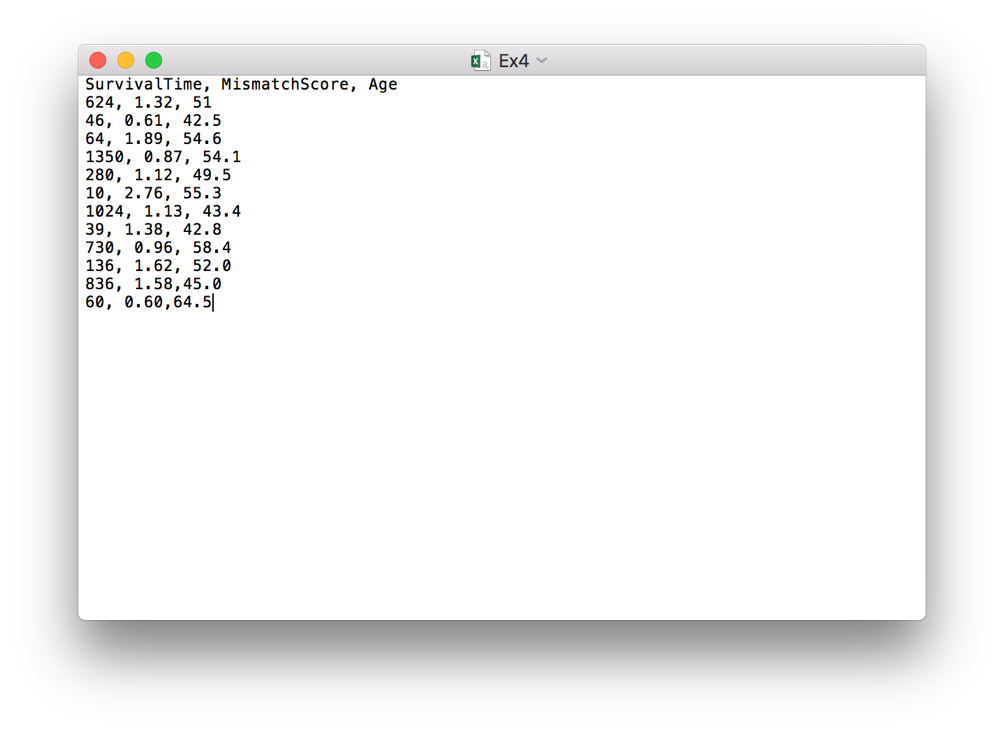
**Problem 4:**



Below is the R program used for this problem.

# read csv file

mydata <- read.csv("/Users/macbookpro/Desktop/Ex4.csv")

x1<- mydata$MismatchScore;

x2<- mydata$Age;

t <- mydata$SurvivalTime;

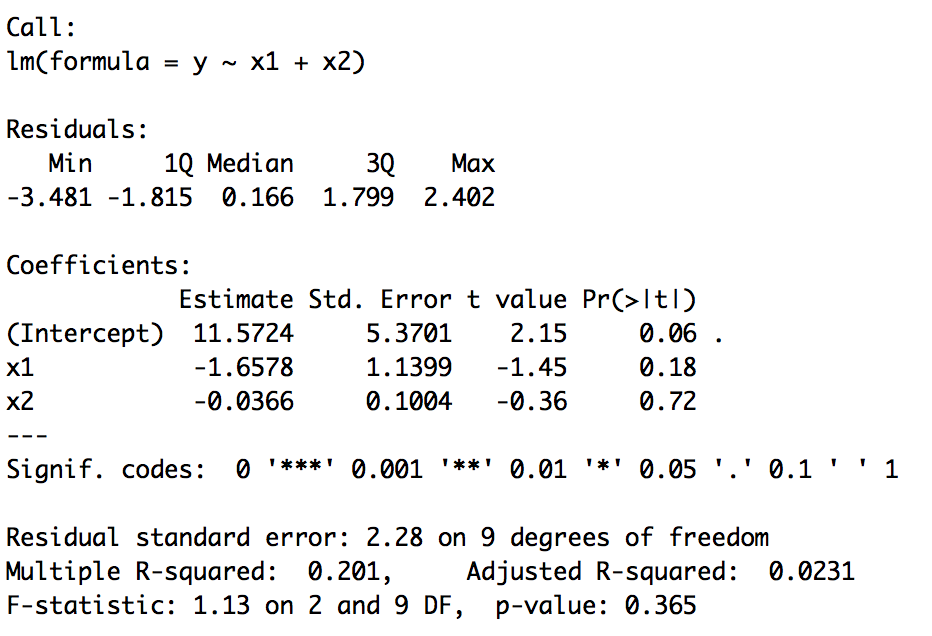
y <- log2(t);

