Homework 5

Molly Olson

November 8, 2015

# Question 1

*Import the HAART dataset (haart.csv) from the GitHub repository into R, and perform the following manipulations:*

setwd("~/Documents/Vanderbilt\_2015\_Fall/Statistical\_Computing/Bios6301-master/datasets")  
haart <- read.csv("haart.csv", header=TRUE)  
haart2 <- read.csv("haart2.csv",header=TRUE)  
  
head(haart)

## male age aids cd4baseline logvl weight hemoglobin init.reg init.date  
## 1 1 25 0 NA NA NA NA 3TC,AZT,EFV 7/1/03  
## 2 1 49 0 143 NA 58.0608 11 3TC,AZT,EFV 11/23/04  
## 3 1 42 1 102 NA 48.0816 1 3TC,AZT,EFV 4/30/03  
## 4 0 33 0 107 NA 46.0000 NA 3TC,AZT,NVP 3/25/06  
## 5 1 27 0 52 4 NA NA 3TC,D4T,EFV 9/1/04  
## 6 0 34 0 157 NA 54.8856 NA 3TC,AZT,NVP 12/2/03  
## last.visit death date.death  
## 1 2/26/07 0 <NA>  
## 2 2/22/08 0 <NA>  
## 3 11/21/05 1 1/11/06  
## 4 5/5/06 1 5/7/06  
## 5 11/13/07 0 <NA>  
## 6 2/28/08 0 <NA>

1. *Convert date columns into a usable (for analysis) format. Use the table command to display the counts of the year from init.date.*

class(haart[,9])

## [1] "factor"

#as.Date(haart[,'init.date'], format="%m/%d/%y")  
haart[,9] <- as.Date(haart[,9], format="%m/%d/%y")  
class(haart[,9])

## [1] "Date"

haart[,10] <- as.Date(haart[,10], format="%m/%d/%y")  
haart[,12] <- as.Date(haart[,12], format="%m/%d/%y")

require(lubridate)

## Loading required package: lubridate

table(year(haart[,'init.date']))

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

1. *Create an indicator variable (one which takes the values 0 or 1 only) to represent death within 1 year of the initial visit. How many observations died in year 1?*

#I tested that I was getting the correct results, with NA's  
haart[,'deathWithinYear'] <- rep(0,1000)  
for(i in 1:length(haart[,'deathWithinYear'])){  
 ifelse(as.numeric(difftime(haart[i,12], haart[i,9], units='days')) <= 365, haart[i,'deathWithinYear'] <- 1, haart[i,'deathWithinYear'] <- 0)  
}  
#head(haart)  
  
sum(haart[,'deathWithinYear'],na.rm=TRUE)

## [1] 92

#There are 92 observations that died within 1 year.

1. *Use the init.date, last.visit and death.date columns to calculate a followup time (in days), which is the difference between the first and either the last visit or a death event (whichever comes first). If these times are longer than 1 year, censor them (this means if the value is above 365, set followup to 365). Print the quantile for this new variable.*

haart[,'followup.time'] <- rep(0,1000)  
  
for(i in 1:length(haart[,'followup.time'])){  
 last.init.diff <- NULL  
 death.init.diff <- NULL  
   
 if(!is.na(haart[i,'last.visit'])) death.init.diff <- difftime(haart[i,'last.visit'],haart[i,'init.date'],units='days')  
 if(!is.na(haart[i,'date.death'])) last.init.diff <- difftime(haart[i,'date.death'],haart[i,'init.date'],units='days')  
   
haart[i,'followup.time'] <- min(365,last.init.diff,death.init.diff)  
}  
#head(haart)  
  
quantile(haart[,'followup.time'])

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

1. *Create another indicator variable representing loss to followup; this means the observation is not known to be dead but does not have any followup visits after the first year. How many records are lost-to-followup?*

haart[,'loss.to.followup'] <- rep(0,1000)  
  
#if not dead and followup is within 1 year of init.date   
for(i in 1:length(haart[,'loss.to.followup'])){  
 if(!is.na(haart[i,'last.visit'])) dif <- difftime(haart[i,'last.visit'],haart[i,'init.date'],units='days')  
   
 if(dif <= 365 & haart[i,'death'] == 0){  
 haart[i,'loss.to.followup'] <- 1  
 }  
 }  
 head(haart)

## 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  
## 6 0 34 0 157 NA 54.8856 NA 3TC,AZT,NVP  
## init.date last.visit death date.death deathWithinYear followup.time  
## 1 2003-07-01 2007-02-26 0 <NA> 0 365  
## 2 2004-11-23 2008-02-22 0 <NA> 0 365  
## 3 2003-04-30 2005-11-21 1 2006-01-11 0 365  
## 4 2006-03-25 2006-05-05 1 2006-05-07 1 41  
## 5 2004-09-01 2007-11-13 0 <NA> 0 365  
## 6 2003-12-02 2008-02-28 0 <NA> 0 365  
## loss.to.followup  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 0  
## 6 0

sum(haart[,'loss.to.followup'])

