Homework 07 Spring 2019

Daniel Smith

April 14, 2019

## Homework 07 Spring 2019 - DUE April 17, 2019

**NOTE TO DR. HIGGINS: I’m not sure I did any of this correctly. Any and all feedback is greatly appreciated. Thanks, Daniel**

#load everything needed   
library(NHANES)  
library(dplyr)  
library(gmodels)  
library(ROCR)  
library(rpart)  
library(partykit)  
library(tidyverse)  
library(RColorBrewer)  
library(reshape)  
library(plot3D)  
library(parallel)  
library(randomForestSRC)  
library(ggRandomForests)  
library(class)  
library(mosaic)  
library(mice)  
library(ggplot2)  
library(car)  
library(stargazer)  
library(reshape2)

## Course Material to Review

Recall the NHANES dataset that we used in Lesson 12 on March 27, 2019, <https://htmlpreview.github.io/?https://github.com/vhertzb/ml_supervised/blob/master/ML_supervised.html>. And more on supervised learning on April 10, 2019, <https://htmlpreview.github.io/?https://github.com/vhertzb/more-supervised-learning/blob/master/More_Supervised_Learning.html>.

Also review the logistic regression examples in Homework 6 assignment, see <https://htmlpreview.github.io/?https://github.com/melindahiggins2000/N741_Homework06_regression/blob/master/homework6.html>.

## Assignment

In the NHANES dataset there is a discrete variable called Depressed indicating whether each participant had “None”, “Several”, “Majority” or “AlmostAll” days in a month where the pariticpant felt down, depressed or hopeless. You are going to build a set of classifiers for this dependent variable. You may use any (set of) independent variable(s) you like except for the variable callsed DaysMentHlthBad (self-reported days that the participant’s mental health was not good out of 30 days).

Run this R code to get started and create 2 groups that either were depressed “None” versus more than “None” - the new variable is depressedYes.

# add depressedYes to NHANES dataset  
NHANES <- NHANES %>%  
 mutate(depressedYes = Depressed != "None")  
   
# check recoding that "Several" and "Most"  
# are coded as TRUE for depressedYes  
# and "None" are coded FALSE for depressedYes  
NHANES %>%  
 select(Depressed, depressedYes) %>%  
 with(table(Depressed, depressedYes))

## depressedYes  
## Depressed FALSE TRUE  
## None 5246 0  
## Several 0 1009  
## Most 0 418

PROBLEM 1: Run 4 classifier models for depressedYes:

* logistic regression A)Build the Classifier

#summarize the data set   
summary(NHANES)

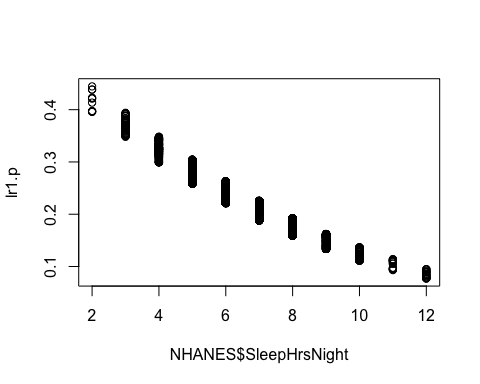
## ID SurveyYr Gender Age   
## Min. :51624 2009\_10:5000 female:5020 Min. : 0.00   
## 1st Qu.:56904 2011\_12:5000 male :4980 1st Qu.:17.00   
## Median :62160 Median :36.00   
## Mean :61945 Mean :36.74   
## 3rd Qu.:67039 3rd Qu.:54.00   
## Max. :71915 Max. :80.00   
##   
## AgeDecade AgeMonths Race1 Race3   
## 40-49 :1398 Min. : 0.0 Black :1197 Asian : 288   
## 0-9 :1391 1st Qu.:199.0 Hispanic: 610 Black : 589   
## 10-19 :1374 Median :418.0 Mexican :1015 Hispanic: 350   
## 20-29 :1356 Mean :420.1 White :6372 Mexican : 480   
## 30-39 :1338 3rd Qu.:624.0 Other : 806 White :3135   
## (Other):2810 Max. :959.0 Other : 158   
## NA's : 333 NA's :5038 NA's :5000   
## Education MaritalStatus HHIncome   
## 8th Grade : 451 Divorced : 707 more 99999 :2220   
## 9 - 11th Grade: 888 LivePartner : 560 75000-99999:1084   
## High School :1517 Married :3945 25000-34999: 958   
## Some College :2267 NeverMarried:1380 35000-44999: 863   
## College Grad :2098 Separated : 183 45000-54999: 784   
## NA's :2779 Widowed : 456 (Other) :3280   
## NA's :2769 NA's : 811   
## HHIncomeMid Poverty HomeRooms HomeOwn   
## Min. : 2500 Min. :0.000 Min. : 1.000 Own :6425   
## 1st Qu.: 30000 1st Qu.:1.240 1st Qu.: 5.000 Rent :3287   
## Median : 50000 Median :2.700 Median : 6.000 Other: 225   
## Mean : 57206 Mean :2.802 Mean : 6.249 NA's : 63   
## 3rd Qu.: 87500 3rd Qu.:4.710 3rd Qu.: 8.000   
## Max. :100000 Max. :5.000 Max. :13.000   
## NA's :811 NA's :726 NA's :69   
## Work Weight Length HeadCirc   
## Looking : 311 Min. : 2.80 Min. : 47.10 Min. :34.20   
## NotWorking:2847 1st Qu.: 56.10 1st Qu.: 75.70 1st Qu.:39.58   
## Working :4613 Median : 72.70 Median : 87.00 Median :41.45   
## NA's :2229 Mean : 70.98 Mean : 85.02 Mean :41.18   
## 3rd Qu.: 88.90 3rd Qu.: 96.10 3rd Qu.:42.92   
## Max. :230.70 Max. :112.20 Max. :45.40   
## NA's :78 NA's :9457 NA's :9912   
## Height BMI BMICatUnder20yrs BMI\_WHO   
## Min. : 83.6 Min. :12.88 UnderWeight: 55 12.0\_18.5 :1277   
## 1st Qu.:156.8 1st Qu.:21.58 NormWeight : 805 18.5\_to\_24.9:2911   
## Median :166.0 Median :25.98 OverWeight : 193 25.0\_to\_29.9:2664   
## Mean :161.9 Mean :26.66 Obese : 221 30.0\_plus :2751   
## 3rd Qu.:174.5 3rd Qu.:30.89 NA's :8726 NA's : 397   
## Max. :200.4 Max. :81.25   
## NA's :353 NA's :366   
## Pulse BPSysAve BPDiaAve BPSys1   
## Min. : 40.00 Min. : 76.0 Min. : 0.00 Min. : 72.0   
## 1st Qu.: 64.00 1st Qu.:106.0 1st Qu.: 61.00 1st Qu.:106.0   
## Median : 72.00 Median :116.0 Median : 69.00 Median :116.0   
## Mean : 73.56 Mean :118.2 Mean : 67.48 Mean :119.1   
## 3rd Qu.: 82.00 3rd Qu.:127.0 3rd Qu.: 76.00 3rd Qu.:128.0   
## Max. :136.00 Max. :226.0 Max. :116.00 Max. :232.0   
## NA's :1437 NA's :1449 NA's :1449 NA's :1763   
## BPDia1 BPSys2 BPDia2 BPSys3   
## Min. : 0.00 Min. : 76.0 Min. : 0.00 Min. : 76.0   
## 1st Qu.: 62.00 1st Qu.:106.0 1st Qu.: 60.00 1st Qu.:106.0   
## Median : 70.00 Median :116.0 Median : 68.00 Median :116.0   
## Mean : 68.28 Mean :118.5 Mean : 67.66 Mean :117.9   
## 3rd Qu.: 76.00 3rd Qu.:128.0 3rd Qu.: 76.00 3rd Qu.:126.0   
## Max. :118.00 Max. :226.0 Max. :118.00 Max. :226.0   
## NA's :1763 NA's :1647 NA's :1647 NA's :1635   
## BPDia3 Testosterone DirectChol TotChol   
## Min. : 0.0 Min. : 0.25 Min. :0.390 Min. : 1.530   
## 1st Qu.: 60.0 1st Qu.: 17.70 1st Qu.:1.090 1st Qu.: 4.110   
## Median : 68.0 Median : 43.82 Median :1.290 Median : 4.780   
## Mean : 67.3 Mean : 197.90 Mean :1.365 Mean : 4.879   
## 3rd Qu.: 76.0 3rd Qu.: 362.41 3rd Qu.:1.580 3rd Qu.: 5.530   
## Max. :116.0 Max. :1795.60 Max. :4.030 Max. :13.650   
## NA's :1635 NA's :5874 NA's :1526 NA's :1526   
## UrineVol1 UrineFlow1 UrineVol2 UrineFlow2   
## Min. : 0.0 Min. : 0.0000 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 50.0 1st Qu.: 0.4030 1st Qu.: 52.0 1st Qu.: 0.475   
## Median : 94.0 Median : 0.6990 Median : 95.0 Median : 0.760   
## Mean :118.5 Mean : 0.9793 Mean :119.7 Mean : 1.149   
## 3rd Qu.:164.0 3rd Qu.: 1.2210 3rd Qu.:171.8 3rd Qu.: 1.513   
## Max. :510.0 Max. :17.1670 Max. :409.0 Max. :13.692   
## NA's :987 NA's :1603 NA's :8522 NA's :8524   
## Diabetes DiabetesAge HealthGen DaysPhysHlthBad   
## No :9098 Min. : 1.00 Excellent: 878 Min. : 0.000   
## Yes : 760 1st Qu.:40.00 Vgood :2508 1st Qu.: 0.000   
## NA's: 142 Median :50.00 Good :2956 Median : 0.000   
## Mean :48.42 Fair :1010 Mean : 3.335   
## 3rd Qu.:58.00 Poor : 187 3rd Qu.: 3.000   
## Max. :80.00 NA's :2461 Max. :30.000   
## NA's :9371 NA's :2468   
## DaysMentHlthBad LittleInterest Depressed nPregnancies   
## Min. : 0.000 None :5103 None :5246 Min. : 1.000   
## 1st Qu.: 0.000 Several:1130 Several:1009 1st Qu.: 2.000   
## Median : 0.000 Most : 434 Most : 418 Median : 3.000   
## Mean : 4.127 NA's :3333 NA's :3327 Mean : 3.027   
## 3rd Qu.: 4.000 3rd Qu.: 4.000   
## Max. :30.000 Max. :32.000   
## NA's :2466 NA's :7396   
## nBabies Age1stBaby SleepHrsNight SleepTrouble  
## Min. : 0.000 Min. :14.00 Min. : 2.000 No :5799   
## 1st Qu.: 2.000 1st Qu.:19.00 1st Qu.: 6.000 Yes :1973   
## Median : 2.000 Median :22.00 Median : 7.000 NA's:2228   
## Mean : 2.457 Mean :22.65 Mean : 6.928   
## 3rd Qu.: 3.000 3rd Qu.:26.00 3rd Qu.: 8.000   
## Max. :12.000 Max. :39.00 Max. :12.000   
## NA's :7584 NA's :8116 NA's :2245   
## PhysActive PhysActiveDays TVHrsDay CompHrsDay   
## No :3677 Min. :1.000 2\_hr :1275 0\_to\_1\_hr:1409   
## Yes :4649 1st Qu.:2.000 1\_hr : 884 0\_hrs :1073   
## NA's:1674 Median :3.000 3\_hr : 836 1\_hr :1030   
## Mean :3.744 0\_to\_1\_hr: 638 2\_hr : 589   
## 3rd Qu.:5.000 More\_4\_hr: 615 3\_hr : 347   
## Max. :7.000 (Other) : 611 (Other) : 415   
## NA's :5337 NA's :5141 NA's :5137   
## TVHrsDayChild CompHrsDayChild Alcohol12PlusYr AlcoholDay   
## Min. :0.000 Min. :0.000 No :1368 Min. : 1.000   
## 1st Qu.:1.000 1st Qu.:0.000 Yes :5212 1st Qu.: 1.000   
## Median :2.000 Median :1.000 NA's:3420 Median : 2.000   
## Mean :1.939 Mean :2.198 Mean : 2.914   
## 3rd Qu.:3.000 3rd Qu.:6.000 3rd Qu.: 3.000   
## Max. :6.000 Max. :6.000 Max. :82.000   
## NA's :9347 NA's :9347 NA's :5086   
## AlcoholYear SmokeNow Smoke100 Smoke100n SmokeAge   
## Min. : 0.0 No :1745 No :4024 Non-Smoker:4024 Min. : 6.00   
## 1st Qu.: 3.0 Yes :1466 Yes :3211 Smoker :3211 1st Qu.:15.00   
## Median : 24.0 NA's:6789 NA's:2765 NA's :2765 Median :17.00   
## Mean : 75.1 Mean :17.83   
## 3rd Qu.:104.0 3rd Qu.:19.00   
## Max. :364.0 Max. :72.00   
## NA's :4078 NA's :6920   
## Marijuana AgeFirstMarij RegularMarij AgeRegMarij HardDrugs   
## No :2049 Min. : 1.00 No :3575 Min. : 5.00 No :4700   
## Yes :2892 1st Qu.:15.00 Yes :1366 1st Qu.:15.00 Yes :1065   
## NA's:5059 Median :16.00 NA's:5059 Median :17.00 NA's:4235   
## Mean :17.02 Mean :17.69   
## 3rd Qu.:19.00 3rd Qu.:19.00   
## Max. :48.00 Max. :52.00   
## NA's :7109 NA's :8634   
## SexEver SexAge SexNumPartnLife SexNumPartYear   
## No : 223 Min. : 9.00 Min. : 0.00 Min. : 0.000   
## Yes :5544 1st Qu.:15.00 1st Qu.: 2.00 1st Qu.: 1.000   
## NA's:4233 Median :17.00 Median : 5.00 Median : 1.000   
## Mean :17.43 Mean : 15.09 Mean : 1.342   
## 3rd Qu.:19.00 3rd Qu.: 12.00 3rd Qu.: 1.000   
## Max. :50.00 Max. :2000.00 Max. :69.000   
## NA's :4460 NA's :4275 NA's :5072   
## SameSex SexOrientation PregnantNow depressedYes   
## No :5353 Bisexual : 119 Yes : 72 Mode :logical   
## Yes : 415 Heterosexual:4638 No :1573 FALSE:5246   
## NA's:4232 Homosexual : 85 Unknown: 51 TRUE :1427   
## NA's :5158 NA's :8304 NA's :3327   
##   
##   
##

