Stats 101A Homework #6

Damien Ha

2023-03-08

Problem 1

```
cleveland <- read.csv("Cleveland_Data.csv")
```

Part A

```
cleveland$exand = as.factor(cleveland$exand)
cleveland$exand = -as.numeric(cleveland$exand)+2
m1 <- glm(cleveland$exand ~ cleveland$maxheartrate, family=binomial)
summary(m1)
```

```
##
## Call:
## glm(formula = cleveland$exand ~ cleveland$maxheartrate, family = binomial)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.1399  -0.8091  -0.6071   1.0599   2.1508
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             4.893946   0.917425   5.334 9.58e-08 ***
## cleveland$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: 382.1  on 301  degrees of freedom
## Residual deviance: 337.7  on 300  degrees of freedom
## AIC: 341.7
##
## Number of Fisher Scoring iterations: 3
```

Part B

H_0: There is not a significant relationship between heart rate and the probability of exand

H_A: There is a significant relationship between heart rate and the probability of exand

Part C

```
library(aod)
```

```
## Warning: package 'aod' was built under R version 4.1.2
```

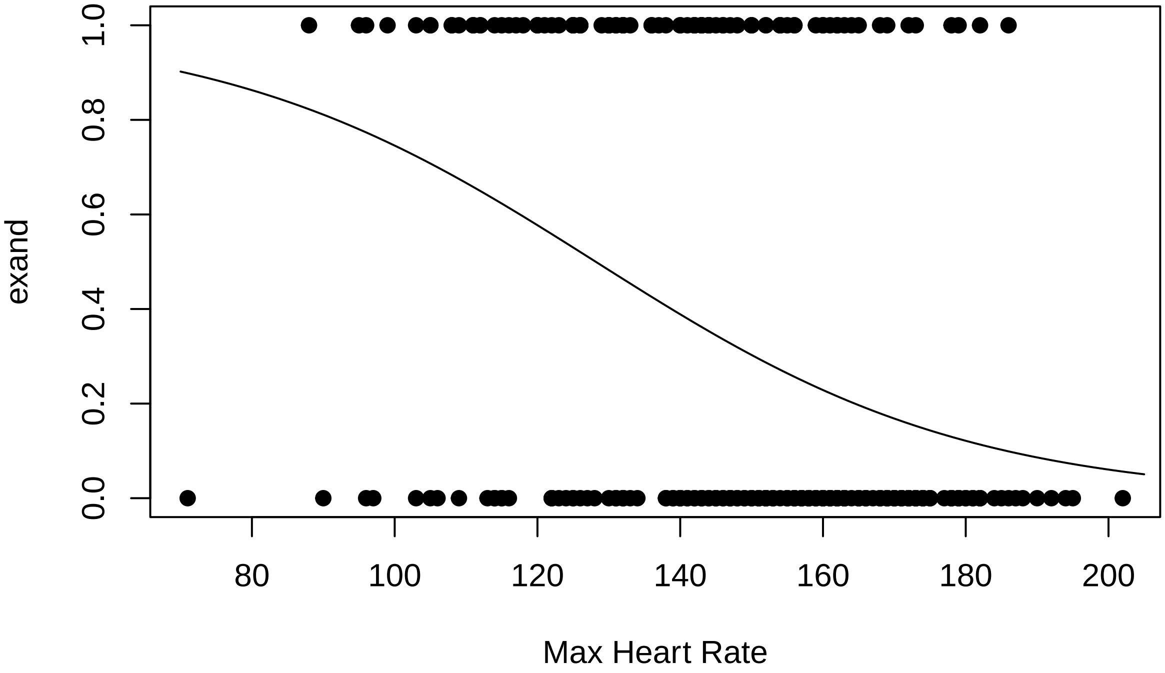
```
wald.test(vcov(m1), coef(m1), 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 37.3, df = 1, P(> X2) = 1e-09
```

The p-value is 0.00000001 which is significant at a level of $\alpha = 0.05$ because $0.00000001 < 0.05$. There is then sufficient evidence to reject the null hypothesis and conclude that there exists a significant relationship between heart rate and probability of exand.

