Week One Homework

Applied Logistic Regression

Adrian Cuyugan

Exercise One

For this exercise, you will need the Myopia Study dataset. Download the MYOPIA.dta Stata file, or you can also access the data through this CSV file.

```
require(RCurl)
```

```
## Loading required package: RCurl
## Loading required package: bitops
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
# Load the dataset
myopia <- getURL("https://d396qusza40orc.cloudfront.net/logisticregression/data/MYOPIA-fixed.csv",
                 ssl.verifypeer=0L, followlocation=1L)
writeLines(myopia, "myopia.csv")
myopia <- read.csv("myopia.csv", head=T, sep=",")
```

One variable that is clearly important is the initial value of spherical equivalent refraction (SPHEQ).

Complete the following:

1. Write down the equation for the logistic regression model of SPHEQ on MYOPIA. Write down the equation for the logit transformation of this logistic regression model. What characteristic of the outcome variable, MYOPIA, leads us to consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between MYOPIA and SPHEQ? Discuss your response in the homework forum.

   The probability of each record having a value of MYOPIC $= 1$,

   $\pi(x) = E(y|x) = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}}$

   The odds ratio of the equation is:

   $OR = \frac{\pi(x)}{(1-\pi(x))}$

   Converting the odds ratio in its natural logarithm function:

   $ln\left(\frac{\pi(x)}{(1-\pi(x))}\right) = \beta_0 + \beta_1 x$

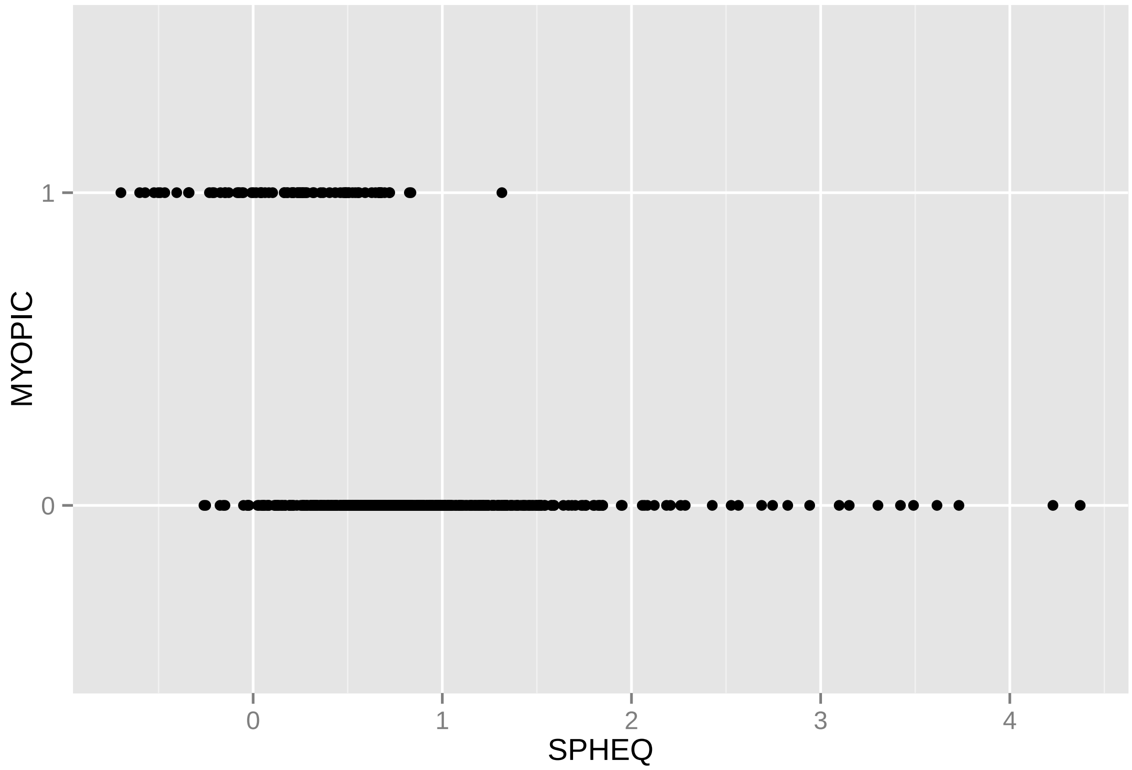
   In linear regression, it assumes that the distribution is normally distributed:

   $y = E(y|x) + \varepsilon$ where $\varepsilon \rightarrow N(0, \sigma^2)$ and therefore $y|x\ N(E(y|x), \sigma^2)$

   This is not applicable in a logistic regression model because it follows a binomial distribution as previously formulated above.

2. Form a scatterplot of MYOPIA vs. SPHEQ.

```
ggplot(myopia, aes(x=SPHEQ, y=factor(MYOPIC))) + geom_point() +
    ylab("MYOPIC")
```



3. Write down an expression for the likelihood and log likelihood for the logistic regression model in part (A) using the ungrouped, n=618, data. Obtain expressions for the two likelihood equations.