#Split the data into a training and test dataset (90/10 split) based on a fix seed  
set.seed(123456)  
lr1 <- glm(depressedYes ~ SleepHrsNight + Age, data=NHANES, family=binomial)  
summary

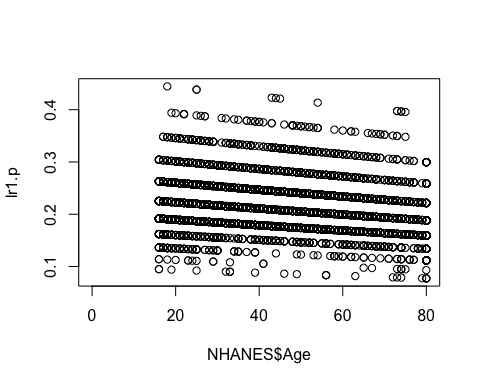
## standardGeneric for "summary" defined from package "base"  
##   
## function (object, ...)   
## standardGeneric("summary")  
## <environment: 0x7f83d713c2e0>  
## Methods may be defined for arguments: object  
## Use showMethods("summary") for currently available ones.

B)Report Effectiveness on NHANES Dataset AND C) Appropriate Visualizations Not a very good model. Sensitivity Is not good, Sepcificity is Acceptable.

#How did LR1 do in prediction?  
lr1.p <- predict(lr1, newdata=NHANES, type = "response")  
#plot for continuous predictor SleepHrsNight  
plot(NHANES$SleepHrsNight, lr1.p) #plot tells us we need a probability of outcome around 0.25.



plot(NHANES$Age, lr1.p)



#Confusion Matrix   
CrossTable(NHANES$depressedYes, lr1.p > 0.25)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 6660   
##   
##   
## | lr1.p > 0.25   
## NHANES$depressedYes | FALSE | TRUE | Row Total |   
## --------------------|-----------|-----------|-----------|  
## FALSE | 4353 | 885 | 5238 |   
## | 4.377 | 18.081 | |   
## | 0.831 | 0.169 | 0.786 |   
## | 0.812 | 0.682 | |   
## | 0.654 | 0.133 | |   
## --------------------|-----------|-----------|-----------|  
## TRUE | 1009 | 413 | 1422 |   
## | 16.122 | 66.601 | |   
## | 0.710 | 0.290 | 0.214 |   
## | 0.188 | 0.318 | |   
## | 0.152 | 0.062 | |   
## --------------------|-----------|-----------|-----------|  
## Column Total | 5362 | 1298 | 6660 |   
## | 0.805 | 0.195 | |   
## --------------------|-----------|-----------|-----------|  
##   
##

#OR we can get TPR and FPR with a 0.25 probability   
#confusion matrix   
t1 <- table(lr1.p > 0.25, NHANES$depressedYes)  
t1 #gives same results as the CrossTable() funciton above.

