**STAT 40001/STAT 50001 Statistical Computing Fall 2024**

**Lab -19**

**Q.N. 1)** In American football points can be scored by a placekicker kicking a ball through a target area at an end of the field. A success occurs when the football is kicked over the crossbar and between the two uprights of the goal posts. The placekicker’s team receives 1 or 3 points, where a point after touchdown (PAT) receives 1 point and a field goal receives 3 points. A placekick that is not successful receives 0 points. Bilder and Loughin (1998) use logistic regression to estimate the probability of success for a placekick (PAT or field goal) in the National Football League (NFL). They examine a number of explanatory variables, including:

week: Week of the season

distance: Distance of the placekick in yards

change: Binary variable denoting lead-change (1) vs. non-lead-change (0) placekicks; lead changing placekicks are those that have the potential to change which team is winning the game (for example, if a field goal is attempted by a team that is losing by 3 points or less, they will no longer be losing if the kick is successful)

elap30: Number of minutes remaining before the end of the half, with overtime placekicks receiving a value of 0

PAT: Binary variable denoting the type of placekick, where a PAT attempt is a 1 and a field goal attempt is a 0

type: Binary variable denoting outdoor (1) vs. dome (0) placekicks

field: Binary variable denoting grass (1) vs. artificial turf (0) placekicks

wind: Binary variable for placekicks attempted in windy conditions (1) vs. non-windy conditions (0); we define windy as a wind stronger than 15 miles per hour at kickoff in an outdoor stadium

good: Binary variable denoting successful (1) vs. failed (0) placekicks; this is our response variable

The data are available in the link below as well as in the **Brightspace.**

<http://www.chrisbilder.com/categorical/programs_and_data.html>

1. Model the binary variable **good** using **distance** as a regressor variable**.**
2. Use your model to predict the probability of success from 37 yards.

> # 0.0 31.0 0.0 32.0 0.0 33.0

> # 1.0 34.0 0.0 35.0 0.0 35.0

> # 1.0 35.0 1.0 36.0 0.0 37.0

> # 0.0 38.0 1.0 39.0 0.0 40.0

> # 1.0 40.0 1.0 40.0 0.0 41.0

> # 1.0 42.0 1.0 43.0 1.0 44.0

> # 0.0 45.0 1.0 45.0 1.0 45.0

> # 0.0 46.0 1.0 47.0 1.0 48.0

> # 0.0 49.0 1.0 50.0 1.0 50.0

> #

> data = scan()

1: 0.0 30.0 1.0 30.0 0.0 30.0

7: 0.0 31.0 0.0 32.0 0.0 33.0

13: 1.0 34.0 0.0 35.0 0.0 35.0

19: 1.0 35.0 1.0 36.0 0.0 37.0

25: 0.0 38.0 1.0 39.0 0.0 40.0

31: 1.0 40.0 1.0 40.0 0.0 41.0

37: 1.0 42.0 1.0 43.0 1.0 44.0

43: 0.0 45.0 1.0 45.0 1.0 45.0

49: 0.0 46.0 1.0 47.0 1.0 48.0

55: 0.0 49.0 1.0 50.0 1.0 50.0

61:

Read 60 items

> fdata = matrix(data, ncol = 2, byrow = T)

> head(fdata)

[,1] [,2]

[1,] 0 30

[2,] 1 30

[3,] 0 30

[4,] 0 31

[5,] 0 32

[6,] 0 33

> y = fdata[,1]

> x = fdata[,2]

> plot(x,y)

> model = glm(y~x, family = binomial)

> model

Call: glm(formula = y ~ x, family = binomial)

Coefficients:

(Intercept) x

-4.8075 0.1251

Degrees of Freedom: 29 Total (i.e. Null); 28 Residual

Null Deviance: 41.46

Residual Deviance: 37.46 AIC: 41.46

> cat("Beta0 = -4.8075, Beta1 = 0.1251")

Beta0 = -4.8075, Beta1 = 0.1251

> cat("Fitted Logistic Model:

+ pi\_hat = [1 + exp(4.8075 - 0.1251\*x)]^-1")

Fitted Logistic Model:

pi\_hat = [1 + exp(4.8075 - 0.1251\*x)]^-1

> curve(predict(model, data.frame(x=x), type = 'resp'),add = T)

> predict(model, data.frame(x=2)) # gives beta0 and beta1

1

-4.557357

> predict(model, data.frame(x=2), type = "resp") # gives the probability

1

0.01038085

> # install.packages("popbio")

> library(popbio)

> logi.hist.plot(x,y,boxp=FALSE,type="hist",col="gray")

> summary(model)

Call:

glm(formula = y ~ x, family = binomial)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -4.80751 2.65576 -1.810 0.0703 .

x 0.12508 0.06676 1.874 0.0610 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

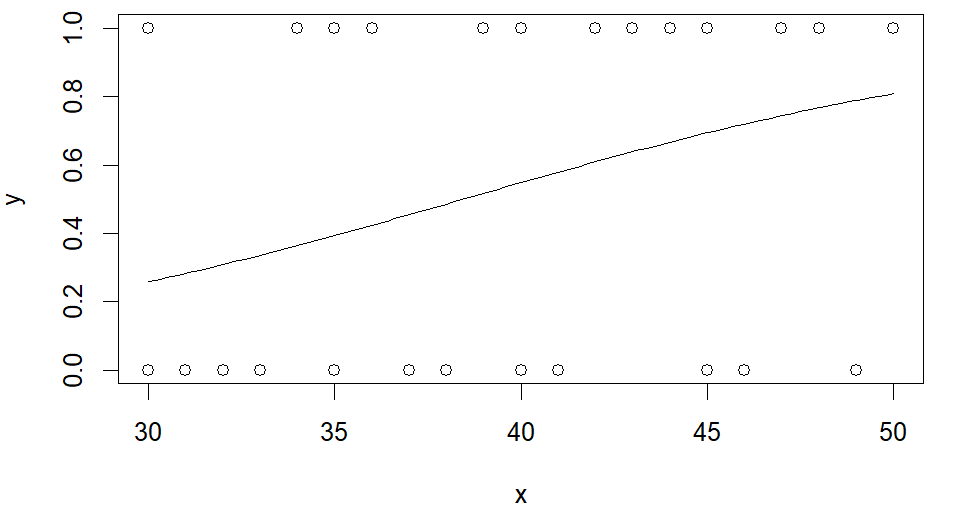
(Dispersion parameter for binomial family taken to be 1)

