Math 70 Homework 8

Min Hyung (Daniel) Kang

Due May 27th, 2015

1. To evaluate the success of a recently released movie a survey has been conducted: 1000 people were asked if they like the movie. The data are in file movieFINAL.txt (y = 1: like the movie; y = 0: do not like the movie). Besides, the information about age and viewers education was obtained (ed=0:no hight school; ed=1: high school; ed=2: BC degree; ed=3: MS degree or higher). Use read.table with option header=T to download the data.

2. Use education status as a continuous variable (as is) and as a dummy variable to set up and run the logistic regression model. Test the hypothesis that the dummy variable approach is equivalent to treating education status continuosly by testing $B_1 - B_0 = B_2 - B_1 = B_3 - B_2$ by likelihood-ratio test (B_0 is the coefficient at no high school, B_1 is the coefficient at high school, etc.).

3. Plot y versus age for viewers with high school degree and superimpose with the fitted model values.

4. Estimate the proportion of people who liked the movie on average. Compute this proportion using two methods: (1) using y observations; (2) using the estimated model. Do the results match?

5. What is the chance that a 32 years old person with a college degree likes the movie? Compute the 95% CI for this proportion on the logit scale and then transform to the probability scale.

6. Compute the p-value for the null hypothesis that people with BC and MS education of the same age equally like the movie. Use the Wald test.

1 Problem 1

We can use the following R-Code to download the data to the code.

```
data = read.table("C:\\RCode(Math70)\\Homework 8\\movieFinal.txt",
      header=TRUE)
```

2 Problem 2

We use education status as a continuous variable first. Then we can simply do a glm : glm(data\$y $\sim$ data\$Age + data\$ed)

```
Call:
glm(formula = data$y ~ data$Age + data$ed, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3005  -1.0760  -0.9379   1.2537   1.4969

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.390615   0.207055   1.887  0.05922 .
data$Age    -0.006612   0.003697  -1.788  0.07373 .
data$ed     -0.211150   0.064286  -3.285  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1370.4  on 999  degrees of freedom
Residual deviance: 1356.2  on 997  degrees of freedom
AIC: 1362.2

Number of Fisher Scoring iterations: 4
```

Using a dummy variable, to test the hypothesis $H_0 : B_1 - B_0 = B_2 - B_1 = B_3 - B_2$, we consider the following :

$$\begin{bmatrix} Age_{1,0} & 1 & 0 \\ Age_{2,0} & 1 & 0 \\ Age_{3,0} & 1 & 0 \\ \cdots & \cdots & \cdots \\ Age_{1,1} & 1 & 1 \\ Age_{2,1} & 1 & 1 \\ \cdots & \cdots & \cdots \\ Age_{1,2} & 1 & 2 \\ Age_{2,2} & 1 & 2 \\ \cdots & \cdots & \cdots \\ Age_{1,3} & 1 & 3 \\ Age_{2,3} & 1 & 3 \end{bmatrix} \begin{bmatrix} a \\ B \\ \delta \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix}$$

```
Call:
glm(formula = ordereddata$y ~ ordereddata$Age + noHS + delta -
    1, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.3005  -1.0760  -0.9379   1.2537   1.4969

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
ordereddata$Age -0.006612   0.003697  -1.788  0.07373 .
noHS             0.390615   0.207055   1.887  0.05922 .
delta           -0.211150   0.064286  -3.285  0.00102 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1386.3  on 1000  degrees of freedom
Residual deviance: 1356.2  on  997  degrees of freedom
AIC: 1362.2

Number of Fisher Scoring iterations: 4
```

After getting the log likelihood of each model, note again that likelihood ratio test is computed as follows :

$$-2(maxl(\theta_{10}, \theta_2) - maxl(\theta, \theta_2)) \sim \chi^2(1)$$

```
[1] "Log Likelihood of Continuous model :  -678.1192"
[1] "Log Likelihood of Dummy model -678.1192"
[1] "Chisq :  0"
```

We see that two approaches are exactly the same.