##   
## FALSE TRUE  
## FALSE 4353 1009  
## TRUE 885 413

#calculate sensitivity   
tpr <- t1[2,2]/(t1[2,2]+t1[1,2])  
tpr #not very good; only 19.5%

## [1] 0.290436

#calculate specificity   
tnr <- t1[1,1]/(t1[1,1]+t1[2,1])  
tnr #Pretty good at 89.7%

## [1] 0.8310424

#Look at Area under the curve   
lr1.pr <- prediction(lr1.p, NHANES$depressedYes)  
lr1.prf <- performance(lr1.pr, measure = "tpr", x.measure= "fpr") #I can't get this to run properly and have spent too much time on it. I can't figure out why I keep getting an error/how to fix it...   
plot(lr1.prf)  
abline(a=0, b=1, col="red")  
#AUC   
auc <- performance(lr1.pr, measure = "auc")  
auc <- auc@y.values[[1]]  
auc #auc "not enough distinct predicitons to compute area under the ROC curve" However, I have a feeling the AUC is not very good for this model.

1. Interpret

#Get ORs from lr1  
exp(coef(lr1))

## (Intercept) SleepHrsNight Age   
## 1.2820980 0.8155201 0.9964828

For everyone one hour increase in the number of sleep per night, the odds of being classified as depressed decreases by 0.1 when controlling for age. Controlling for number of hours of sleep/night, age pracitcally has no effect on the odds of being classified as depressed.

* decision tree

1. Build the Classifier

#Use Logisitc Regression Model from Above   
summary(NHANES)

## ID SurveyYr Gender Age   
## Min. :51624 2009\_10:5000 female:5020 Min. : 0.00   
## 1st Qu.:56904 2011\_12:5000 male :4980 1st Qu.:17.00   
## Median :62160 Median :36.00   
## Mean :61945 Mean :36.74   
## 3rd Qu.:67039 3rd Qu.:54.00   
## Max. :71915 Max. :80.00   
##   
## AgeDecade AgeMonths Race1 Race3   
## 40-49 :1398 Min. : 0.0 Black :1197 Asian : 288   
## 0-9 :1391 1st Qu.:199.0 Hispanic: 610 Black : 589   
## 10-19 :1374 Median :418.0 Mexican :1015 Hispanic: 350   
## 20-29 :1356 Mean :420.1 White :6372 Mexican : 480   
## 30-39 :1338 3rd Qu.:624.0 Other : 806 White :3135   
## (Other):2810 Max. :959.0 Other : 158   
## NA's : 333 NA's :5038 NA's :5000   
## Education MaritalStatus HHIncome   
## 8th Grade : 451 Divorced : 707 more 99999 :2220   
## 9 - 11th Grade: 888 LivePartner : 560 75000-99999:1084   
## High School :1517 Married :3945 25000-34999: 958   
## Some College :2267 NeverMarried:1380 35000-44999: 863   
## College Grad :2098 Separated : 183 45000-54999: 784   
## NA's :2779 Widowed : 456 (Other) :3280   
## NA's :2769 NA's : 811   
## HHIncomeMid Poverty HomeRooms HomeOwn   
## Min. : 2500 Min. :0.000 Min. : 1.000 Own :6425   
## 1st Qu.: 30000 1st Qu.:1.240 1st Qu.: 5.000 Rent :3287   
## Median : 50000 Median :2.700 Median : 6.000 Other: 225   
## Mean : 57206 Mean :2.802 Mean : 6.249 NA's : 63   
## 3rd Qu.: 87500 3rd Qu.:4.710 3rd Qu.: 8.000   
## Max. :100000 Max. :5.000 Max. :13.000   
## NA's :811 NA's :726 NA's :69   
## Work Weight Length HeadCirc   
## Looking : 311 Min. : 2.80 Min. : 47.10 Min. :34.20   
## NotWorking:2847 1st Qu.: 56.10 1st Qu.: 75.70 1st Qu.:39.58   
## Working :4613 Median : 72.70 Median : 87.00 Median :41.45   
## NA's :2229 Mean : 70.98 Mean : 85.02 Mean :41.18   
## 3rd Qu.: 88.90 3rd Qu.: 96.10 3rd Qu.:42.92   
## Max. :230.70 Max. :112.20 Max. :45.40   
## NA's :78 NA's :9457 NA's :9912   
## Height BMI BMICatUnder20yrs BMI\_WHO   
## Min. : 83.6 Min. :12.88 UnderWeight: 55 12.0\_18.5 :1277   
## 1st Qu.:156.8 1st Qu.:21.58 NormWeight : 805 18.5\_to\_24.9:2911   
## Median :166.0 Median :25.98 OverWeight : 193 25.0\_to\_29.9:2664   
## Mean :161.9 Mean :26.66 Obese : 221 30.0\_plus :2751   
## 3rd Qu.:174.5 3rd Qu.:30.89 NA's :8726 NA's : 397   
## Max. :200.4 Max. :81.25   
## NA's :353 NA's :366   
## Pulse BPSysAve BPDiaAve BPSys1   
## Min. : 40.00 Min. : 76.0 Min. : 0.00 Min. : 72.0   
## 1st Qu.: 64.00 1st Qu.:106.0 1st Qu.: 61.00 1st Qu.:106.0   
## Median : 72.00 Median :116.0 Median : 69.00 Median :116.0   
## Mean : 73.56 Mean :118.2 Mean : 67.48 Mean :119.1   
## 3rd Qu.: 82.00 3rd Qu.:127.0 3rd Qu.: 76.00 3rd Qu.:128.0   
## Max. :136.00 Max. :226.0 Max. :116.00 Max. :232.0   
## NA's :1437 NA's :1449 NA's :1449 NA's :1763   
## BPDia1 BPSys2 BPDia2 BPSys3   
## Min. : 0.00 Min. : 76.0 Min. : 0.00 Min. : 76.0   
## 1st Qu.: 62.00 1st Qu.:106.0 1st Qu.: 60.00 1st Qu.:106.0   
## Median : 70.00 Median :116.0 Median : 68.00 Median :116.0   
## Mean : 68.28 Mean :118.5 Mean : 67.66 Mean :117.9   
## 3rd Qu.: 76.00 3rd Qu.:128.0 3rd Qu.: 76.00 3rd Qu.:126.0   
## Max. :118.00 Max. :226.0 Max. :118.00 Max. :226.0   
## NA's :1763 NA's :1647 NA's :1647 NA's :1635   
## BPDia3 Testosterone DirectChol TotChol   
## Min. : 0.0 Min. : 0.25 Min. :0.390 Min. : 1.530   
## 1st Qu.: 60.0 1st Qu.: 17.70 1st Qu.:1.090 1st Qu.: 4.110   
## Median : 68.0 Median : 43.82 Median :1.290 Median : 4.780   
## Mean : 67.3 Mean : 197.90 Mean :1.365 Mean : 4.879   
## 3rd Qu.: 76.0 3rd Qu.: 362.41 3rd Qu.:1.580 3rd Qu.: 5.530   
## Max. :116.0 Max. :1795.60 Max. :4.030 Max. :13.650   
## NA's :1635 NA's :5874 NA's :1526 NA's :1526   
## UrineVol1 UrineFlow1 UrineVol2 UrineFlow2   
## Min. : 0.0 Min. : 0.0000 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 50.0 1st Qu.: 0.4030 1st Qu.: 52.0 1st Qu.: 0.475   
## Median : 94.0 Median : 0.6990 Median : 95.0 Median : 0.760   
## Mean :118.5 Mean : 0.9793 Mean :119.7 Mean : 1.149   
## 3rd Qu.:164.0 3rd Qu.: 1.2210 3rd Qu.:171.8 3rd Qu.: 1.513   
## Max. :510.0 Max. :17.1670 Max. :409.0 Max. :13.692   
## NA's :987 NA's :1603 NA's :8522 NA's :8524   
## Diabetes DiabetesAge HealthGen DaysPhysHlthBad   
## No :9098 Min. : 1.00 Excellent: 878 Min. : 0.000   
## Yes : 760 1st Qu.:40.00 Vgood :2508 1st Qu.: 0.000   
## NA's: 142 Median :50.00 Good :2956 Median : 0.000   
## Mean :48.42 Fair :1010 Mean : 3.335   
## 3rd Qu.:58.00 Poor : 187 3rd Qu.: 3.000   
## Max. :80.00 NA's :2461 Max. :30.000   
## NA's :9371 NA's :2468   
## DaysMentHlthBad LittleInterest Depressed nPregnancies   
## Min. : 0.000 None :5103 None :5246 Min. : 1.000   
## 1st Qu.: 0.000 Several:1130 Several:1009 1st Qu.: 2.000   
## Median : 0.000 Most : 434 Most : 418 Median : 3.000   
## Mean : 4.127 NA's :3333 NA's :3327 Mean : 3.027   
## 3rd Qu.: 4.000 3rd Qu.: 4.000   
## Max. :30.000 Max. :32.000   
## NA's :2466 NA's :7396   
## nBabies Age1stBaby SleepHrsNight SleepTrouble  
## Min. : 0.000 Min. :14.00 Min. : 2.000 No :5799   
## 1st Qu.: 2.000 1st Qu.:19.00 1st Qu.: 6.000 Yes :1973   
## Median : 2.000 Median :22.00 Median : 7.000 NA's:2228   
## Mean : 2.457 Mean :22.65 Mean : 6.928   
## 3rd Qu.: 3.000 3rd Qu.:26.00 3rd Qu.: 8.000   
## Max. :12.000 Max. :39.00 Max. :12.000   
## NA's :7584 NA's :8116 NA's :2245   
## PhysActive PhysActiveDays TVHrsDay CompHrsDay   
## No :3677 Min. :1.000 2\_hr :1275 0\_to\_1\_hr:1409   
## Yes :4649 1st Qu.:2.000 1\_hr : 884 0\_hrs :1073   
## NA's:1674 Median :3.000 3\_hr : 836 1\_hr :1030   
## Mean :3.744 0\_to\_1\_hr: 638 2\_hr : 589   
## 3rd Qu.:5.000 More\_4\_hr: 615 3\_hr : 347   
## Max. :7.000 (Other) : 611 (Other) : 415   
## NA's :5337 NA's :5141 NA's :5137   
## TVHrsDayChild CompHrsDayChild Alcohol12PlusYr AlcoholDay   
## Min. :0.000 Min. :0.000 No :1368 Min. : 1.000   
## 1st Qu.:1.000 1st Qu.:0.000 Yes :5212 1st Qu.: 1.000   
## Median :2.000 Median :1.000 NA's:3420 Median : 2.000   
## Mean :1.939 Mean :2.198 Mean : 2.914   
## 3rd Qu.:3.000 3rd Qu.:6.000 3rd Qu.: 3.000   
## Max. :6.000 Max. :6.000 Max. :82.000   
## NA's :9347 NA's :9347 NA's :5086   
## AlcoholYear SmokeNow Smoke100 Smoke100n SmokeAge   
## Min. : 0.0 No :1745 No :4024 Non-Smoker:4024 Min. : 6.00   
## 1st Qu.: 3.0 Yes :1466 Yes :3211 Smoker :3211 1st Qu.:15.00   
## Median : 24.0 NA's:6789 NA's:2765 NA's :2765 Median :17.00   
## Mean : 75.1 Mean :17.83   
## 3rd Qu.:104.0 3rd Qu.:19.00   
## Max. :364.0 Max. :72.00   
## NA's :4078 NA's :6920   
## Marijuana AgeFirstMarij RegularMarij AgeRegMarij HardDrugs   
## No :2049 Min. : 1.00 No :3575 Min. : 5.00 No :4700   
## Yes :2892 1st Qu.:15.00 Yes :1366 1st Qu.:15.00 Yes :1065   
## NA's:5059 Median :16.00 NA's:5059 Median :17.00 NA's:4235   
## Mean :17.02 Mean :17.69   
## 3rd Qu.:19.00 3rd Qu.:19.00   
## Max. :48.00 Max. :52.00   
## NA's :7109 NA's :8634   
## SexEver SexAge SexNumPartnLife SexNumPartYear   
## No : 223 Min. : 9.00 Min. : 0.00 Min. : 0.000   
## Yes :5544 1st Qu.:15.00 1st Qu.: 2.00 1st Qu.: 1.000   
## NA's:4233 Median :17.00 Median : 5.00 Median : 1.000   
## Mean :17.43 Mean : 15.09 Mean : 1.342   
## 3rd Qu.:19.00 3rd Qu.: 12.00 3rd Qu.: 1.000   
## Max. :50.00 Max. :2000.00 Max. :69.000   
## NA's :4460 NA's :4275 NA's :5072   
## SameSex SexOrientation PregnantNow depressedYes   
## No :5353 Bisexual : 119 Yes : 72 Mode :logical   
## Yes : 415 Heterosexual:4638 No :1573 FALSE:5246   
## NA's:4232 Homosexual : 85 Unknown: 51 TRUE :1427   
## NA's :5158 NA's :8304 NA's :3327   
##   
##   
##

