Homework 6

Juwon Lee, Economics and Statistics, UCLA

2023-03-07

tinytex::install_tinytex()

1.

(a) Write a fitted logistic regression model to predict exercise induced angina (exand) as a function of maximum heart rate.

```
hw6 <- read.csv("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regression/Cle
hw6$exand[hw6$exand=='exercise indicued angina'] <- 1
hw6$exand[hw6$exand=='no exercise indicued angina'] <- 0
hw6$exand <- as.integer(hw6$exand)
```

$\log(\frac{\hat{\theta}(x_i)}{1-\hat{\theta}(x_i)}) = \hat{\beta}_0 + \hat{\beta}_1 * \text{maximum heart rate.}$

(b) State the null hypothesis and alternative hypothesis to test the significance of maxheartrate predictor.

$H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

(c) Find the Wald test statistics and its corresponding p-value. Make a conclusion for the test for (b).

```
length(hw6$exand[hw6$maxheartrate == 108])
```

```
## [1] 2
```

```
sum(hw6$exand[hw6$maxheartrate == 108])
```

```
## [1] 2
```

```
maxheartrate <- c() ; exand <- c() ; m <- c()

for(i in min(hw6$maxheartrate):max(hw6$maxheartrate)) {
  maxheartrate[i-70] = i
  exand[i-70] = sum(hw6$exand[hw6$maxheartrate == i])
  m[i-70] = length(hw6$exand[hw6$maxheartrate == i])
}

hw6 <- data.frame(maxheartrate, exand, m, 'm-y'=m-exand, 'theta'=exand / m, '1-theta' = 1-exand/m)

m1 <- glm(cbind(hw6$exand, hw6$m.y)~maxheartrate, family=binomial)
summary(m1)
```

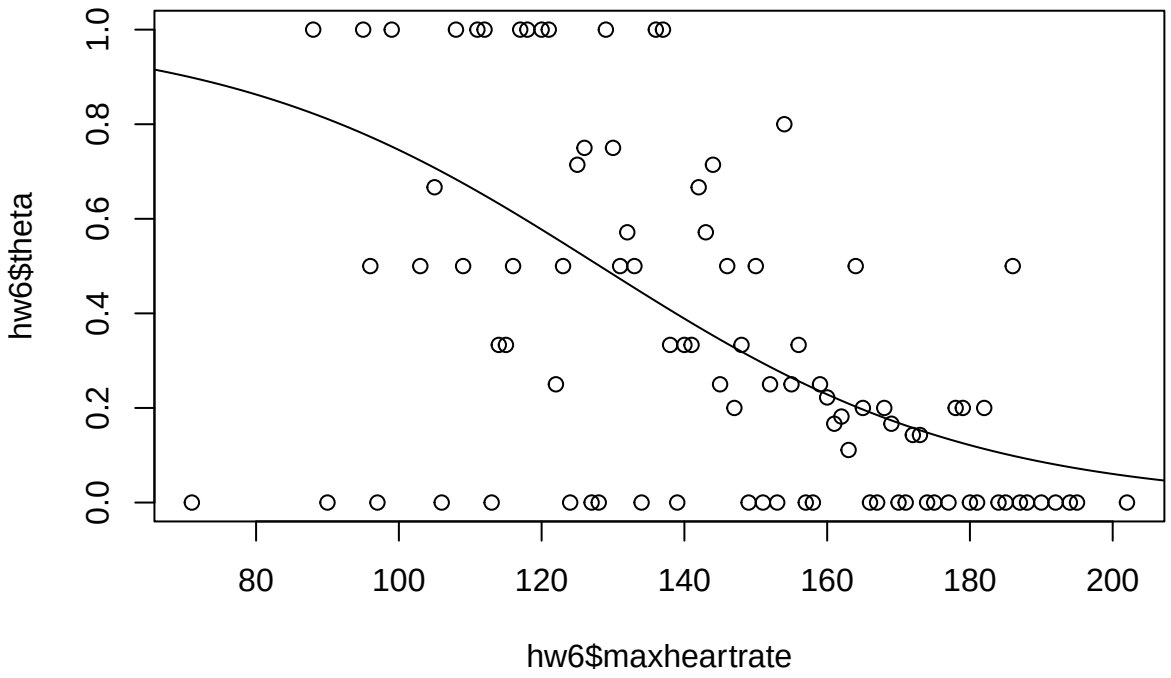
```
##
```

```
## Call:
## glm(formula = cbind(hw6$exand, hw6$m.y) ~ maxheartrate, family = binomial)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.13987  -0.46701   0.00000   0.05637   2.46141
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   4.893946   0.917451   5.334 9.59e-08 ***
## maxheartrate -0.038187   0.006254  -6.106 1.02e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 139.253  on 90  degrees of freedom
## Residual deviance:  94.848  on 89  degrees of freedom
## AIC: 181.26
##
## Number of Fisher Scoring iterations: 4
```

Thus, $Z = \frac{\hat{\beta_1}}{estimated\ se(\hat{\beta_1})} = \frac{-0.038187}{0.006254} = -6.106$,
and p-value $= 1.02e-09 < 0.05$, so that the predictor is significant ($\beta_1 \neq 0$).