3 Problem 3

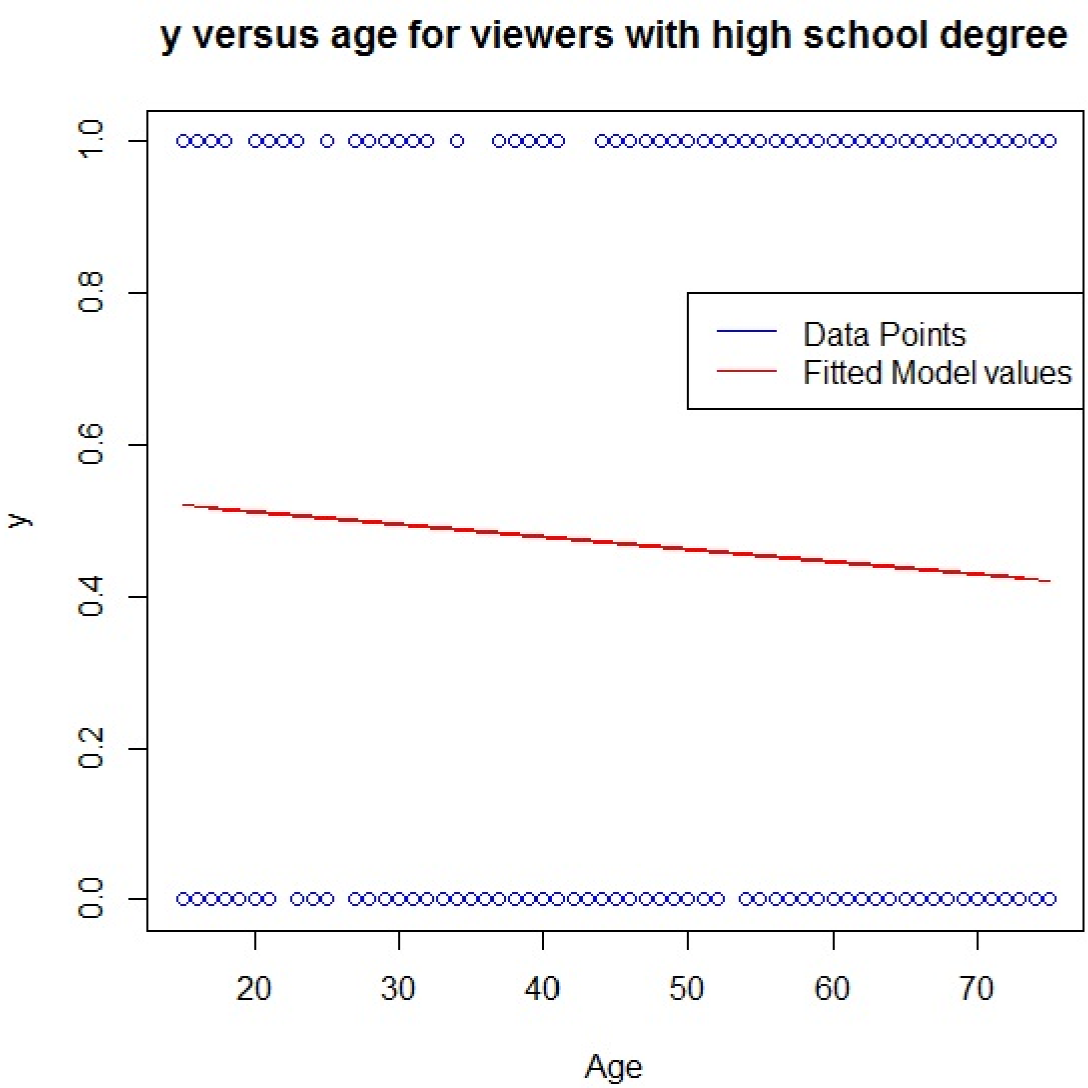
We select data from the table where the value of education is 1. Then using glm model we compute the fitted model values.

$$glm(y \sim Age + ed, family = binomial)$$

gives us coefficients $coeff_{Age}, coeff_{ed}$.

Then we construct the model as following :

$$\hat{y} = \frac{e^{intercept+coeff_{Age}*Age+coeff_{ed}*ed}}{1+e^{intercept+coeff_{Age}*Age+coeff_{ed}*ed}}$$



4 Problem 4

We estimate the proportion of people who liked the movie. Using the observations, we can simply divide the number of people who liked the movie by total number of people.

$$\bar{p} = \frac{\text{Number of people who liked the movie}}{n}$$

We estimate the proportions by computing the mean of age and mean of education, and plugging them into the regression model which we computed previously..

$$\hat{p} = intercept + coeff_{Age} * \bar{Age} + coeff_{ed} * \bar{ed}$$

```
[1] 0.437
(Intercept)
  0.4361309
```

We see that both methods give almost exactly the same result.