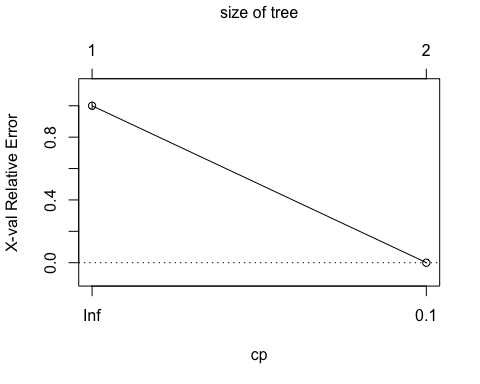
#grow tree  
fitd <- rpart(depressedYes~., method="class", data = NHANES) #decided to include all possible predictors because my tree was only a "root" with any combination that I came up with. In the end, only one variable was used to make the tree. Does not seem like a good fit given that depression was used ot make depressedYes.   
class(fitd)

## [1] "rpart"

#display results   
printcp(fitd)

##   
## Classification tree:  
## rpart(formula = depressedYes ~ ., data = NHANES, method = "class")  
##   
## Variables actually used in tree construction:  
## [1] Depressed  
##   
## Root node error: 1427/6673 = 0.21385  
##   
## n=6673 (3327 observations deleted due to missingness)  
##   
## CP nsplit rel error xerror xstd  
## 1 1.00 0 1 1 0.023472  
## 2 0.01 1 0 0 0.000000

#Visualize Cross-Validation Restuls   
plotcp(fitd)

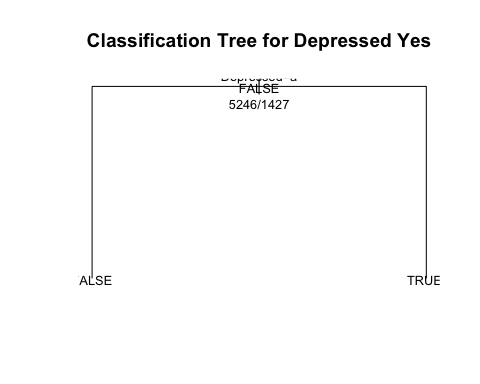


#Summary of Splits   
summary(fitd)

## Call:  
## rpart(formula = depressedYes ~ ., data = NHANES, method = "class")  
## n=6673 (3327 observations deleted due to missingness)  
##   
## CP nsplit rel error xerror xstd  
## 1 1.00 0 1 1 0.02347154  
## 2 0.01 1 0 0 0.00000000  
##   
## Variable importance  
## Depressed LittleInterest DaysMentHlthBad HealthGen   
## 68 15 15 1   
##   
## Node number 1: 6673 observations, complexity param=1  
## predicted class=FALSE expected loss=0.2138468 P(node) =1  
## class counts: 5246 1427  
## probabilities: 0.786 0.214   
## left son=2 (5246 obs) right son=3 (1427 obs)  
## Primary splits:  
## Depressed splits as LRR, improve=2243.68100, (0 missing)  
## LittleInterest splits as LRR, improve= 625.24820, (8 missing)  
## DaysMentHlthBad < 2.5 to the left, improve= 540.91780, (3 missing)  
## HealthGen splits as LLLRR, improve= 86.13322, (0 missing)  
## SleepTrouble splits as LR, improve= 80.23511, (0 missing)  
## Surrogate splits:  
## LittleInterest splits as LRR, agree=0.834, adj=0.226, (0 split)  
## DaysMentHlthBad < 9.5 to the left, agree=0.834, adj=0.221, (0 split)  
## HealthGen splits as LLLLR, agree=0.789, adj=0.014, (0 split)  
## ID < 51638.5 to the right, agree=0.787, adj=0.003, (0 split)  
##   
## Node number 2: 5246 observations  
## predicted class=FALSE expected loss=0 P(node) =0.7861532  
## class counts: 5246 0  
## probabilities: 1.000 0.000   
##   
## Node number 3: 1427 observations  
## predicted class=TRUE expected loss=0 P(node) =0.2138468  
## class counts: 0 1427  
## probabilities: 0.000 1.000

1. Report Effectiveness Was not effective on the NHANES data set. When including all variables, the only varibale included in the decision tree was “Depressed” which was used to make the “depressedYes” variable.
2. Visualization

# Plot the tree  
plot(fitd, uniform = TRUE, main = "Classification Tree for Depressed Yes")  
text(fitd, use.n = TRUE, all = TRUE, cex = 0.8)



1. Interpret Results… There isn’t much that is meaningful from this tree since the variable “depressed” was used to create “depressedYes”… Did I do something wrong?