## [1] 173

#173 observations lost to followup

5.*Recall our work in class, which separated the init.reg field into a set of indicator variables, one for each unique drug. Create these fields and append them to the database as new columns. Which drug regimen are found over 100 times?*

reg\_list <- strsplit(as.character(haart[,'init.reg']), ',')  
head(sapply(reg\_list, function(x) 'D4T' %in% x))

## [1] FALSE FALSE FALSE FALSE TRUE FALSE

all\_drugs <- unique(unlist(reg\_list))  
reg\_drugs <- matrix(nrow=nrow(haart), ncol=length(all\_drugs))  
for(i in seq\_along(all\_drugs)){  
 # + makes this 1/0 instead of T/F  
 reg\_drugs[,i] <- +sapply(reg\_list, function(x) all\_drugs[i] %in% x)  
}  
colnames(reg\_drugs) <- all\_drugs  
haart <- cbind(haart, reg\_drugs)  
head(haart)

## 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  
## 6 0 34 0 157 NA 54.8856 NA 3TC,AZT,NVP  
## init.date last.visit death date.death deathWithinYear followup.time  
## 1 2003-07-01 2007-02-26 0 <NA> 0 365  
## 2 2004-11-23 2008-02-22 0 <NA> 0 365  
## 3 2003-04-30 2005-11-21 1 2006-01-11 0 365  
## 4 2006-03-25 2006-05-05 1 2006-05-07 1 41  
## 5 2004-09-01 2007-11-13 0 <NA> 0 365  
## 6 2003-12-02 2008-02-28 0 <NA> 0 365  
## loss.to.followup 3TC AZT EFV NVP D4T ABC DDI IDV LPV RTV SQV FTC TDF DDC  
## 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 2 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 3 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 4 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## 5 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0  
## 6 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## NFV T20 ATV FPV  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0

#16:33 are the drug columns in haart  
for(i in 16:33){  
 sum <- sum(haart[,i])  
 if(sum > 100){  
 print(colnames(haart)[i])  
 }  
}

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

1. *The dataset haart2.csv contains a few additional observations for the same study. Import these and append them to your master dataset (if you were smart about how you coded the previous steps, cleaning the additional observations should be easy!). Show the first five records and the last five records of the complete (and clean) data set*

haart2[,9] <- as.Date(haart2[,9], format="%m/%d/%y")  
haart2[,10] <- as.Date(haart2[,10], format="%m/%d/%y")  
haart2[,12] <- as.Date(haart2[,12], format="%m/%d/%y")

haart2[,'deathWithinYear'] <- rep(0,4)  
for(i in 1:length(haart2[,'deathWithinYear'])){  
 ifelse(as.numeric(difftime(haart2[i,12], haart2[i,9], units='days')) <= 365, haart2[i,'deathWithinYear'] <- 1, haart2[i,'deathWithinYear'] <- 0)  
}

haart2[,'followup.time'] <- rep(0,4)  
  
for(i in 1:length(haart2[,'followup.time'])){  
 last.init.diff <- NULL  
 death.init.diff <- NULL  
   
 if(!is.na(haart2[i,'last.visit'])) death.init.diff <- difftime(haart2[i,'last.visit'],haart2[i,'init.date'],units='days')  
 if(!is.na(haart2[i,'date.death'])) last.init.diff <- difftime(haart2[i,'date.death'],haart2[i,'init.date'],units='days')  
   
haart2[i,'followup.time'] <- min(365,last.init.diff,death.init.diff)  
}

haart2[,'loss.to.followup'] <- rep(0,4)  
  
#if not dead and followup is within 1 year of init.date   
for(i in 1:length(haart2[,'loss.to.followup'])){  
 if(!is.na(haart2[i,'last.visit'])) dif <- difftime(haart2[i,'last.visit'],haart2[i,'init.date'],units='days')  
   
 if(dif <= 365 & haart2[i,'death'] == 0){  
 haart2[i,'loss.to.followup'] <- 1  
 }  
 }

reg\_list <- strsplit(as.character(haart2[,'init.reg']), ',')  
head(sapply(reg\_list, function(x) 'D4T' %in% x))

## [1] FALSE FALSE FALSE TRUE

reg\_drugs <- matrix(nrow=nrow(haart2), ncol=length(all\_drugs))  
for(i in seq\_along(all\_drugs)){  
 # + makes this 1/0 instead of T/F  
 reg\_drugs[,i] <- +sapply(reg\_list, function(x) all\_drugs[i] %in% x)  
}  
colnames(reg\_drugs) <- all\_drugs  
haart2 <- cbind(haart2, reg\_drugs)

head(haart)