Part D

```
plot(cleveland$maxheartrate, cleveland$exand, pch = 19,
     xlab = "Max Heart Rate", ylab = "exand")
x <- seq(70,205,0.5)
y <- 1/(1+exp(-1*(m1$coeff[1] + m1$coeff[2]*x)))
lines(x,y)
```



Part E

```
(1/(1 + exp(-1 * (m1$coeff[1] + m1$coeff[2]*5)))) -
  (1/(1 + exp(-1*(m1$coeff[1] + m1$coeff[2]*0))))
```

```
##  (Intercept)
## -0.001550332
```

So on average, with an increase of 5 in max heart rate, we would have a 0.155% decrease in exand

Part F

```
pchisq(m1$null.deviance - m1$deviance,1,lower=FALSE)
```

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

Part G

```
1 - m1$deviance/m1$null.deviance
```

```
## [1] 0.1162119
```

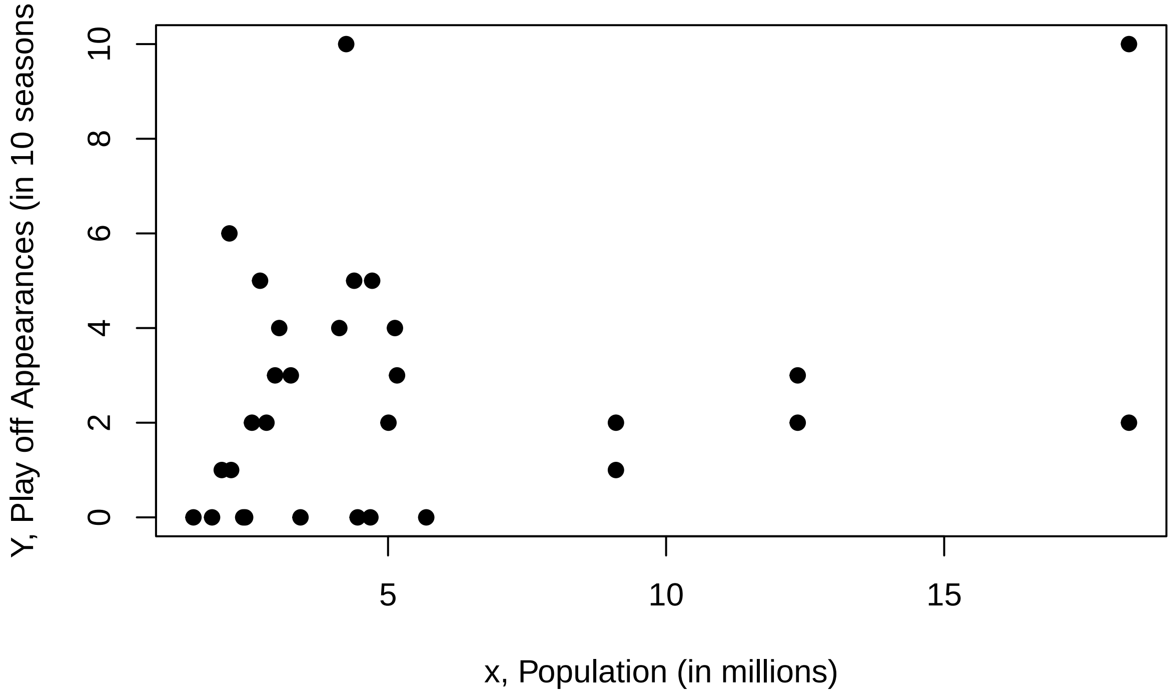
The R^2 value is 0.116 which suggests that 11.6% of the deviance is explained by the model. This is a fairly low percentage and suggest that there could be a better fit of model to the data.