5 Question 5

We consider a 32 year old person with a college degree.

First, we use the model before to get the value of probability of him liking the movie.

$$p = \frac{e^s}{1 + e^s}$$

where

$$s = intercept + coeff_{Age} * Age + coeff_{ed} * ed$$

and

$$Age = 32, ed = 2$$

Computing this equation, we get the following result : `0.4394813`

Now we compute the 95% CI.

The $(1 - \alpha)$ CI for a value s can be evaluated as :

$$s \pm Z_{1-\alpha/2}\sqrt{x'Cx}$$

.

Here, C could be understood as $(X'DX)^{-1}$, where D is the following matrix :

$$D = \begin{bmatrix} \frac{e^{B'x_1}}{(1+e^{B'x_1})^2} & 0 & 0 & \cdots & 0 \\ 0 & \frac{e^{B'x_2}}{(1+e^{B'x_2})^2} & 0 & \cdots & 0 \\ 0 & 0 & \frac{e^{B'x_3}}{(1+e^{B'x_3})^2} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & \frac{e^{B'x_n}}{(1+e^{B'x_n})^2} \end{bmatrix}$$

And we can compute in the probability scale by the following :

$$\frac{e^s}{1+e^s}$$

Using R, we get the following :

```
[1] "95% CI on the logit scale : ( -0.4069 , -0.0797 )"
[1] "95% CI on the probability scale : ( 0.3997 , 0.4801 )"
```

6 Problem 6

We know that when we use wald test, the following is used :

$$Z = \frac{\hat{\theta}_1 - \theta_{10}}{\sqrt{I_{11}^{-1}(\theta)/n}} \simeq N(0,1)$$

We leave out BC education and construct glm.

$$\begin{bmatrix} Age_{1,0} & 1 & 0 & 0 \\ Age_{2,0} & 1 & 0 & 0 \\ Age_{3,0} & 1 & 0 & 0 \\ \cdots & \cdots & \cdots & \\ Age_{1,1} & 0 & 1 & 0 \\ Age_{2,1} & 0 & 1 & 0 \\ \cdots & \cdots & \cdots & \\ Age_{1,2} & 0 & 0 & 0 \\ Age_{2,2} & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \\ Age_{1,3} & 0 & 0 & 1 \\ Age_{2,3} & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a \\ \delta_0 \\ \delta_1 \\ \delta_3 \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \cdots \\ y_n \end{bmatrix}$$

$$glm(y \sim Age + ed_1 + ed_2 + ed_3)$$

where each δ_i is the dummy variable for education level.

```
Call:
glm(formula = ordereddata$y ~ ordereddata$Age + ed1 + ed2 + ed3,
    family = binomial)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.397  -1.050  -0.942   1.281   1.488

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      0.047339   0.199017   0.238  0.81199
ordereddata$Age -0.006214   0.003713  -1.673  0.09426 .
ed1              0.555065   0.201531   2.754  0.00588 **
ed2             -0.112764   0.164133  -0.687  0.49207
ed3             -0.300553   0.175090  -1.717  0.08606 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1370.4  on 999  degrees of freedom
Residual deviance: 1349.9  on 995  degrees of freedom
AIC: 1359.9

Number of Fisher Scoring iterations: 4
```

Note that here, the intercept indicates the mean of BC education. Hence, we can look at the coefficient for ed_3, using its std.error and estimate in the wald test.

Hence,

$$Z = \frac{\hat{\theta}_1 - \theta_{10}}{\sqrt{I_{11}^{-1}(\theta)/n}} = \frac{-0.3006 - 0}{0.1751} = -1.717 \simeq N(0,1)$$

The p-value is $0.0861 > 0.05$. Hence, we fail to reject the null hypothesis. Hence, people with BC and MS education of same age equally like the movie.

