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:

```
#read in the data
library(knitr)
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"
```

```
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")
```

```r
require(lubridate)
```

```
## Loading required package: lubridate
```

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

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

2. *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?*

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

sum(haart[,'deathWithinYear'],na.rm=TRUE) #sum how many 'successes'
```

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

```r
 #There are 92 observations that died within 1 year.
```

3. *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.*

```r
haart[,'followup.time'] <- rep(0,1000)

for(i in 1:length(haart[,'followup.time'])){
  last.init.diff <- NULL #initialize
  death.init.diff <- NULL

  if(!is.na(haart[i,'last.visit'])) death.init.diff <-  difftime(haart[i,'last.visit'],haart[i,'init.da
  #evaluate the difference in time for last visit and initial visit, assign it to last.init.diff
  if(!is.na(haart[i,'date.death'])) last.init.diff <- difftime(haart[i,'date.death'],haart[i,'init.date
  #evaluate the difference in time for the date of death and initial visit, assign it to death.init

haart[i,'followup.time'] <- min(365,last.init.diff,death.init.diff) #take the minimum value between
#365, difference in death/init and last visit/init.
}
#head(haart)

quantile(haart[,'followup.time'])
```

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

4. *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='da

  if(dif <= 365 & haart[i,'death'] == 0){
    haart[i,'loss.to.followup'] <- 1 #if the difference in last visit and intit date is less than a yea
    # and we don't have record of them dying,then censor them
  }
  }
  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
## 2                0   1   1   1   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
## 4                0   1   1   0   1   0   0   0   0   0   0   0   0   0
## 5                0   1   0   1   0   1   0   0   0   0   0   0   0   0
## 6                0   1   1   0   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
## 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"
```

6. *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*

Doing parts 1-5 for haart2.csv

```
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,'deathWithinYea
}
```

```
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.
  if(!is.na(haart2[i,'date.death'])) last.init.diff <- difftime(haart2[i,'date.death'],haart2[i,'init.da

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=

  if(dif <= 365 & haart2[i,'death'] == 0){
    haart2[i,'loss.to.followup'] <- 1
  }
  }
```

```
reg_list2 <- strsplit(as.character(haart2[,'init.reg']), ',')
head(sapply(reg_list2, function(x) 'D4T' %in% x))
```

```
## [1] FALSE FALSE FALSE  TRUE
```

```
#we use the all_drugs from part 5 because we need all drug variables from haart, since
#haart2 only has a subset of them
reg_drugs2 <- matrix(nrow=nrow(haart2), ncol=length(all_drugs))
for(i in seq_along(all_drugs)){
  # + makes this 1/0 instead of T/F
  reg_drugs2[,i] <- +sapply(reg_list2, function(x) all_drugs[i] %in% x)
}
colnames(reg_drugs2) <- all_drugs
haart2 <- cbind(haart2, reg_drugs2)
```

```
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:
`logistic <- function(x) 1 / (1 + exp(-x))`

```
predictors <- haart[,c("death","weight","hemoglobin","cd4baseline")] #we only care about these predicto
predictors <- predictors[complete.cases(predictors),] #we only want the complete cases of the data with


x <- predictors[2:4] #this is the predictor variables
y <- predictors[1] #this is the response variable


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

  # Logistic function
  logistic <- function(x) 1 / (1 + exp(-x))
    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
## weight      -0.046210552
## hemoglobin  -0.350642786
## cd4baseline  0.002092582
```

```
# Compare with R's built-in linear regression
#g <- glm(disease ~ test1 + test2, data=data, family=binomial(logit))
#print(g$coefficients)
```

Question 3

Import the addr.txt file from the GitHub repository. This file contains a listing of names and addresses (thanks google). Parse each line to create a data.frame with the following columns: lastname, firstname, streetno, streetname, city, state, zip. Keep middle initials or abbreviated names in the firstname column. Print out the entire data.frame.