#fit log model

fit <- lm(y ~ x1 + x2)

#Results of the model

summary(fit)



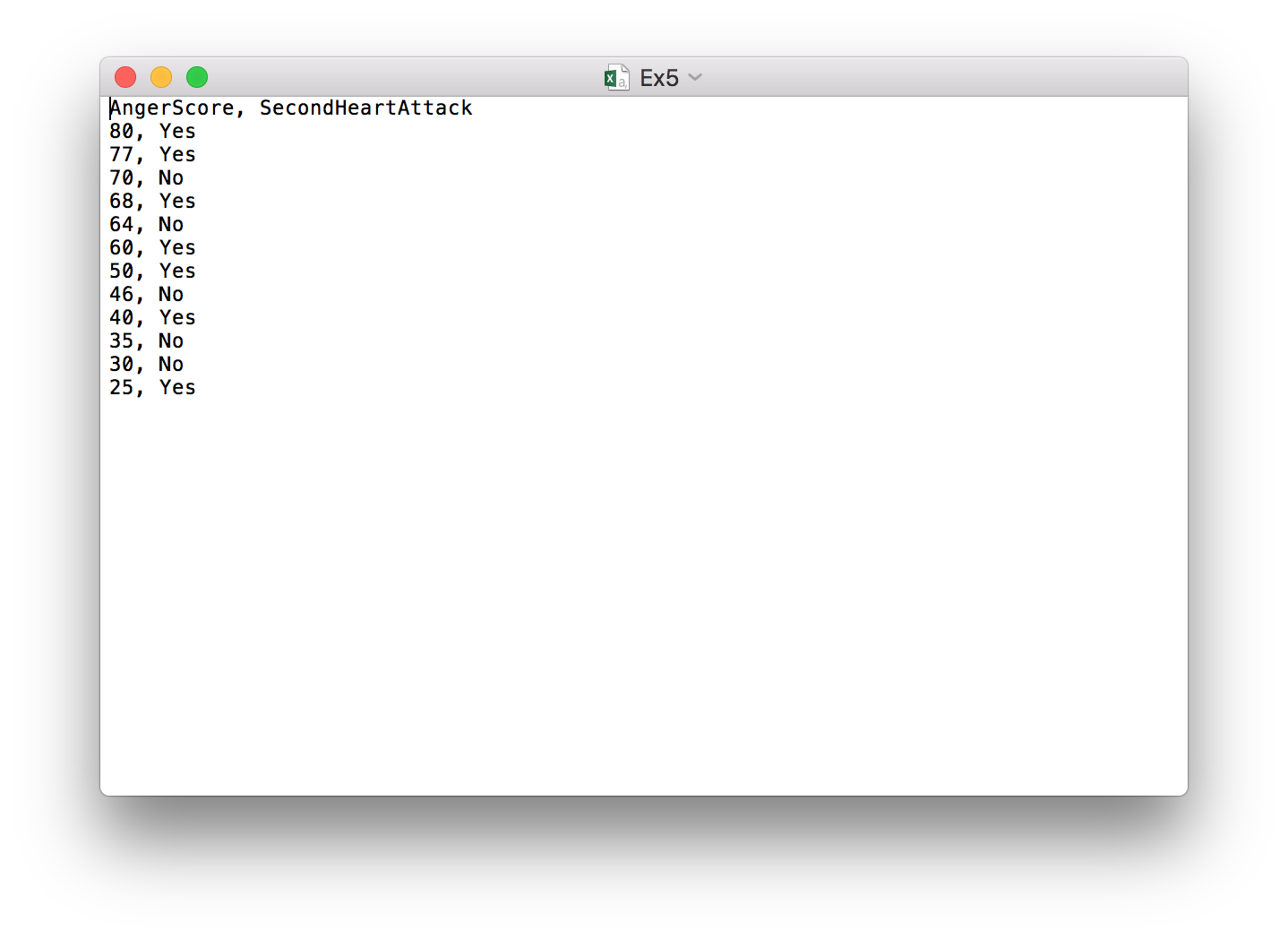
1. **Let the dependent variable be the logarithm of Survival time. Fit a multiple linear regression on the independent variables of Mismatch score and Age.**

