Homework 05

Ya-Chen Lin (Lisa)

2015 M11 5

Question 01

```
library("lubridate")
```

1.

```
haart <- read.csv("https://raw.githubusercontent.com/fonnesbeck/Bios6301/master/datasets/haart.csv")
haart[,'init.date'] <- as.Date(haart[,'init.date'], format="%m/%d/%y")
year <- format(haart$init.date,'%Y')
table(year)
```

```
## year
## 1998 2000 2001 2002 2003 2004 2005 2006 2007
##    1    5   17   60  270  292  207  104   44
```

2.

```
haart[,'date.death'] <- as.Date(haart[,'date.death'], format="%m/%d/%y")
death <- which(haart$death == 1)
indi <- c()
for(i in seq_along(death)){
  if(haart$date.death[death[i]] - haart$init.date[death[i]] <= 365){
    indi <- c(indi, 1)
  }else{
    indi <- c(indi, 0)
  }
}
sum(indi == 1)
```

```
## [1] 92
```

3.

```
haart[,'init.date'] <- as.Date(haart[,'init.date'], format="%m/%d/%y")
haart[,'date.death'] <- as.Date(haart[,'date.death'], format="%m/%d/%y")
haart[,'last.visit'] <- as.Date(haart[,'last.visit'], format="%m/%d/%y")
date.diff <- rep(NA, length(haart[,'init.date']))
for(i in seq_along(haart[,'init.date'])){
  if(is.na(haart[i,'last.visit']) == FALSE){
    date.diff[i] <- as.numeric(haart[i,'last.visit']-haart[i,'init.date'])
  }else{
    date.diff[i] <- as.numeric(haart[i,'date.death'] -haart[i,'init.date'])
  }
```

```
  if(date.diff[i] > 365){
    date.diff[i] <- 365
  }
}
quantile(date.diff)
```

```
##     0%    25%    50%    75%   100%
##   0.00 320.75 365.00 365.00 365.00
```

4.

```
haart[,'last.visit'] <- as.Date(haart[,'last.visit'], format="%m/%d/%y")
loss.follow.up <- rep(NA, length(haart[,'init.date']))
for(i in seq_along(haart[,'init.date'])){
  if(haart[i,'death'] != 1 && haart[i,'last.visit'] -haart[i,'init.date'] <= 365){
    loss.follow.up[i] <- TRUE
  }else{
    loss.follow.up[i] <- FALSE
  }
}
sum(loss.follow.up == TRUE)
```

```
## [1] 173
```

We can see that there are 173 records lost follow-ups.
5.

```
reg_list <- strsplit(as.character(haart[,'init.reg']), ',')
all_drugs <- unique(unlist(reg_list))
reg_drugs <- matrix(nrow=nrow(haart), ncol=length(all_drugs))
for(i in seq_along(all_drugs)){
  reg_drugs[,i] <- sapply(reg_list, function(x) all_drugs[i] %in% x)
}
colnames(reg_drugs) <- all_drugs
haart <- cbind(haart,reg_drugs)
usage <- rep(NA, length(all_drugs))
for(i in seq_along(all_drugs)){
  usage[i] <- sum(reg_drugs[,i] == TRUE)
}
all_drugs[which(usage > 100)]
```

```
## [1] "3TC" "AZT" "EFV" "NVP" "D4T"
```