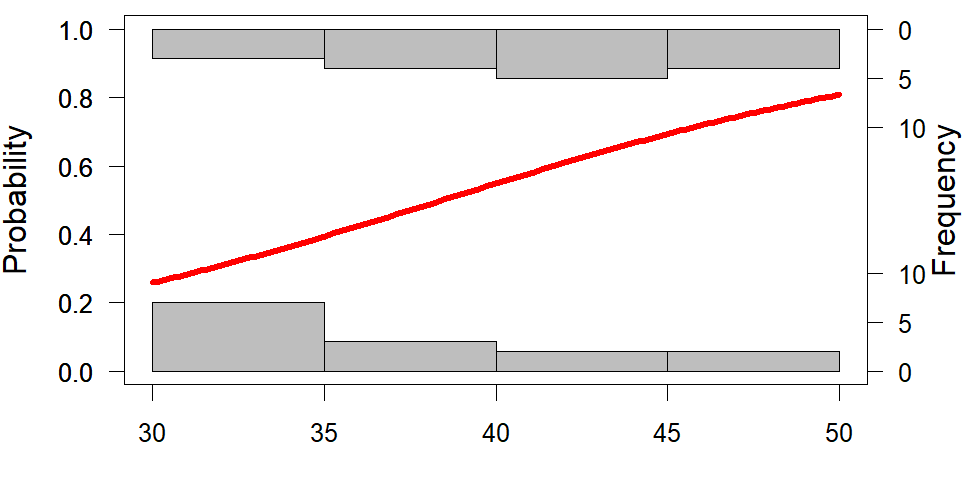
Null deviance: 41.455 on 29 degrees of freedom

Residual deviance: 37.465 on 28 degrees of freedom

AIC: 41.465

Number of Fisher Scoring iterations: 4

****

****

**Q.N. 2)** An important challenge in clinical trials is patients who drop out before the trial is completed. This can cost pharmaceutical companies millions of dollars. Can we predict who will drop out of the study early? The link below provides the Age and the Hamilton Rating Depression Scale (HRDS) and whether or not they completed the study (Drop: 1=Yes and 0=No)

[https://media.pearsoncmg.com/aw/aw\_sharpe\_business\_3/datasets/txt/Clinical Trials.txt](https://media.pearsoncmg.com/aw/aw_sharpe_business_3/datasets/txt/Clinical%20Trials.txt)

1. Import the data in R and determine its dimension.
2. The missing values are left blank. Please clean data by removing missing values.
3. Fit a multiple logistic regression model using **Age** and **HDRS** as a predictor variables.
4. What is the predicted dropout probability of a 30-year-old patient with HDRS score of 30?

> #Q2

> Q2 <- read.table("https://media.pearsoncmg.com/aw/aw\_sharpe\_business\_3/datasets/txt/Clinical%20Trials.txt", header = T, sep = "\t", na.strings = " ")

> head(Q2, 15)

DRP AGE HD2114

1 0 35 22

2 0 48 15

3 0 21 26

4 0 30 15

5 0 29 22

6 1 46 22

7 0 26 21

8 1 27 15

9 0 28 23

10 0 27 19

11 0 24 21

12 1 31 NA

13 0 19 14

14 1 44 12

15 1 23 26

> dim(Q2)

[1] 428 3

> names(Q2)

[1] "DRP" "AGE" "HD2114"

> data = na.omit(Q2)

> dim(data)

[1] 400 3

> attach(data)

> model2 <- glm(DRP~AGE+HD2114, family = binomial, data = data)

> model2

Call: glm(formula = DRP ~ AGE + HD2114, family = binomial, data = data)

Coefficients:

(Intercept) AGE HD2114

-0.44197 -0.03790 0.04682

Degrees of Freedom: 399 Total (i.e. Null); 397 Residual

Null Deviance: 464.6

Residual Deviance: 445.8 AIC: 451.8

> cat("Logisitc Regression Fitted Model:

+ DRP = [1+exp(0.44197+0.03790\*AGE - 0.04682\*HD2114)]^-1")

Logisitc Regression Fitted Model:

DRP = [1+exp(0.44197+0.03790\*AGE - 0.04682\*HD2114)]^-1

> summary(model2)

Call:

glm(formula = DRP ~ AGE + HD2114, family = binomial, data = data)

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -0.44197 0.48827 -0.905 0.365370

AGE -0.03790 0.01151 -3.293 0.000992 \*\*\*

HD2114 0.04682 0.01590 2.944 0.003241 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 464.61 on 399 degrees of freedom

Residual deviance: 445.80 on 397 degrees of freedom

AIC: 451.8

Number of Fisher Scoring iterations: 4

> predict(model2, data.frame(AGE = 30, HD2114 = 30), type = "resp")

1

0.4564631

> cat("Probablility of Success = 45.643%")

Probablility of Success = 45.643%