* random forest

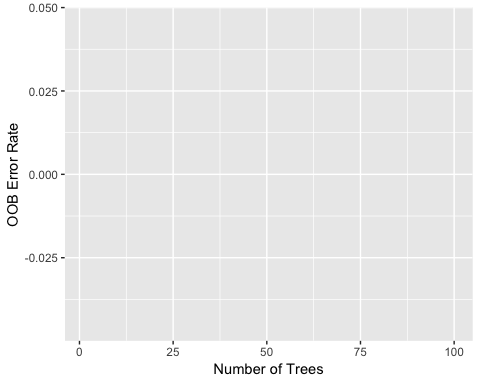
1. Build Classifier

NHANES.df <- as.data.frame(NHANES)  
set.seed(456789)  
# Random Forest for the NHANES dataset  
fitallrf <- rfsrc(depressedYes~., data=NHANES.df, ntree = 100, tree.err=TRUE, na.action = c("na.impute"))  
# view the results  
fitallrf

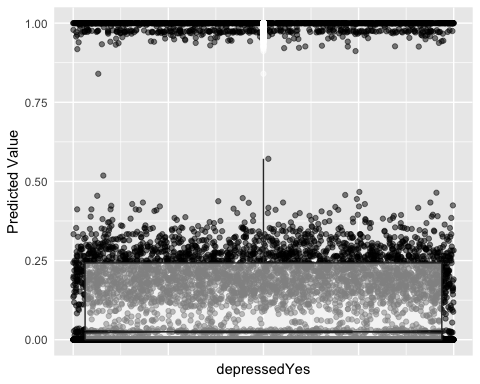
## Sample size: 10000  
## Was data imputed: yes  
## Number of trees: 100  
## Forest terminal node size: 5  
## Average no. of terminal nodes: 23.15  
## No. of variables tried at each split: 26  
## Total no. of variables: 76  
## Resampling used to grow trees: swr  
## Resample size used to grow trees: 10000  
## Analysis: RF-R  
## Family: regr  
## Splitting rule: mse \*random\*  
## Number of random split points: 10  
## % variance explained: 99.95  
## Error rate: 0

1. Report its effectiveness on the NHANES dataset and C) Make an approriate visualization of the model (I think this is answered here?)

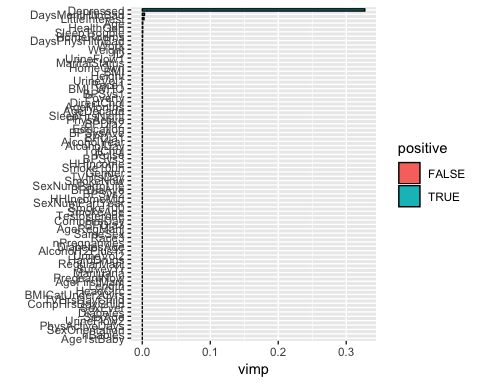
# Plot the OOB errors against the growth of the forest  
gg\_e <- gg\_error(fitallrf) #only one tree reported an error value and it was tree '100' and the rate was very small.   
plot(gg\_e)



# Plot the predicted depressedYes values  
plot(gg\_rfsrc(fitallrf), alpha = 0.5)



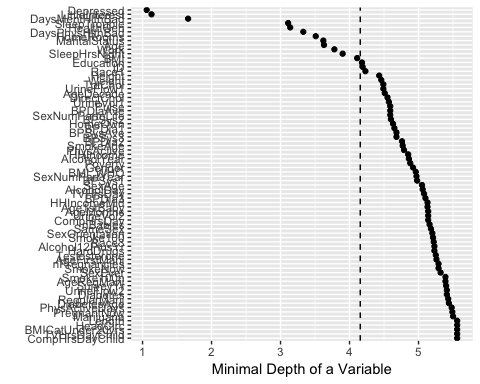
#Plot VIMP rankings of independent variables   
plot(gg\_vimp(fitallrf))



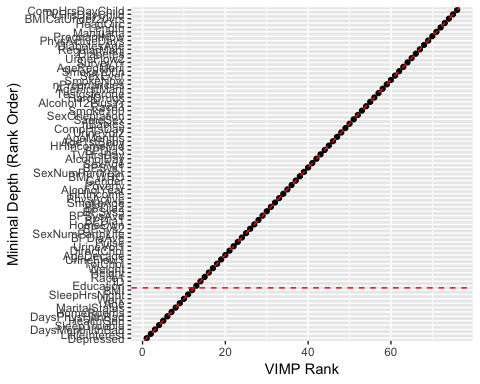
#minimal depth   
varsel\_depressedYes <- var.select(fitallrf)

## minimal depth variable selection ...  
##   
##   
## -----------------------------------------------------------  
## family : regr   
## var. selection : Minimal Depth   
## conservativeness : medium   
## x-weighting used? : TRUE   
## dimension : 76   
## sample size : 10000   
## ntree : 100   
## nsplit : 10   
## mtry : 26   
## nodesize : 5   
## refitted forest : FALSE   
## model size : 12   
## depth threshold : 4.1553   
## PE (true OOB) : 1e-04   
##   
##   
## Top variables:  
## depth vimp  
## Depressed 1.06 NA  
## LittleInterest 1.13 NA  
## DaysMentHlthBad 1.66 NA  
## SleepTrouble 3.11 NA  
## HealthGen 3.14 NA  
## DaysPhysHlthBad 3.33 NA  
## HomeRooms 3.51 NA  
## MaritalStatus 3.62 NA  
## Age 3.63 NA  
## Work 3.78 NA  
## SleepHrsNight 3.90 NA  
## BMI 4.11 NA  
## -----------------------------------------------------------

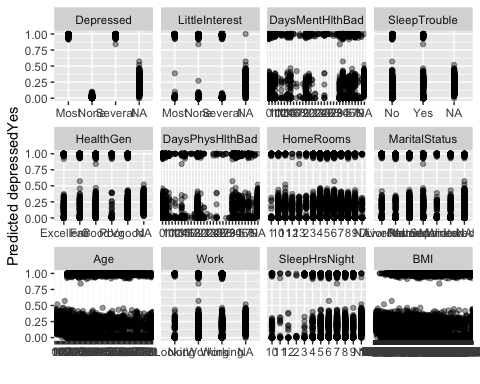
# Save the gg\_minimal\_depth object for later use  
gg\_md <- gg\_minimal\_depth(varsel\_depressedYes)  
# Plot the object  
plot(gg\_md)



# Plot minimal depth v VIMP  
gg\_mdVIMP <- gg\_minimal\_vimp(gg\_md)  
plot(gg\_mdVIMP) #Honestly, I don't know why I have two lines... BUT if we go off the diagonal line, the measurements are in agreement. #Create the variable dependence object from the random forest



gg\_v <- gg\_variable(fitallrf)  
# Use the top ranked minimal depth variables only, plotted in minimal depth rank order  
xvar <- gg\_md$topvars  
# Plot the variable list in a single panel plot  
plot(gg\_v, xvar = xvar, panel = TRUE, alpha = 0.4) +  
 labs(y="Predicted depressedYes", x="")



1. Interpretation According to the minimal depth, the top 3 important variables in the prediciton of depressedYes are: Depressed, LittleInterest, and DaysMentHlthBad. This makes sense since depressedYes is a derivative of Depressed and DaysMentHlthBad is known to correlate with depressedYes. Clinically, LittleInterest ebing important in the prediction of depressedYes since having little interest in things you previously enjoyed is part of the clinical diagnosis of depression.