6.

```
haart2 <- read.csv("https://raw.githubusercontent.com/fonnesbeck/Bios6301/master/datasets/haart2.csv")
reg_list <- strsplit(as.character(haart[,'init.reg']), ',')
all_drugs <- unique(unlist(reg_list))
reg_list2 <- strsplit(as.character(haart2[,'init.reg']), ',')
reg_drugs2 <- matrix(nrow=nrow(haart2), ncol=length(all_drugs))
```

```
for(i in seq_along(all_drugs)){
  reg_drugs2[,i] <- sapply(reg_list2, function(x) all_drugs[i] %in% x)
}
colnames(reg_drugs2) <- all_drugs
haart2 <- cbind(haart2,reg_drugs2)
newdataframe <- rbind(haart, haart2)
newdataframe[1:5,]
```

```
##   male age aids cd4baseline logvl  weight hemoglobin    init.reg
## 1    1  25    0          NA    NA      NA         NA 3TC,AZT,EFV
## 2    1  49    0         143    NA 58.0608         11 3TC,AZT,EFV
## 3    1  42    1         102    NA 48.0816          1 3TC,AZT,EFV
## 4    0  33    0         107    NA 46.0000         NA 3TC,AZT,NVP
## 5    1  27    0          52     4      NA         NA 3TC,D4T,EFV
##    init.date last.visit death date.death  3TC   AZT   EFV   NVP   D4T
## 1 2003-07-01 2007-02-26     0       <NA> TRUE  TRUE  TRUE FALSE FALSE
## 2 2004-11-23 2008-02-22     0       <NA> TRUE  TRUE  TRUE FALSE FALSE
## 3 2003-04-30 2005-11-21     1 2006-01-11 TRUE  TRUE  TRUE FALSE FALSE
## 4 2006-03-25 2006-05-05     1 2006-05-07 TRUE  TRUE FALSE  TRUE FALSE
## 5 2004-09-01 2007-11-13     0       <NA> TRUE FALSE  TRUE FALSE  TRUE
##     ABC   DDI   IDV   LPV   RTV   SQV   FTC   TDF   DDC   NFV   T20   ATV
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 2 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 3 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 4 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 5 FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##     FPV
## 1 FALSE
## 2 FALSE
## 3 FALSE
## 4 FALSE
## 5 FALSE
```

```
newdataframe[1000:1004,]
```

```
##      male      age aids cd4baseline    logvl  weight hemoglobin
## 1000    0 40.00000    1         131       NA 46.2672          8
## 1001    0 27.00000    0         232       NA      NA         NA
## 1002    1 38.72142    0         170       NA 84.0000         NA
## 1003    1 23.00000   NA         154 3.995635 65.5000         14
## 1004    0 31.00000    0         236       NA 45.8136         NA
##         init.reg  init.date last.visit death date.death  3TC   AZT   EFV
## 1000 3TC,D4T,NVP 2003-07-03 2008-02-29     0       <NA> TRUE FALSE FALSE
## 1001 3TC,AZT,NVP 0012-01-03 0001-05-04     0       <NA> TRUE  TRUE FALSE
## 1002 3TC,AZT,NVP       <NA>       <NA>     0       <NA> TRUE  TRUE FALSE
## 1003 3TC,DDI,EFV       <NA>       <NA>     0       <NA> TRUE FALSE  TRUE
## 1004 3TC,D4T,NVP 0012-03-03 0010-11-07     0       <NA> TRUE FALSE FALSE
##        NVP   D4T   ABC   DDI   IDV   LPV   RTV   SQV   FTC   TDF   DDC
## 1000  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1001  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1002  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1003 FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## 1004  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
##        NFV   T20   ATV   FPV
## 1000 FALSE FALSE FALSE FALSE
## 1001 FALSE FALSE FALSE FALSE
## 1002 FALSE FALSE FALSE FALSE
## 1003 FALSE FALSE FALSE FALSE
## 1004 FALSE FALSE FALSE FALSE
```

Question 02

```
haart <- read.csv("https://raw.githubusercontent.com/fonnesbeck/Bios6301/master/datasets/haart.csv")
data <- haart[,c('cd4baseline', 'weight', 'hemoglobin','death')]
data <- data[complete.cases(data),]
# Logistic function
logistic <- function(x) 1 / (1 + exp(-x))
x <- data[1:3]
y <- data[4]

estimate_logistic <- function(x, y, MAX_ITER=10) {

  n <- dim(x)[1]
  k <- dim(x)[2]

  x <- as.matrix(cbind(rep(1, n), x))
  y <- as.matrix(y)

  # Initialize fitting parameters
  theta <- rep(0, k+1)

  J <- rep(0, MAX_ITER)

  for (i in 1:MAX_ITER) {

    # Calculate linear predictor
    z <- x %*% theta
    # Apply logit function
    h <- logistic(z)

    # Calculate gradient
    grad <- t((1/n)*x) %*% as.matrix(h - y)
    # Calculate Hessian
    H <- t((1/n)*x) %*% diag(array(h)) %*% diag(array(1-h)) %*% x

    # Calculate log likelihood
    J[i] <- (1/n) %*% sum(-y * log(h) - (1-y) * log(1-h))

    # Newton's method
    theta <- theta - solve(H) %*% grad
  }

  return(theta)
}
```

```
estimate_logistic(x,y)
```

```
##                      [,1]
## rep(1, n)    3.576411744
## cd4baseline  0.002092582
## weight      -0.046210552
## hemoglobin  -0.350642786
```

```
g <- glm(death ~ cd4baseline+ weight+hemoglobin, data=haart,family=binomial(logit))
print(g$coefficients)
```

```
##  (Intercept)  cd4baseline       weight   hemoglobin
##  3.576411744  0.002092582 -0.046210552 -0.350642786
```