```
setwd("~/Documents/Vanderbilt_2015_Fall/Statistical_Computing/Bios6301-master/datasets")
addr <- readLines("addr.txt") #read in the data
addr <- strsplit(addr, split = "  +") #split by two or more spaces so we keep the correct things togeth
```

```
dataframe <- do.call(rbind.data.frame,addr) #put into dataframe
names(dataframe) <- c("LastName","FirstName","Address","City","State","Zip") #names
dataframe[] <- lapply(dataframe, as.character) #strings as charaacters
```

```
library(stringr)
dataframe$StreetNo <- str_split_fixed(dataframe$Address, " ",n=2) #putting street numbers into column

for(i in 1:length(addr)){
  dataframe$StreetName[i] <- str_split_fixed(dataframe$Address[i], " ",n=2)[2] #putting street names in
}

dataframe$Address <- NULL #eliminate address column
```

```
dataframe
```

```
##       LastName  FirstName       City State         Zip StreetNo.1
## 1        Bania  Thomas M.     Boston    MA       02215         725
## 2      Barnaby      David   Wms. Bay    WI       53191         373
## 3       Bausch       Judy   Wms. Bay    WI       53191         373
## 4      Bolatto    Alberto     Boston    MA       02215         725
## 5    Carlstrom       John    Chicago    IL       60637         933
## 6   Chamberlin Richard A.       Hilo    HI       96720         111
## 7        Chuss       Dave   Evanston    IL 60208-3112        2145
## 8        Davis      E. J.    Chicago    IL       60637         933
## 9        Depoy     Darren   Columbus    OH       43210         174
## 10     Griffin       Greg Pittsburgh    PA       15213        5000
## 11   Halvorsen       Nils    Chicago    IL       60637         933
## 12      Harper         Al   Wms. Bay    WI       53191         373
## 13       Huang     Maohai     Boston    MA       02215         725
## 14     Ingalls   James G.     Boston    MA       02215         725
## 15     Jackson   James M.     Boston    MA       02215         725
## 16     Knudsen      Scott   Wms. Bay    WI       53191         373
## 17       Kovac       John    Chicago    IL       60637        5640
## 18   Landsberg      Randy    Chicago    IL       60637        5640
## 19          Lo  Kwok-Yung     Urbana    IL       61801        1002
## 20 Loewenstein  Robert F.   Wms. Bay    WI       53191         373
## 21       Lynch       John  Arlington    VA       22230        4201
## 22     Martini       Paul   Columbus    OH       43210         174
## 23       Meyer    Stephan    Chicago    IL       60637         933
## 24      Mrozek       Fred   Wms. Bay    WI       53191         373
## 25     Newcomb       Matt Pittsburgh    PA       15213        5000
## 26       Novak      Giles   Evanston    IL 60208-3112        2145
## 27      Odalen      Nancy   Wms. Bay    WI       53191         373
## 28      Pernic       Dave   Wms. Bay    WI       53191         373
## 29      Pernic        Bob   Wms. Bay    WI       53191         373
## 30    Peterson    Jeffrey Pittsburgh    PA       15213        5000
## 31       Pryke       Clem    Chicago    IL       60637         933
## 32      Rebull      Luisa    Chicago    IL       60637        5640
## 33   Renbarger     Thomas   Evanston    IL 60208-3112        2145
## 34     Rottman        Joe  Littleton    CO       80125        8730
## 35   Schartman      Ethan    Chicago    IL       60637         933
## 36       Spotz        Bob   Wms. Bay    WI       53191         373
## 37       Thoma       Mark   Wms. Bay    WI       53191         373
## 38      Walker      Chris     Tucson    AZ       85721         933
## 39      Wehrer     Cheryl Pittsburgh    PA       15213        5000
## 40       Wirth      Jesse   Wms. Bay    WI       53191         373
## 41      Wright       Greg    Holmdel    NY 07733-1988         791
## 42     Zingale    Michael    Chicago    IL       60637        5640
##             StreetNo.2          StreetName
## 1    Commonwealth Ave.   Commonwealth Ave.
## 2        W. Geneva St.       W. Geneva St.
## 3        W. Geneva St.       W. Geneva St.
## 4    Commonwealth Ave.   Commonwealth Ave.
## 5          E. 56th St.         E. 56th St.
## 6           Nowelo St.          Nowelo St.
## 7          Sheridan Rd         Sheridan Rd
## 8          E. 56th St.         E. 56th St.
```