* k-nearest neighbor A)Build the Classifier

#Create a dataset from NHANES   
NHANES2 <- NHANES %>%  
 dplyr::select(SleepHrsNight, PhysActiveDays, depressedYes) %>%  
 na.omit()  
glimpse(NHANES2)

## Observations: 3,333  
## Variables: 3  
## $ SleepHrsNight <int> 8, 8, 8, 7, 5, 4, 7, 7, 7, 8, 6, 6, 6, 5, 7, 6, 7…  
## $ PhysActiveDays <int> 5, 5, 5, 7, 5, 1, 7, 7, 7, 3, 3, 3, 2, 4, 1, 7, 6…  
## $ depressedYes <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, FALSE, …

#Convert to numeric  
NHANES2$depressedYes <- as.numeric(NHANES2$depressedYes)

1. Report effectiveness 100% prediciton using k of 1, 3, 5, 20, 50

# Apply knn procedure to predict Diabetes  
# Let's try different values of k to see how that affects performance  
knn.1 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k = 1)  
knn.3 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k = 3)  
knn.5 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k = 5)  
knn.20 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k =20)  
knn.50 <- knn(train = NHANES2, test = NHANES2, cl = NHANES2$depressedYes, k =50)  
#knn.1 amount correctly predicted  
100\*sum(NHANES2$depressedYes == knn.1)/length(knn.1)

## [1] 100

#knn.3 correct prediciton   
100\*sum(NHANES2$depressedYes == knn.1)/length(knn.3)

## [1] 100

#knn.5 correct   
100\*sum(NHANES2$depressedYes == knn.1)/length(knn.5)

## [1] 100

#knn.20 correct prediciton   
100\*sum(NHANES2$depressedYes == knn.1)/length(knn.20)

## [1] 100

#perfect prediction for all values of K... Let's try knn.50?  
100\*sum(NHANES2$depressedYes == knn.1)/length(knn.50)#still 100

## [1] 100

1. Appropriate Visualization

#create grid   
active <- range(~ SleepHrsNight, data = NHANES2)  
sleep <- range(~ PhysActiveDays, data = NHANES2)  
res <- 100  
fake\_grid <- expand.grid(  
 PhysActiveDays = seq(from = active[1], to = active[2], length.out = res),  
 SleepHrsNight = seq(from = sleep[1], to = sleep[2], length.out = res))  
# K-nearest neighbor prediction  
pred\_knn <- NHANES2 %>%  
 select(SleepHrsNight, PhysActiveDays) %>%  
 knn(test=select(fake\_grid, SleepHrsNight, PhysActiveDays), cl = NHANES2$depressedYes, k=5) %>%  
 as.numeric() - 1  
#build the data frame  
res <- fake\_grid %>%  
 mutate(  
 "K-nearest neighbor" = pred\_knn  
 ) %>%  
 gather(k="model", value = "y\_hat", -SleepHrsNight, -PhysActiveDays)  
#plot   
 ggplot(data = res, aes(x = SleepHrsNight, y = PhysActiveDays)) +  
 geom\_tile(aes(fill=y\_hat), color = NA) +  
 geom\_count(aes(color = depressedYes), alpha = 0.4, data = NHANES2) +  
 scale\_fill\_gradient(low = "white", high = "blue") +  
 scale\_color\_manual(values = c("gray", "gold")) +  
 scale\_size(range = c(0,2)) +  
 scale\_x\_continuous(expand = c(0.02, 0)) +  
 scale\_y\_continuous(expand = c(0.02, 0)) +  
 facet\_wrap(~model)  
 #Trying to plot this gives me the error: Error: Continuous value supplied to discrete scale. I'm not sure what I'm doing wrong. I went back and changed my predicotr variable from PhysActive (Y/N) to PhysActiveDays, but that didn't fix it.

1. Interpret the Results. What have you learned about people who self-report being depressed? The knn method overfits the model with 100% perfect prediction at the various levels of k.

For each model do the following:

1. Build the classifier.
2. Report its effectiveness on the NHANES dataset.
3. Make an appropriate visualization of this model.
4. Interpret the results. What have you learned about people who self-report being depressed?

PROBLEM 2: Repeat problem 1 except now use the quantitative variable called DaysMentHlthBad as your outcome variable. Run 3 models:

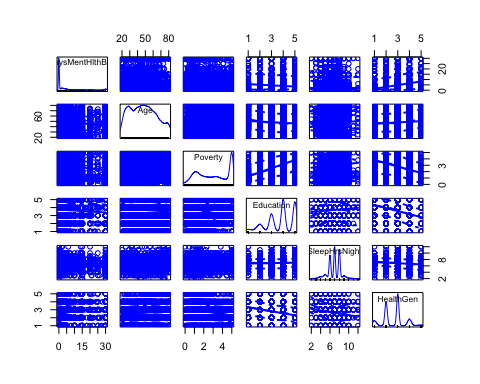
* multiple linear regression,

1. Build the classifier.

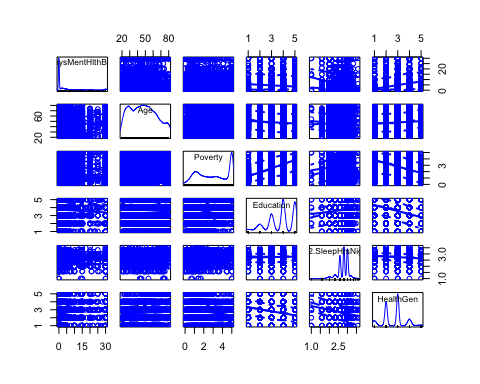
#Explore the data   
summary(NHANES)