(d) Draw the plot of max heart rate vs. the probability of exercise-induced angina to show that the logistic model is appropriate. Adjust the scale of the exand to see the scatter plot with the fitted logistic regression line.

```
x <- seq(60, 220, 0.1)
y <- 1/(1+exp(-1*(m1$coefficients[1] + m1$coefficients[2] * x)))
plot(hw6$maxheartrate, hw6$theta)
lines(x,y)
```



(e) If we increase maximum heart rate by five units, what change do you expect to have on exercise induced angina(exand)?

Because $\log(\frac{\hat{\theta}(x_i)}{1-\hat{\theta}(x_i)}) = 4.893946 + -0.038187 *$ maximum heart rate,
$\log(\frac{\hat{\theta}(x_i)}{1-\hat{\theta}(x_i)}) = 4.893946 + -0.038187 * 5 = 4.703013$,
$\frac{\hat{\theta}(x_i)}{1-\hat{\theta}(x_i)} = e^{4.703013} = 110.278943$,
$\hat{\theta}(x_i) = 110.278943 - 110.278943 * \hat{\theta}(x_i)$,
$\hat{\theta}(x_i) = \frac{110.278943}{111.278943} = 0.991014$.

(f) Using the difference in deviance G^2's of the model, test the significance of the model.

```
Gdiff <- m1$null.deviance - m1$deviance

pchisq(Gdiff, 1, lower=FALSE)
```

```
## [1] 2.670219e-11
```

Thus, we can reject the null, so that we can conclude that
the logistic model with the predictor is appropriate.

2.

Chapter 6 of Bradbury (2007), a book on baseball, uses regression analysis to compare the success of the 30 Major League Baseball teams. For example, the author considers the relationship between x_i, market size (i.e., the population in millions of the city associated with each team) and Y_i, the number of times team i made the post-season playoffs in the $m_i = 10$ seasons between 1995 and 2004.

The author found that "it is hard to find much correlation between market size and ... success in making the playoffs. The relationship ... is quite weak." The data is plotted in Figure 8.16 and it can be found on the book website in the file playoffs.txt. The output below provides the analysis implied by the author's comments.

(a) Describe in detail two major concerns that potentially threaten the validity of the analysis implied by the author's comments.

(1) The problem of scales, because $Y_i \leq 10$ and $x_i > 1000000$. \ (2) The problem of multicollinearity, because AverageWins and PlayoffAppearances may be strongly correlated.

(b) Using an analysis which is appropriate for the data, show that there is very strong evidence of a relationship between Y and x.

```
baseball <- read.table("/Users/user/Desktop/Yonsei/Junior/3-2/Introduction to Data Analysis and Regressi

sort <- sort(baseball$Population, decreasing=T)

sort <- sort[-c(1,3,5)]

Population <- c() ; PlayoffAppearances <- c() ; m <- c()

for(i in 1:length(sort)) {
  Population[i] = sort[i]
  PlayoffAppearances[i] = sum(baseball$PlayoffAppearances[baseball$Population == sort[i]])
  m[i] = sum(baseball$n[baseball$Population == sort[i]])
}

hw6_2 <- data.frame(Population, PlayoffAppearances, m, 'm-y'=m-PlayoffAppearances,
                    'theta'= PlayoffAppearances / m, '1-theta' = 1- PlayoffAppearances/m)

m2 <- glm(cbind(hw6_2$PlayoffAppearances, hw6_2$m.y)~Population, family=binomial)
summary(m2)
```

```
##
## Call:
## glm(formula = cbind(hw6_2$PlayoffAppearances, hw6_2$m.y) ~ Population,
##     family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4876  -2.1682  -0.1894   1.0385   5.3057
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.45843    0.21102  -6.911  4.8e-12 ***
## Population   0.07807    0.02751   2.838  0.00455 **
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106.518  on 26  degrees of freedom
## Residual deviance:  98.644  on 25  degrees of freedom
## AIC: 150.48
##
## Number of Fisher Scoring iterations: 4
```

Thus, $Z = \frac{\hat{\beta_1}}{estimated\ se(\hat{\beta_1})} = \frac{0.07807}{0.02751} = 2.838$,
and p-value $= 0.00455 < 0.05$, so that the predictor is significant ($\beta_1 \neq 0$).