Question 03

```
data <- read.csv("https://raw.githubusercontent.com/lisa8191/Bios6301/master/datasets/addr.txt", sep=''
data <- data.frame(data)
newdata <- matrix(NA, ncol=7, nrow=43)
#lastname, firstname, streetno, streetname, city, state, zip.
colnames(newdata) <- c("lastname", "firstname", "streetno", "streetname", "city", "state", "zip")
#lastname
for(i in 1:43){
  newdata[i,1] <- as.character(data[i,1])
}
#streetno
a <- grep("[0-9]",data[,3])
for(i in seq_along(a)){
  newdata[a[i],3] <- as.character(data[a[i],3])
}
b <- grep("[0-9]",data[,4])
for(i in seq_along(b)){
  newdata[b[i],3] <- as.character(data[b[i],4])
}
#zip
b <- grep("[0-9]",data[,8])
for(i in seq_along(b)){
  newdata[b[i],7] <- as.character(data[b[i],8])
}
c <- grep("[0-9]",data[,9])
for(i in seq_along(c)){
  newdata[c[i],7] <- as.character(data[c[i],9])
}
d <- grep("[0-9]",data[,10])
for(i in seq_along(d)){
  newdata[d[i],7] <- as.character(data[d[i],10])
}
#State
s7 <- grep("[A-Z]{2}", data[,7])
for(i in seq_along(s7)){
```

```
  newdata[s7[i],6] <- as.character(data[s7[i],7])
}
s8 <- grep("[A-Z]{2}", data[,8])
for(i in seq_along(s8)){
  newdata[s8[i],6] <- as.character(data[s8[i],8])
}
s9 <- grep("[A-Z]{2}", data[,9])
for(i in seq_along(s9)){
  newdata[s9[i],6] <- as.character(data[s9[i],9])
}
s10 <- grep("[A-Z]{2}", data[,10])
for(i in seq_along(s10)){
  newdata[s10[i],6] <- as.character(data[s10[i],10])
}
#city
b<- grep("Ave.|Rd.|St.|Ln|Blvd|Rd", data[,5])
c<- grep("Ave.|Rd.|St.|Ln|Blvd|Rd", data[,6])
d<- grep("Ave.|Rd.|St.|Ln|Blvd|Rd", data[,7])
wms7 <- which(data[,7] == "Wms.")
wms8 <- which(data[,8] == "Wms.")
for(i in seq_along(b)){
  newdata[b[i],5] <- as.character(data[b[i],6])
}
for(i in seq_along(c)){
  newdata[c[i],5] <- as.character(data[c[i],7])
}
for(i in seq_along(d)){
  newdata[d[i],5] <- as.character(data[d[i],8])
}
for(i in seq_along(wms7)){
  newdata[wms7[i],5] <- paste(as.character(data[wms7[i],7]),as.character(data[wms7[i],8]))
}
for(i in seq_along(wms8)){
  newdata[wms8[i],5] <- paste(as.character(data[wms8[i],8]),as.character(data[wms8[i],9]))
}
#first.name
a <- grep("[A-Z]", data[,3])
b <- grep("^[^A-Z]", data[,3])
for(i in seq_along(a)){
  newdata[a[i],2] <- paste(as.character(data[a[i],2]),as.character(data[a[i],3]))
}
for(i in seq_along(b)){
  newdata[b[i],2] <- as.character(data[b[i],2])
}
#street.name
a <- grep("[A-Z]", data[,3])#4th row=number
b<- grep("Ave.|Rd.|St.|Ln|Blvd|Rd", data[,5])
c<- grep("Ave.|Rd.|St.|Ln|Blvd|Rd", data[,6])
at <- a[which(a %in% c == TRUE)]
af <- c[which(c %in% a == FALSE)]
d<- grep("Ave.|Rd.|St.|Ln|Blvd|Rd", data[,7])
att <- a[which(a %in% d == TRUE)]
atf <- d[which(d %in% a == FALSE)]
```

```
for(i in seq_along(b)){
  newdata[b[i],4] <- paste(as.character(data[b[i],4]), as.character(data[b[i],5]))
}
for(i in seq_along(at)){
  newdata[at[i],4] <- paste(as.character(data[at[i],5]), as.character(data[at[i],6]))
}
for(i in seq_along(af)){
  newdata[af[i],4] <- paste(as.character(data[af[i],4]), as.character(data[af[i],5]),as.character(data[a
}
for(i in seq_along(att)){
  newdata[att[i],4] <- paste(as.character(data[att[i],5]), as.character(data[att[i],6]),as.character(da
}
for(i in seq_along(atf)){
  newdata[atf[i],4] <- paste(as.character(data[atf[i],4]), as.character(data[atf[i],5]),as.character(da
}

newdata[20,7] <- as.character(data[21,1])
newdata <- newdata[-21,]
newdata[21,6] <- as.character(data[22,7])
newdata <- data.frame(newdata)
newdata
```

```