Problem 2

Exercise 8.3.1

```
playoffs <- read.table("playoffs.txt", header = T)
```

```
plot(x = playoffs$Population, y = playoffs$PlayoffAppearances, pch = 19,
xlab = "x, Population (in millions)", ylab = "Y, Play off Appearances (in 10 seasons")
```

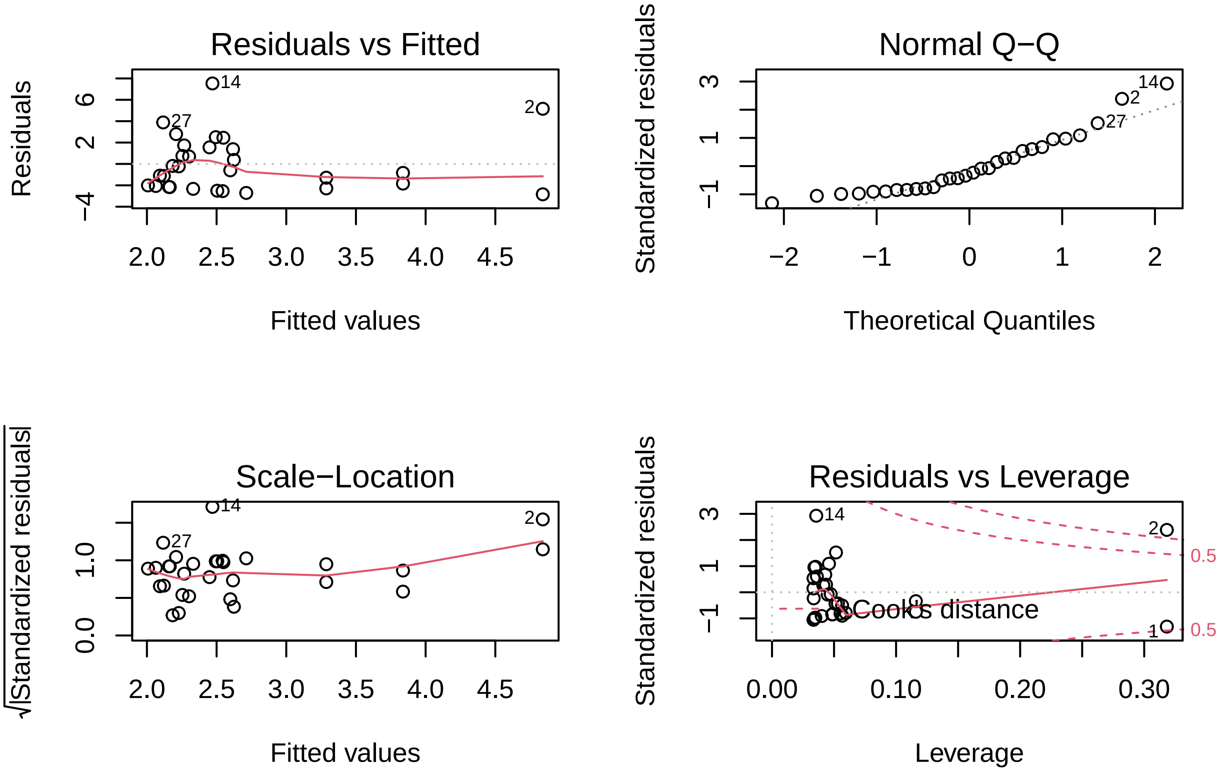


```
m2 <- lm(PlayoffAppearances ~ Population, data = playoffs)
summary(m2)
```

```
##
## Call:
## lm(formula = PlayoffAppearances ~ Population, data = playoffs)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8409 -2.1347 -0.7179  1.5085  7.5298
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.7547     0.7566   2.319   0.0279 *
## Population    0.1684     0.1083   1.555   0.1311
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.619 on 28 degrees of freedom
## Multiple R-squared:  0.07952,    Adjusted R-squared:  0.04664
## F-statistic: 2.419 on 1 and 28 DF,  p-value: 0.1311
```

```r
par(mfrow = c(2, 2))
plot(m2)
```



Part A

One major concern is correlation. The regression analysis shows there may be a positive linear trend between population and playoff appearances, but it is a weak correlation. It's possible outside factors play a role in playoff appearances, and these other variables must be taken into account. Another major concern is sample size. The sample size of 30 is not particularly large, so it may not be generalizable to all baseball teams or sports.