   $\ell(\beta) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i}$

   Converted into natural logarithm function:

   $Ln(\beta) = ln(\ell(\beta)) = \sum_{i=1}^{n} y_i ln[\pi(x_i)] + (1 - y_i) ln[1 - \pi(x_i)]$

   Maximum likelihood equations:

   $\sum_{i=1}^{n} (y_i - \pi(x_i)) = 0$ $\sum_{i=1}^{n} x_i (y_i - \pi(x_i)) = 0$

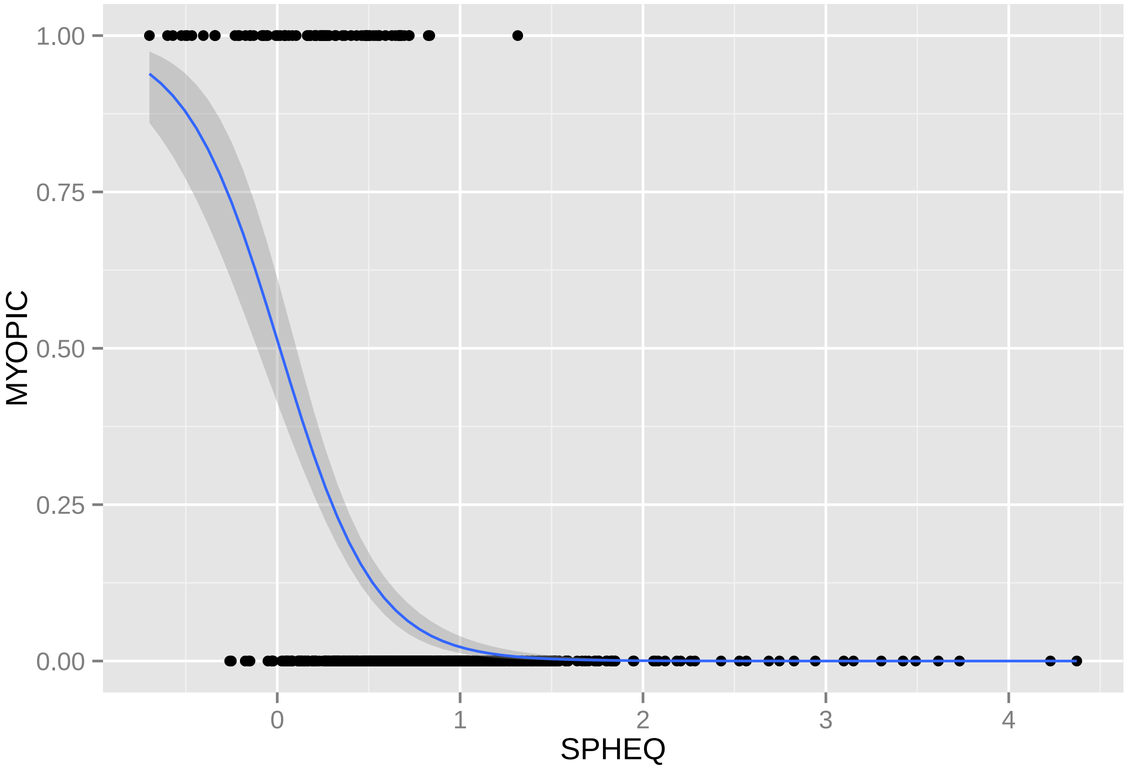
4. Using Stata, obtain the maximum likelihood estimates of the parameters of the logistic regression model in part (A). These estimates should be based on the ungrouped, n=618, data. Using these estimates, write down the equation for the fitted values, that is, the estimated logistic probabilities. Plot the equation for the fitted values on the axes used in the scatterplots in parts (B) and (C).

```
summary(glm(as.factor(MYOPIC) ~ SPHEQ, data=myopia, family="binomial"))
```

```
##
## Call:
## glm(formula = as.factor(MYOPIC) ~ SPHEQ, family = "binomial",
##     data = myopia)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6435  -0.4533  -0.2681  -0.1029   3.1602
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.05397    0.20675   0.261    0.794
## SPHEQ       -3.83310    0.41837  -9.162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 480.08  on 617  degrees of freedom
## Residual deviance: 337.34  on 616  degrees of freedom
## AIC: 341.34
##
## Number of Fisher Scoring iterations: 6
```

Based on the coefficients provided by the logistic regression model, here are the estimates:
$\pi_e(x) = \frac{e^{(0.05397-3.8331x)}}{(1+e^{(0.05397-3.8331x)})}$

```
ggplot(myopia, aes(x=SPHEQ, y=MYOPIC)) + geom_point() +
    stat_smooth(method="glm", family="binomial")
```



Exercise Two

For this exercise, you will need the ICU dataset. Download the icu.dta Stata file, or you can also access the data through this CSV file.