## 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  
## 6 0 34 0 157 NA 54.8856 NA 3TC,AZT,NVP  
## init.date last.visit death date.death deathWithinYear followup.time  
## 1 2003-07-01 2007-02-26 0 <NA> 0 365  
## 2 2004-11-23 2008-02-22 0 <NA> 0 365  
## 3 2003-04-30 2005-11-21 1 2006-01-11 0 365  
## 4 2006-03-25 2006-05-05 1 2006-05-07 1 41  
## 5 2004-09-01 2007-11-13 0 <NA> 0 365  
## 6 2003-12-02 2008-02-28 0 <NA> 0 365  
## loss.to.followup 3TC AZT EFV NVP D4T ABC DDI IDV LPV RTV SQV FTC TDF DDC  
## 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 2 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 3 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 4 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## 5 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0  
## 6 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## NFV T20 ATV FPV  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0

head(haart2)

## male age aids cd4baseline logvl weight hemoglobin init.reg  
## 1 0 27.00000 0 232 NA NA NA 3TC,AZT,NVP  
## 2 1 38.72142 0 170 NA 84.0000 NA 3TC,AZT,NVP  
## 3 1 23.00000 NA 154 3.995635 65.5000 14 3TC,DDI,EFV  
## 4 0 31.00000 0 236 NA 45.8136 NA 3TC,D4T,NVP  
## init.date last.visit death date.death deathWithinYear followup.time  
## 1 2003-12-01 2004-01-05 0 <NA> 0 35  
## 2 2002-09-26 2004-03-29 0 <NA> 0 365  
## 3 2007-01-31 2007-04-16 0 <NA> 0 75  
## 4 2003-12-03 2007-10-11 0 <NA> 0 365  
## loss.to.followup 3TC AZT EFV NVP D4T ABC DDI IDV LPV RTV SQV FTC TDF DDC  
## 1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## 2 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## 3 1 1 0 1 0 0 0 1 0 0 0 0 0 0 0  
## 4 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0  
## NFV T20 ATV FPV  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0

haart3 <- rbind(haart,haart2)

head(haart3)

## 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  
## 6 0 34 0 157 NA 54.8856 NA 3TC,AZT,NVP  
## init.date last.visit death date.death deathWithinYear followup.time  
## 1 2003-07-01 2007-02-26 0 <NA> 0 365  
## 2 2004-11-23 2008-02-22 0 <NA> 0 365  
## 3 2003-04-30 2005-11-21 1 2006-01-11 0 365  
## 4 2006-03-25 2006-05-05 1 2006-05-07 1 41  
## 5 2004-09-01 2007-11-13 0 <NA> 0 365  
## 6 2003-12-02 2008-02-28 0 <NA> 0 365  
## loss.to.followup 3TC AZT EFV NVP D4T ABC DDI IDV LPV RTV SQV FTC TDF DDC  
## 1 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 2 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 3 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0  
## 4 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## 5 0 1 0 1 0 1 0 0 0 0 0 0 0 0 0  
## 6 0 1 1 0 1 0 0 0 0 0 0 0 0 0 0  
## NFV T20 ATV FPV  
## 1 0 0 0 0  
## 2 0 0 0 0  
## 3 0 0 0 0  
## 4 0 0 0 0  
## 5 0 0 0 0  
## 6 0 0 0 0

tail(haart3)

## male age aids cd4baseline logvl weight hemoglobin  
## 999 0 31.00000 0 102 NA 61.6896 11  
## 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 deathWithinYear  
## 999 3TC,AZT,NVP 2003-05-22 2008-03-07 0 <NA> 0  
## 1000 3TC,D4T,NVP 2003-07-03 2008-02-29 0 <NA> 0  
## 1001 3TC,AZT,NVP 2003-12-01 2004-01-05 0 <NA> 0  
## 1002 3TC,AZT,NVP 2002-09-26 2004-03-29 0 <NA> 0  
## 1003 3TC,DDI,EFV 2007-01-31 2007-04-16 0 <NA> 0  
## 1004 3TC,D4T,NVP 2003-12-03 2007-10-11 0 <NA> 0  
## followup.time loss.to.followup 3TC AZT EFV NVP D4T ABC DDI IDV LPV  
## 999 365 0 1 1 0 1 0 0 0 0 0  
## 1000 365 0 1 0 0 1 1 0 0 0 0  
## 1001 35 1 1 1 0 1 0 0 0 0 0  
## 1002 365 0 1 1 0 1 0 0 0 0 0  
## 1003 75 1 1 0 1 0 0 0 1 0 0  
## 1004 365 0 1 0 0 1 1 0 0 0 0  
## RTV SQV FTC TDF DDC NFV T20 ATV FPV  
## 999 0 0 0 0 0 0 0 0 0  
## 1000 0 0 0 0 0 0 0 0 0  
## 1001 0 0 0 0 0 0 0 0 0  
## 1002 0 0 0 0 0 0 0 0 0  
## 1003 0 0 0 0 0 0 0 0 0  
## 1004 0 0 0 0 0 0 0 0 0

# Exercise 2

*Obtain the code for using Newton's Method to estimate logistic regression parameters (logistic.r) and modify it to predict death from weight, hemoglobin and cd4baseline in the HAART dataset. Use complete cases only. Report the estimates for each parameter, including the intercept.*

*Note: The original script logistic\_debug.r is in the exercises folder. It needs modification, specifically, the logistic function should be defined:*