```
## 9          W. 18th Ave.         W. 18th Ave.
## 10           Forbes Ave.          Forbes Ave.
## 11           E. 56th St.          E. 56th St.
## 12         W. Geneva St.        W. Geneva St.
## 13 W. Commonwealth Ave. W. Commonwealth Ave.
## 14 W. Commonwealth Ave. W. Commonwealth Ave.
## 15 W. Commonwealth Ave. W. Commonwealth Ave.
## 16         W. Geneva St.        W. Geneva St.
## 17         S. Ellis Ave.        S. Ellis Ave.
## 18         S. Ellis Ave.        S. Ellis Ave.
## 19          W. Green St.         W. Green St.
## 20         W. Geneva St.        W. Geneva St.
## 21           Wilson Blvd          Wilson Blvd
## 22          W. 18th Ave.         W. 18th Ave.
## 23           E. 56th St.          E. 56th St.
## 24         W. Geneva St.        W. Geneva St.
## 25           Forbes Ave.          Forbes Ave.
## 26           Sheridan Rd          Sheridan Rd
## 27         W. Geneva St.        W. Geneva St.
## 28         W. Geneva St.        W. Geneva St.
## 29         W. Geneva St.        W. Geneva St.
## 30           Forbes Ave.          Forbes Ave.
## 31           E. 56th St.          E. 56th St.
## 32         S. Ellis Ave.        S. Ellis Ave.
## 33           Sheridan Rd          Sheridan Rd
## 34  W. Mountain View Ln  W. Mountain View Ln
## 35           E. 56th St.          E. 56th St.
## 36         W. Geneva St.        W. Geneva St.
## 37         W. Geneva St.        W. Geneva St.
## 38         N. Cherry St.        N. Cherry St.
## 39           Forbes Ave.          Forbes Ave.
## 40         W. Geneva St.        W. Geneva St.
## 41  Holmdel-Keyport Rd.  Holmdel-Keyport Rd.
## 42         S. Ellis Ave.        S. Ellis Ave.
```

Question 4

The first argument to most functions that fit linear models are formulas. The following example defines the response variable death and allows the model to incorporate all other variables as terms. . is used to mean all columns not otherwise in the formula.

```
url <- "https://github.com/fonnesbeck/Bios6301/raw/master/datasets/haart.csv"
haart_df <- read.csv(url)[,c('death','weight','hemoglobin','cd4baseline')]
coef(summary(glm(death ~ ., data=haart_df, family=binomial(logit))))
```

```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  3.576411744 1.226870535  2.915069 0.0035561039
## weight      -0.046210552 0.022556001 -2.048703 0.0404911395
## hemoglobin  -0.350642786 0.105064078 -3.337418 0.0008456055
## cd4baseline  0.002092582 0.001811959  1.154872 0.2481427160
```

Now imagine running the above several times, but with a different response and data set each time. Here's a function:

```
myfun <- function(dat, response) {
  form <- as.formula(response ~ .)
  coef(summary(glm(form, data=dat, family=binomial(logit))))
}
```

Unfortunately, it doesn't work. tryCatch is "catching" the error so that this file can be knit to PDF.

```
tryCatch(myfun(haart_df, death), error = function(e) e)
```

```
## <simpleError in eval(expr, envir, enclos): object 'death' not found>
```

What do you think is going on? Consider using debug to trace the problem.

The error given says that `death` is not found. When debugging, it says the object `form` not found. When I call `?as.formula()`, it says that it takes an object as it's argument. When I run `class(response ~ .)`, it says that it is a formula. So, I think that we are not passing in the formula correctly, thus form isn't going to be defined. When doing regression, you also need to specify where the variables are coming from, otherwise R won't know what to do with them. Death is the response, so the response isn't being passed correctly. We need to define which dataset it is coming from.

Bonus points

```
myfun2 <- function(dat, response) {
  dat$response2 <- dat[,response] #calls the column of the dataset that we want the
  #response to be and assigns it to response2 in the data
  coef(summary(glm(response2 ~ ., data=dat, family=binomial(logit)))) #replace the original `form`, wit
}
myfun2(haart_df, 'death')
```

```
## Warning: glm.fit: algorithm did not converge
```

```
##                  Estimate  Std. Error       z value  Pr(>|z|)
## (Intercept) -2.656607e+01 115935.1724 -2.291459e-04 0.9998172
## death        5.313214e+01  69028.4183  7.697140e-04 0.9993859
## weight      -4.499694e-15   1939.0571 -2.320558e-18 1.0000000
## hemoglobin   5.124642e-14   9774.8190  5.242697e-18 1.0000000
## cd4baseline  1.830771e-16    184.0846  9.945271e-19 1.0000000
```