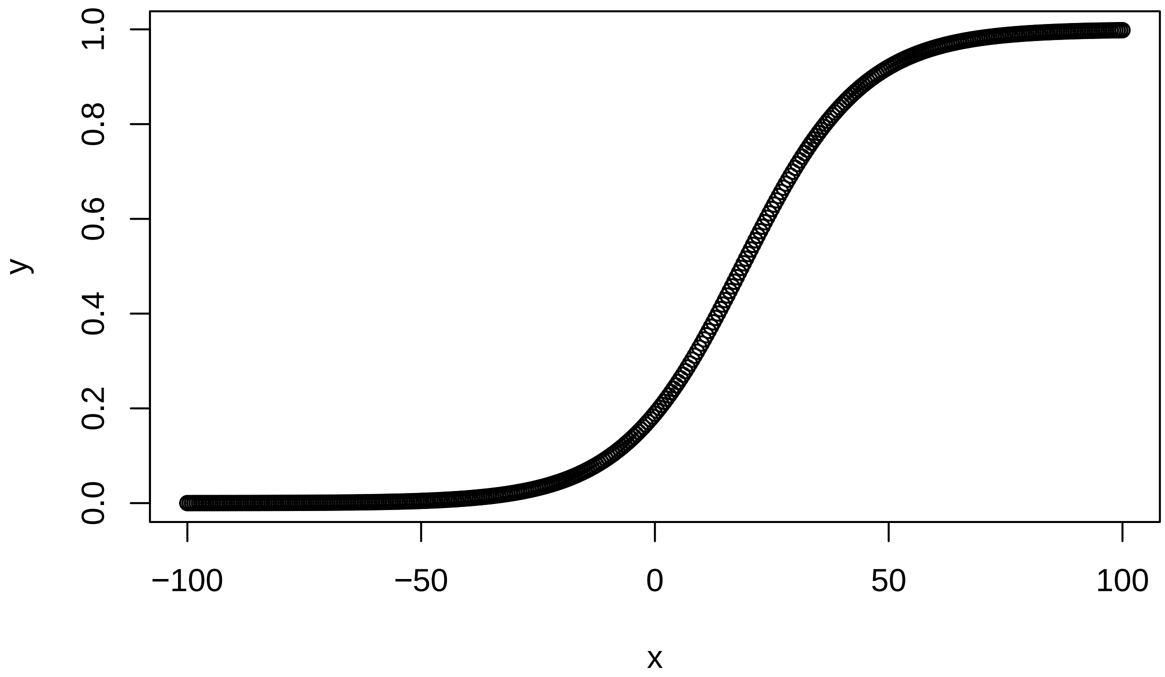
Part B

```
m3 <- glm(cbind(playoffs$PlayoffAppearances, playoffs$n - playoffs$PlayoffAppearances)
          ~ playoffs$Population, family = "binomial")
summary(m3)
```

```
##
## Call:
## glm(formula = cbind(playoffs$PlayoffAppearances, playoffs$n -
##     playoffs$PlayoffAppearances) ~ playoffs$Population, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4876  -2.0968  -0.4703   1.0666   5.3057
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.45843    0.21102  -6.911  4.8e-12 ***
## playoffs$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: 124.10  on 29  degrees of freedom
## Residual deviance: 116.22  on 28  degrees of freedom
## AIC: 170.33
##
## Number of Fisher Scoring iterations: 4
```

```
x <- seq(0,20,0.5)
y <- 1/(1+exp(-1*(m3$coeff[1] + m3$coeff[2]*x)))

x <- seq(-100,100,0.5)
y <- 1/(1+exp(-1*(m3$coeff[1] + m3$coeff[2]*x)))
plot(x,y)
```



```
wald.test(vcov(m3), coef(m3), 2)
```

```
## Wald test:
## ----------
##
## Chi-squared test:
## X2 = 8.1, df = 1, P(> X2) = 0.0045
```

The wald test gives a p-value of 0.0045. At a level of $\alpha = 0.05$, this is significant as $0.0045 < 0.05$. There is evidence to reject the null hypothesis and conclude there is a significant relationship between population and playoff appearances