## ID SurveyYr Gender Age   
## Min. :51624 2009\_10:5000 female:5020 Min. : 0.00   
## 1st Qu.:56904 2011\_12:5000 male :4980 1st Qu.:17.00   
## Median :62160 Median :36.00   
## Mean :61945 Mean :36.74   
## 3rd Qu.:67039 3rd Qu.:54.00   
## Max. :71915 Max. :80.00   
##   
## AgeDecade AgeMonths Race1 Race3   
## 40-49 :1398 Min. : 0.0 Black :1197 Asian : 288   
## 0-9 :1391 1st Qu.:199.0 Hispanic: 610 Black : 589   
## 10-19 :1374 Median :418.0 Mexican :1015 Hispanic: 350   
## 20-29 :1356 Mean :420.1 White :6372 Mexican : 480   
## 30-39 :1338 3rd Qu.:624.0 Other : 806 White :3135   
## (Other):2810 Max. :959.0 Other : 158   
## NA's : 333 NA's :5038 NA's :5000   
## Education MaritalStatus HHIncome   
## 8th Grade : 451 Divorced : 707 more 99999 :2220   
## 9 - 11th Grade: 888 LivePartner : 560 75000-99999:1084   
## High School :1517 Married :3945 25000-34999: 958   
## Some College :2267 NeverMarried:1380 35000-44999: 863   
## College Grad :2098 Separated : 183 45000-54999: 784   
## NA's :2779 Widowed : 456 (Other) :3280   
## NA's :2769 NA's : 811   
## HHIncomeMid Poverty HomeRooms HomeOwn   
## Min. : 2500 Min. :0.000 Min. : 1.000 Own :6425   
## 1st Qu.: 30000 1st Qu.:1.240 1st Qu.: 5.000 Rent :3287   
## Median : 50000 Median :2.700 Median : 6.000 Other: 225   
## Mean : 57206 Mean :2.802 Mean : 6.249 NA's : 63   
## 3rd Qu.: 87500 3rd Qu.:4.710 3rd Qu.: 8.000   
## Max. :100000 Max. :5.000 Max. :13.000   
## NA's :811 NA's :726 NA's :69   
## Work Weight Length HeadCirc   
## Looking : 311 Min. : 2.80 Min. : 47.10 Min. :34.20   
## NotWorking:2847 1st Qu.: 56.10 1st Qu.: 75.70 1st Qu.:39.58   
## Working :4613 Median : 72.70 Median : 87.00 Median :41.45   
## NA's :2229 Mean : 70.98 Mean : 85.02 Mean :41.18   
## 3rd Qu.: 88.90 3rd Qu.: 96.10 3rd Qu.:42.92   
## Max. :230.70 Max. :112.20 Max. :45.40   
## NA's :78 NA's :9457 NA's :9912   
## Height BMI BMICatUnder20yrs BMI\_WHO   
## Min. : 83.6 Min. :12.88 UnderWeight: 55 12.0\_18.5 :1277   
## 1st Qu.:156.8 1st Qu.:21.58 NormWeight : 805 18.5\_to\_24.9:2911   
## Median :166.0 Median :25.98 OverWeight : 193 25.0\_to\_29.9:2664   
## Mean :161.9 Mean :26.66 Obese : 221 30.0\_plus :2751   
## 3rd Qu.:174.5 3rd Qu.:30.89 NA's :8726 NA's : 397   
## Max. :200.4 Max. :81.25   
## NA's :353 NA's :366   
## Pulse BPSysAve BPDiaAve BPSys1   
## Min. : 40.00 Min. : 76.0 Min. : 0.00 Min. : 72.0   
## 1st Qu.: 64.00 1st Qu.:106.0 1st Qu.: 61.00 1st Qu.:106.0   
## Median : 72.00 Median :116.0 Median : 69.00 Median :116.0   
## Mean : 73.56 Mean :118.2 Mean : 67.48 Mean :119.1   
## 3rd Qu.: 82.00 3rd Qu.:127.0 3rd Qu.: 76.00 3rd Qu.:128.0   
## Max. :136.00 Max. :226.0 Max. :116.00 Max. :232.0   
## NA's :1437 NA's :1449 NA's :1449 NA's :1763   
## BPDia1 BPSys2 BPDia2 BPSys3   
## Min. : 0.00 Min. : 76.0 Min. : 0.00 Min. : 76.0   
## 1st Qu.: 62.00 1st Qu.:106.0 1st Qu.: 60.00 1st Qu.:106.0   
## Median : 70.00 Median :116.0 Median : 68.00 Median :116.0   
## Mean : 68.28 Mean :118.5 Mean : 67.66 Mean :117.9   
## 3rd Qu.: 76.00 3rd Qu.:128.0 3rd Qu.: 76.00 3rd Qu.:126.0   
## Max. :118.00 Max. :226.0 Max. :118.00 Max. :226.0   
## NA's :1763 NA's :1647 NA's :1647 NA's :1635   
## BPDia3 Testosterone DirectChol TotChol   
## Min. : 0.0 Min. : 0.25 Min. :0.390 Min. : 1.530   
## 1st Qu.: 60.0 1st Qu.: 17.70 1st Qu.:1.090 1st Qu.: 4.110   
## Median : 68.0 Median : 43.82 Median :1.290 Median : 4.780   
## Mean : 67.3 Mean : 197.90 Mean :1.365 Mean : 4.879   
## 3rd Qu.: 76.0 3rd Qu.: 362.41 3rd Qu.:1.580 3rd Qu.: 5.530   
## Max. :116.0 Max. :1795.60 Max. :4.030 Max. :13.650   
## NA's :1635 NA's :5874 NA's :1526 NA's :1526   
## UrineVol1 UrineFlow1 UrineVol2 UrineFlow2   
## Min. : 0.0 Min. : 0.0000 Min. : 0.0 Min. : 0.000   
## 1st Qu.: 50.0 1st Qu.: 0.4030 1st Qu.: 52.0 1st Qu.: 0.475   
## Median : 94.0 Median : 0.6990 Median : 95.0 Median : 0.760   
## Mean :118.5 Mean : 0.9793 Mean :119.7 Mean : 1.149   
## 3rd Qu.:164.0 3rd Qu.: 1.2210 3rd Qu.:171.8 3rd Qu.: 1.513   
## Max. :510.0 Max. :17.1670 Max. :409.0 Max. :13.692   
## NA's :987 NA's :1603 NA's :8522 NA's :8524   
## Diabetes DiabetesAge HealthGen DaysPhysHlthBad   
## No :9098 Min. : 1.00 Excellent: 878 Min. : 0.000   
## Yes : 760 1st Qu.:40.00 Vgood :2508 1st Qu.: 0.000   
## NA's: 142 Median :50.00 Good :2956 Median : 0.000   
## Mean :48.42 Fair :1010 Mean : 3.335   
## 3rd Qu.:58.00 Poor : 187 3rd Qu.: 3.000   
## Max. :80.00 NA's :2461 Max. :30.000   
## NA's :9371 NA's :2468   
## DaysMentHlthBad LittleInterest Depressed nPregnancies   
## Min. : 0.000 None :5103 None :5246 Min. : 1.000   
## 1st Qu.: 0.000 Several:1130 Several:1009 1st Qu.: 2.000   
## Median : 0.000 Most : 434 Most : 418 Median : 3.000   
## Mean : 4.127 NA's :3333 NA's :3327 Mean : 3.027   
## 3rd Qu.: 4.000 3rd Qu.: 4.000   
## Max. :30.000 Max. :32.000   
## NA's :2466 NA's :7396   
## nBabies Age1stBaby SleepHrsNight SleepTrouble  
## Min. : 0.000 Min. :14.00 Min. : 2.000 No :5799   
## 1st Qu.: 2.000 1st Qu.:19.00 1st Qu.: 6.000 Yes :1973   
## Median : 2.000 Median :22.00 Median : 7.000 NA's:2228   
## Mean : 2.457 Mean :22.65 Mean : 6.928   
## 3rd Qu.: 3.000 3rd Qu.:26.00 3rd Qu.: 8.000   
## Max. :12.000 Max. :39.00 Max. :12.000   
## NA's :7584 NA's :8116 NA's :2245   
## PhysActive PhysActiveDays TVHrsDay CompHrsDay   
## No :3677 Min. :1.000 2\_hr :1275 0\_to\_1\_hr:1409   
## Yes :4649 1st Qu.:2.000 1\_hr : 884 0\_hrs :1073   
## NA's:1674 Median :3.000 3\_hr : 836 1\_hr :1030   
## Mean :3.744 0\_to\_1\_hr: 638 2\_hr : 589   
## 3rd Qu.:5.000 More\_4\_hr: 615 3\_hr : 347   
## Max. :7.000 (Other) : 611 (Other) : 415   
## NA's :5337 NA's :5141 NA's :5137   
## TVHrsDayChild CompHrsDayChild Alcohol12PlusYr AlcoholDay   
## Min. :0.000 Min. :0.000 No :1368 Min. : 1.000   
## 1st Qu.:1.000 1st Qu.:0.000 Yes :5212 1st Qu.: 1.000   
## Median :2.000 Median :1.000 NA's:3420 Median : 2.000   
## Mean :1.939 Mean :2.198 Mean : 2.914   
## 3rd Qu.:3.000 3rd Qu.:6.000 3rd Qu.: 3.000   
## Max. :6.000 Max. :6.000 Max. :82.000   
## NA's :9347 NA's :9347 NA's :5086   
## AlcoholYear SmokeNow Smoke100 Smoke100n SmokeAge   
## Min. : 0.0 No :1745 No :4024 Non-Smoker:4024 Min. : 6.00   
## 1st Qu.: 3.0 Yes :1466 Yes :3211 Smoker :3211 1st Qu.:15.00   
## Median : 24.0 NA's:6789 NA's:2765 NA's :2765 Median :17.00   
## Mean : 75.1 Mean :17.83   
## 3rd Qu.:104.0 3rd Qu.:19.00   
## Max. :364.0 Max. :72.00   
## NA's :4078 NA's :6920   
## Marijuana AgeFirstMarij RegularMarij AgeRegMarij HardDrugs   
## No :2049 Min. : 1.00 No :3575 Min. : 5.00 No :4700   
## Yes :2892 1st Qu.:15.00 Yes :1366 1st Qu.:15.00 Yes :1065   
## NA's:5059 Median :16.00 NA's:5059 Median :17.00 NA's:4235   
## Mean :17.02 Mean :17.69   
## 3rd Qu.:19.00 3rd Qu.:19.00   
## Max. :48.00 Max. :52.00   
## NA's :7109 NA's :8634   
## SexEver SexAge SexNumPartnLife SexNumPartYear   
## No : 223 Min. : 9.00 Min. : 0.00 Min. : 0.000   
## Yes :5544 1st Qu.:15.00 1st Qu.: 2.00 1st Qu.: 1.000   
## NA's:4233 Median :17.00 Median : 5.00 Median : 1.000   
## Mean :17.43 Mean : 15.09 Mean : 1.342   
## 3rd Qu.:19.00 3rd Qu.: 12.00 3rd Qu.: 1.000   
## Max. :50.00 Max. :2000.00 Max. :69.000   
## NA's :4460 NA's :4275 NA's :5072   
## SameSex SexOrientation PregnantNow depressedYes   
## No :5353 Bisexual : 119 Yes : 72 Mode :logical   
## Yes : 415 Heterosexual:4638 No :1573 FALSE:5246   
## NA's:4232 Homosexual : 85 Unknown: 51 TRUE :1427   
## NA's :5158 NA's :8304 NA's :3327   
##   
##   
##

#produce a scatter plot matrix   
scatterplotMatrix(~ DaysMentHlthBad + Age + Poverty + Education + SleepHrsNight + HealthGen, data = NHANES)



#Some variables look weird, but Education and HealthGen are factors so might not be an issue? We'll see... I could possible log2() transform SleepHrsNight?  
scatterplotMatrix(~ DaysMentHlthBad + Age + Poverty + Education + log2(SleepHrsNight) + HealthGen, data = NHANES) #Didn't help. I'll just stick with the original.