**Solution:** Based on the program, we can see the  **equals 11.5724**. The estimate of **Mismatch score (x1) equals -1.6578**. The estimate of **Age (x2) equals -0.0366**.

1. **Compute an estimate of the variance of the error term.**

**Solution:** The residual standard error term will be 2.277 (it is different from the result in screenshot because after several times of running program, the answer is 2.277, not 2.28).

**Problem 5:**



Below is the R program used for this problem.

# read csv file

mydata <- read.csv("/Users/macbookpro/Desktop/Ex5.csv")

y <- mydata$SecondHeartAttack;

x <- mydata$AngerScore;

# change "YES" or "NO" into booleans

for (i in mydata$SecondHeartAttack){

if (i == "Yes"){

i <- 1

}

else{

i <- 0

}

}

# perform logistic regression on dataset

mylogit <- glm(y ~ x, data = mydata, family = "binomial")

# show the summary

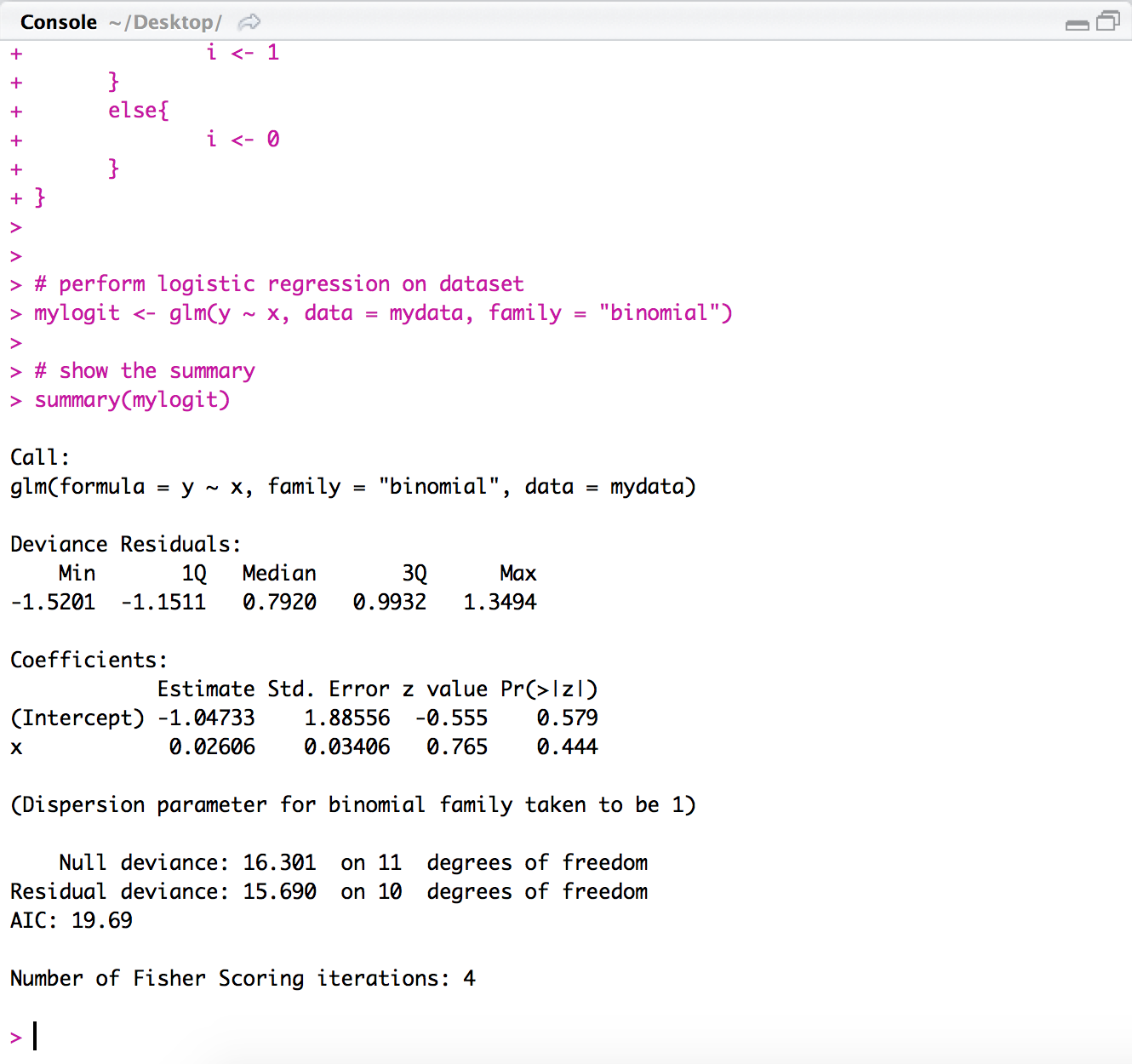
summary(mylogit)