The ICU dataset consists of a sample of 200 subjects who were part of a much larger study on survival of patients following admission to an adult intensive care unit (ICU). The major goal of this study was to develop a logistic regression model to predict the probability of survival to hospital discharge of these patients. A number of publications have appeared which have focused on various facets of this problem.

Complete the following:

1. Write down the equation for the logistic regression model of STA on AGE. Write down the equation for the logit transformation of this logistic regression model. What characteristic of the outcome variable, STA, leads us to consider the logistic regression model as opposed to the usual linear regression model to describe the relationship between STA and AGE?

```r
summary(glm(as.factor(STA) ~ AGE, data=icu, family="binomial"))
```

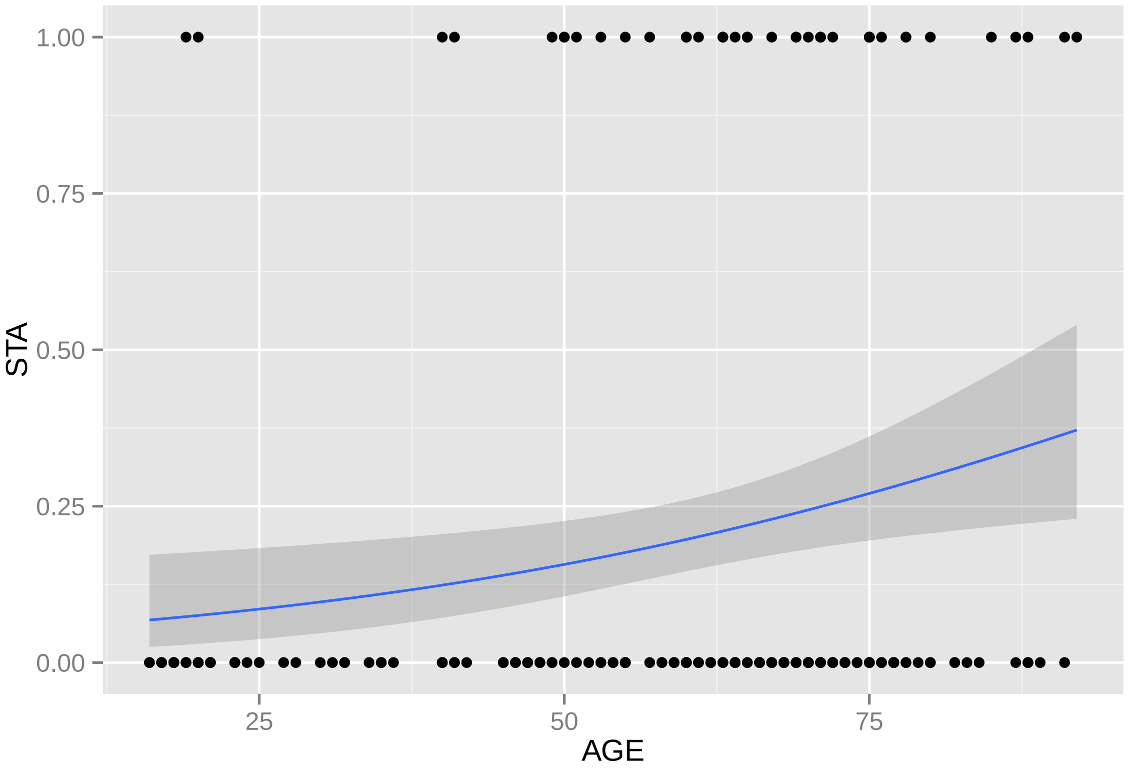
```
## 
## Call:
## glm(formula = as.factor(STA) ~ AGE, family = "binomial", data = icu)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -0.9536  -0.7391  -0.6145  -0.3905   2.2854  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -3.05851    0.69608  -4.394 1.11e-05 ***
## AGE          0.02754    0.01056   2.607  0.00913 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for binomial family taken to be 1)
## 
##     Null deviance: 200.16  on 199  degrees of freedom
## Residual deviance: 192.31  on 198  degrees of freedom
## AIC: 196.31
## 
## Number of Fisher Scoring iterations: 4
```

2. Form a scatterplot of STA versus AGE.

3. Write down an expression for the likelihood and log likelihood for the logistic regression model in part (A) using the ungrouped, n=200, data. Obtain expressions for the two likelihood equations.

4. Using Stata, obtain the maximum likelihood estimates of the parameters of the logistic regression model in part (A). These estimates should be based on the ungrouped, n=200, data. Using these estimates, write down the equation for the fitted values, that is, the estimated logistic probabilities. Plot the equation for the fitted values on the axes used in the scatterplots in part (B).

    ```r
    ggplot(icu, aes(x=AGE, y=STA)) + geom_point() +
        stat_smooth(method="glm", family="binomial")
    ```



5. Summarize (describe in words) the results presented in the plot obtained from parts (B) and (D). Discuss your response in the homework forum.