##       lastname  firstname streetno           streetname       city state
## 1        Bania  Thomas M.      725    Commonwealth Ave.     Boston    MA
## 2      Barnaby      David      373        W. Geneva St.   Wms. Bay    WI
## 3       Bausch       Judy      373        W. Geneva St.   Wms. Bay    WI
## 4      Bolatto    Alberto      725    Commonwealth Ave.     Boston    MA
## 5    Carlstrom       John      933          E. 56th St.    Chicago    IL
## 6   Chamberlin Richard A.      111           Nowelo St.       Hilo    HI
## 7        Chuss       Dave     2145          Sheridan Rd   Evanston    IL
## 8        Davis      E. J.      933          E. 56th St.    Chicago    IL
## 9        Depoy     Darren      174         W. 18th Ave.   Columbus    OH
## 10     Griffin       Greg     5000          Forbes Ave. Pittsburgh    PA
## 11   Halvorsen       Nils      933          E. 56th St.    Chicago    IL
## 12      Harper         Al      373        W. Geneva St.   Wms. Bay    WI
## 13       Huang     Maohai      725 W. Commonwealth Ave.     Boston    MA
## 14     Ingalls   James G.      725 W. Commonwealth Ave.     Boston    MA
## 15     Jackson   James M.      725 W. Commonwealth Ave.     Boston    MA
## 16     Knudsen      Scott      373        W. Geneva St.   Wms. Bay    WI
## 17       Kovac       John     5640        S. Ellis Ave.    Chicago    IL
## 18   Landsberg      Randy     5640        S. Ellis Ave.    Chicago    IL
## 19          Lo  Kwok-Yung     1002         W. Green St.     Urbana    IL
## 20 Loewenstein  Robert F.      373        W. Geneva St.   Wms. Bay    WI
## 21       Lynch       John     4201          Wilson Blvd  Arlington    VA
## 22     Martini       Paul      174         W. 18th Ave.   Columbus    OH
## 23       Meyer    Stephan      933          E. 56th St.    Chicago    IL
## 24      Mrozek       Fred      373        W. Geneva St.   Wms. Bay    WI
## 25     Newcomb       Matt     5000          Forbes Ave. Pittsburgh    PA
## 26       Novak      Giles     2145          Sheridan Rd   Evanston    IL
## 27      Odalen      Nancy      373        W. Geneva St.   Wms. Bay    WI
## 28      Pernic       Dave      373        W. Geneva St.   Wms. Bay    WI
## 29      Pernic        Bob      373        W. Geneva St.   Wms. Bay    WI
## 30    Peterson    Jeffrey     5000          Forbes Ave. Pittsburgh    PA
```

```
## 31       Pryke       Clem      933          E. 56th St.    Chicago    IL
## 32      Rebull      Luisa     5640       S. Ellis Ave.    Chicago    IL
## 33   Renbarger     Thomas     2145          Sheridan Rd   Evanston    IL
## 34     Rottman        Joe     8730 W. Mountain View Ln  Littleton    CO
## 35   Schartman      Ethan      933          E. 56th St.    Chicago    IL
## 36       Spotz        Bob      373       W. Geneva St.   Wms. Bay    WI
## 37       Thoma       Mark      373       W. Geneva St.   Wms. Bay    WI
## 38      Walker      Chris      933       N. Cherry St.     Tucson    AZ
## 39      Wehrer     Cheryl     5000          Forbes Ave. Pittsburgh    PA
## 40       Wirth      Jesse      373       W. Geneva St.   Wms. Bay    WI
## 41      Wright       Greg      791 Holmdel-Keyport Rd.    Holmdel    NY
## 42     Zingale    Michael     5640       S. Ellis Ave.    Chicago    IL
##           zip
## 1       02215
## 2       53191
## 3       53191
## 4       02215
## 5       60637
## 6       96720
## 7  60208-3112
## 8       60637
## 9       43210
## 10      15213
## 11      60637
## 12      53191
## 13      02215
## 14      02215
## 15      02215
## 16      53191
## 17      60637
## 18      60637
## 19      61801
## 20      53191
## 21      22230
## 22      43210
## 23      60637
## 24      53191
## 25      15213
## 26 60208-3112
## 27      53191
## 28      53191
## 29      53191
## 30      15213
## 31      60637
## 32      60637
## 33 60208-3112
## 34      80125
## 35      60637
## 36      53191
## 37      53191
## 38      85721
## 39      15213
## 40      53191
## 41 07733-1988
```

```
## 42       60637
```

Question 04

It seems like when trying to put 'death' as response, the function is reading death as a variable. Thus, it returns error as "death" not found.