# predict the new value

newdata <- data.frame(x = 55 )

pred <- predict.glm(mylogit, newdata, type = "response")

pred

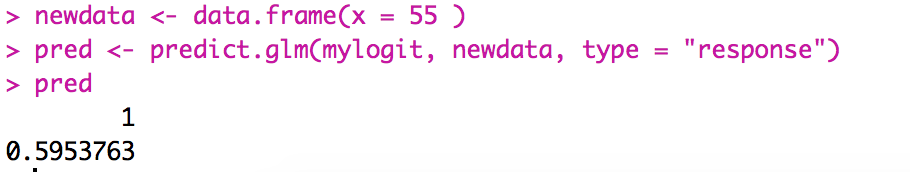


1. **Explain how the relationship between a second heart attack and one’s anger score can be analyzed via a logistic regression model.**

**Solution:** Based on the dataset, the values of “Second Heart Attack” contain ‘Yes’ and ‘No’. They can be categorized as two categories. So to some extent, it is a ‘Binary Classification’ problem. So it can be solved via logistic regression.

1. **Using a software package of your choice, estimate parameters for this model (for example, in Matlab to ﬁt a logistic model consider the command ‘glmﬁt’).**
2. **Estimate the probability that a heart attack patient with an anger score of 55 will have a second heart attack within 5 years.**

**Solution:** Based on the program, the probability is 0.5953763.



**Problem 6:**

**(a) For this data set compute the SVD (singular value decomposition) of the original matrix, and using this SVD discuss the expected results of performing a PCA on this data.**

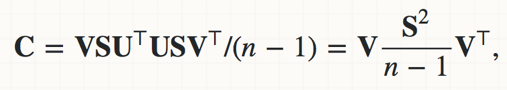
Basically the SVD on scaled matrix is the same as PCA. So the SVD on original matrix, I think it is helpless for the PCA.

**(b) Compute the PCA: First compute the mean(s) for the data, and subtract from the original data; second compute the covariance matrix including the scaling 1/(n − 1); third compute an eigenvalue decomposition and sort both the eigenvalues and eigenvectors in descending order.**

**(c) Plot and discuss the principal components. Discuss how this process and results might differ from a direct SVD of the de-biased, scaled data.**

**Solution:** The covariance matrix C is given by **C = XTX/(n-1)**. It is a symmetric matrix so it can be diagonalized as: **C=VLVT**, where **V** is a matrix of eigenvectors (each column is an eigenvector) and **L** is a diagonal matrix with eigenvalues ***lamda***in the decreasing order on the diagonal.

With Singular Decomposition of X, we can get **X=USVT**, where S is the diagonal matrix of singular values ***si***. So we can use some matrix multiplication rules to get



With the eigenvalue decomposition, **C=VLVT**, so we can obtain ***lamda* = s2/(n-1).** So we can know why this process might differ from SVD.