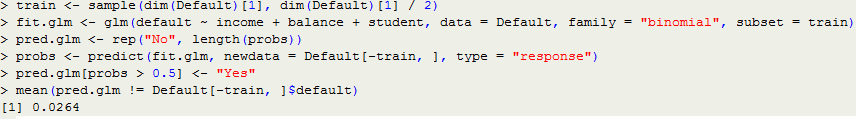
1. **Repeat the process in (b) three times, using three different splits of the observations into a training set and a validation set. Comment on the results obtained.**

****

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.

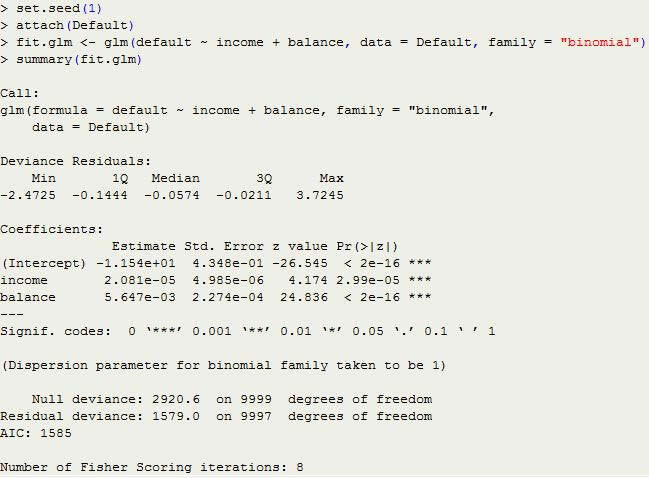
1. **Now consider a logistic regression model that predicts the probability of “default” using “income”, “balance”, and a dummy variable for “student”. Estimate the test error for this model using the validation set approach. Comment on whether or not including a dummy variable for “student” leads to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

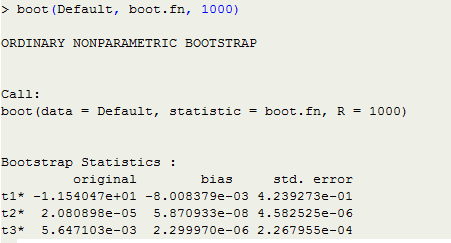
fit <- glm(default ~ income + balance, data = data, family = "binomial", subset = index)

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%