The following R-code was used.

```
function(job=1){
  dump("hw8q1","C:\\RCode(Math70)\\Homework 8\\hw8q1.r")
  data = read.table("C:\\RCode(Math70)\\Homework 8\\movieFinal.txt",
       header=TRUE)
  n = nrow(data)

  #Simple logistic regression
  result = glm(data$y ~ data$Age + data$ed, family=binomial) #age, binomial
  coefAge = coef(result)["data$Age"]
  coefEd = coef(result)["data$ed"]
  intercept = coef(result)["(Intercept)"]

  #Run logistic regression model of y with respect to education status
  if(job == 2){
       #Simple logistic regression
       result = glm(data$y ~ data$Age + data$ed, family=binomial)
       print(summary(result))

       #Using dummy variables
       ordereddata = data[order(data$ed),]
       n0=length(data$ed[data$ed==0])
       n1=length(data$ed[data$ed==1])
       n2=length(data$ed[data$ed==2])
       n3=length(data$ed[data$ed==3])

       #Using dummy variables
       delta = matrix(c(rep(0,n0),rep(1,n1),rep(2,n2),rep(3,n3)),ncol=1)
       noHS = rep(1,n)

       result2 = glm(ordereddata$y ~ ordereddata$Age + noHS + delta -1,family=binomial)
       print(summary(result2))

       #Compute the likelihood test
       print(paste("Log Likelihood of Continuous model : ",round(logLik(result),4)))
       print(paste("Log Likelihood of Dummy model",round(logLik(result2),4)))
       print(paste("Chisq : ",round(pchisq(-2*(logLik(result2)-logLik(result)),df=1),4)))
  }

  #Plot y versus age for viewers with high school degree
  if(job == 3){
       #Select y values with high school degree
       yindex = data$y[data$ed==1]
       ageindex = data$Age[data$ed==1]
       edindex = data$ed[data$ed==1]
       plot(ageindex, yindex, xlab = "Age", ylab = "y",col="BLUE",
```

```
        main="y versus age for viewers with high school degree")

    #construct the model
    modelVal = intercept + coefAge * ageindex + coefEd * edindex
    modelVal = exp(modelVal) / (1 + exp(modelVal))
    lines(ageindex, modelVal,col="RED")

    legend(50,0.8,c("Data Points","Fitted Model values"),
       lwd=c(1,1),col=c("BLUE","RED"))
}

#Compute the proportion of people who liked the movie
if(job == 4){
    count = length(data$y[data$y==1])
    print(count/n)

    #Compute the averages
    ageavg = mean(data$Age)
    edavg = mean(data$ed)

    #Estimate the proportion
    modelVal = intercept + coefAge * ageavg + coefEd * edavg
    modelVal = exp(modelVal) / (1 + exp(modelVal))
    print(modelVal)
}

#Compute the probability that a 32 year old person with college degree
#likes the movie
if(job == 5){

    #construct the model
    modelVal = intercept + coefAge * 32 + coefEd * 2
    modelVal = exp(modelVal) / (1 + exp(modelVal))
    print(modelVal)

    #Compute D Matrix
    Beta = rbind(intercept,coefAge, coefEd);
    D = matrix(0,nrow=n,ncol=n)
    for(i in 1:n){
       xi = matrix(rbind(1,data$Age[i],data$ed[i]),ncol=1)
       D[i,i] = exp(t(Beta)%*%xi)/(1+exp(t(Beta)%*%xi))^2
    }

    #Compute C Matrix
    X = cbind(rep(1,n),data$Age,data$ed)
    C= solve(t(X) %*% D %*% X)
```

```
        #Compute CI
        x = rbind(1,32,2)
        teststat = 1.96 * sqrt(t(x)%*%C%*%x)
        logit = intercept + coefAge * 32 + coefEd * 2
        leftLogit = logit - teststat
        rightLogit = logit + teststat
        print(paste("95% CI on the logit scale : (",
           round(leftLogit,4),",",round(rightLogit,4),")"))
        leftProb = exp(leftLogit)/(1+exp(leftLogit))
        rightProb = exp(rightLogit)/(1+exp(rightLogit))
        print(paste("95% CI on the probability scale : (",
           round(leftProb,4),",",round(rightProb,4),")"))
    }

    if(job==6){
        #Using dummy variables
        ordereddata = data[order(data$ed),]
        n0=length(data$ed[data$ed==0])
        n1=length(data$ed[data$ed==1])
        n2=length(data$ed[data$ed==2])
        n3=length(data$ed[data$ed==3])

        ed1 = matrix(c(rep(1,n0),rep(0,n-n0)),ncol=1)
        ed2 = matrix(c(rep(0,n0),rep(1,n1),rep(0,n2+n3)),ncol=1)
        ed3 = matrix(c(rep(0,n-n3),rep(1,n3)),ncol=1)

        result2 = glm(ordereddata$y ~ ordereddata$Age + ed1 + ed2 + ed3,family=binomial)
        print(summary(result2))
    }
}
```