#Regression Time   
mod1 <- lm(DaysMentHlthBad ~ Age + Poverty + Education + SleepHrsNight + HealthGen, data = NHANES)  
#What did we get?  
summary(mod1)

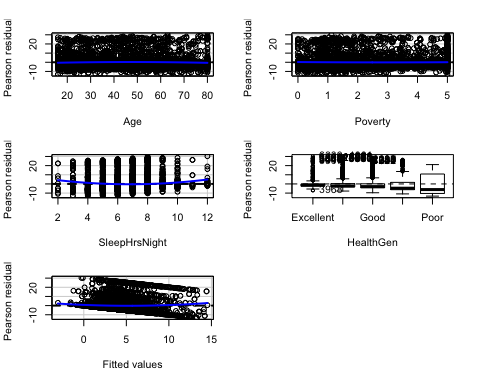
##   
## Call:  
## lm(formula = DaysMentHlthBad ~ Age + Poverty + Education + SleepHrsNight +   
## HealthGen, data = NHANES)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.871 -4.055 -2.402 0.340 30.526   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.130829 0.817405 11.171 < 2e-16 \*\*\*  
## Age -0.032027 0.006008 -5.331 1.01e-07 \*\*\*  
## Poverty -0.379188 0.069757 -5.436 5.67e-08 \*\*\*  
## Education9 - 11th Grade 1.591277 0.524348 3.035 0.00242 \*\*   
## EducationHigh School 0.723231 0.494546 1.462 0.14368   
## EducationSome College 1.310130 0.488675 2.681 0.00736 \*\*   
## EducationCollege Grad 0.983581 0.511253 1.924 0.05442 .   
## SleepHrsNight -0.737600 0.074363 -9.919 < 2e-16 \*\*\*  
## HealthGenVgood 0.913450 0.338051 2.702 0.00691 \*\*   
## HealthGenGood 1.718015 0.335925 5.114 3.25e-07 \*\*\*  
## HealthGenFair 3.849170 0.412295 9.336 < 2e-16 \*\*\*  
## HealthGenPoor 8.246274 0.698271 11.810 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.631 on 6015 degrees of freedom  
## (3973 observations deleted due to missingness)  
## Multiple R-squared: 0.07982, Adjusted R-squared: 0.07813   
## F-statistic: 47.43 on 11 and 6015 DF, p-value: < 2.2e-16

#What if we took Education out since not all factors significant?  
mod2 <- lm(DaysMentHlthBad ~ Age + Poverty + SleepHrsNight + HealthGen, data = NHANES)  
#what did we get?  
summary(mod2)

##   
## Call:  
## lm(formula = DaysMentHlthBad ~ Age + Poverty + SleepHrsNight +   
## HealthGen, data = NHANES)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.3020 -4.0224 -2.4091 0.3955 30.2753   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 10.115942 0.640157 15.802 < 2e-16 \*\*\*  
## Age -0.027602 0.005321 -5.187 2.20e-07 \*\*\*  
## Poverty -0.362991 0.060023 -6.048 1.55e-09 \*\*\*  
## SleepHrsNight -0.774780 0.070170 -11.041 < 2e-16 \*\*\*  
## HealthGenVgood 1.051626 0.319929 3.287 0.00102 \*\*   
## HealthGenGood 1.659435 0.315987 5.252 1.56e-07 \*\*\*  
## HealthGenFair 3.773370 0.388675 9.708 < 2e-16 \*\*\*  
## HealthGenPoor 7.961293 0.676853 11.762 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.556 on 6481 degrees of freedom  
## (3511 observations deleted due to missingness)  
## Multiple R-squared: 0.07422, Adjusted R-squared: 0.07322   
## F-statistic: 74.23 on 7 and 6481 DF, p-value: < 2.2e-16

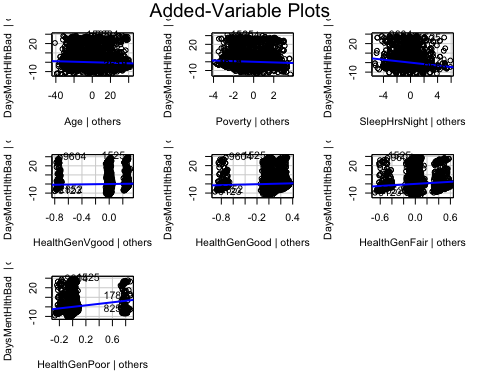
1. Report its effectiveness on the NHANES dataset.

# diagnostics for the secondmodel  
residualPlots(mod2) #Residuals seem to be okay.

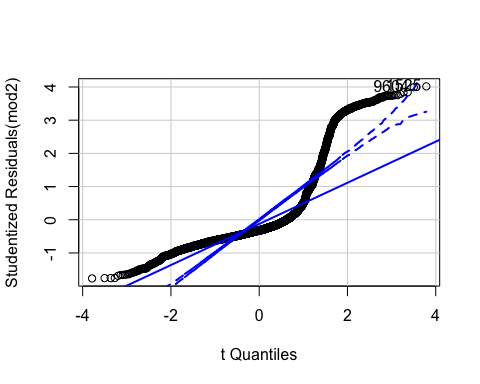


## Test stat Pr(>|Test stat|)   
## Age -2.9381 0.003314 \*\*   
## Poverty 0.9491 0.342600   
## SleepHrsNight 6.4320 1.349e-10 \*\*\*  
## HealthGen   
## Tukey test 4.5667 4.955e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#added variable plots  
avPlots(mod2, id.n=2, id.cex=0.7)



# run the qq-plot  
qqPlot(mod2, id.n=3) #data is not normally distributed.

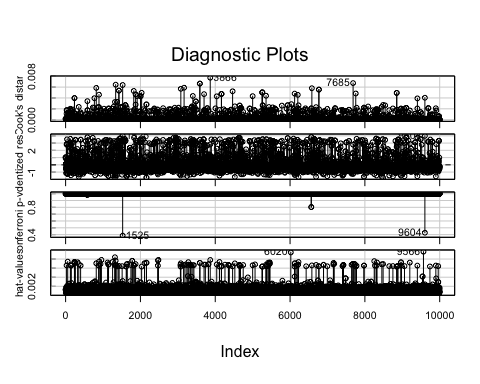


## [1] 1525 9604

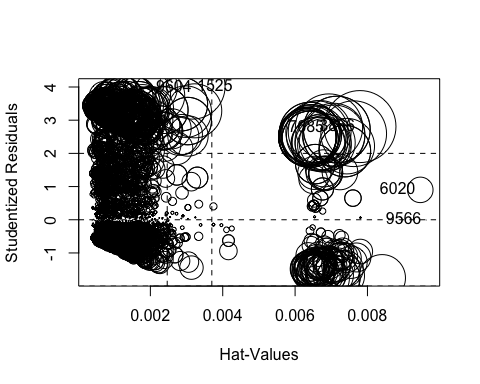
#run Bonferroni test for outliers  
outlierTest(mod2)

## No Studentized residuals with Bonferonni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferonni p  
## 1525 4.017981 5.9364e-05 0.38521

#identify highly influential points  
influenceIndexPlot(mod2, id.n=3)



#make influence plot to help us understand plot above  
influencePlot(mod2, id.n=3)



## StudRes Hat CookD  
## 1525 4.017980527 0.003160513 6.383287e-03  
## 3866 2.807327779 0.007792955 7.729213e-03  
## 6020 0.900551308 0.009460457 9.682331e-04  
## 7685 2.779498940 0.006937014 6.738890e-03  
## 9566 -0.006996614 0.009626763 5.948866e-08  
## 9604 3.993106640 0.002026124 4.037181e-03

#test for heteroskedasticity  
ncvTest(mod2) #P-values are signifianct would need to deal with

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 863.6683, Df = 1, p = < 2.22e-16

#Test for multicollinearity   
vif(mod2) #values are acceptable

## GVIF Df GVIF^(1/(2\*Df))  
## Age 1.045912 1 1.022698  
## Poverty 1.121389 1 1.058957  
## SleepHrsNight 1.022871 1 1.011371  
## HealthGen 1.132255 4 1.015648

1. Make an appropriate visualization of this model.

#Model 2 Results   
stargazer(mod2, title="Results from Model 2",  
 type = "html", align=TRUE, out="model.htm")

|  |  |
| --- | --- |
| **Results from Model 2** | |
|  | |
|  | Dependent variable: |
|  |  |
|  | DaysMentHlthBad |
|  | |
| Age | -0.028\*\*\* |
|  | (0.005) |
|  |  |
| Poverty | -0.363\*\*\* |
|  | (0.060) |
|  |  |
| SleepHrsNight | -0.775\*\*\* |
|  | (0.070) |
|  |  |
| HealthGenVgood | 1.052\*\*\* |
|  | (0.320) |
|  |  |
| HealthGenGood | 1.659\*\*\* |
|  | (0.316) |
|  |  |
| HealthGenFair | 3.773\*\*\* |
|  | (0.389) |
|  |  |
| HealthGenPoor | 7.961\*\*\* |
|  | (0.677) |
|  |  |
| Constant | 10.116\*\*\* |
|  | (0.640) |
|  |  |
|  | |
| Observations | 6,489 |
| R2 | 0.074 |
| Adjusted R2 | 0.073 |
| Residual Std. Error | 7.556 (df = 6481) |
| F Statistic | 74.230\*\*\* (df = 7; 6481) |
|  | |
| Note: | \*p\*\*p\*\*\*p<0.01 |

1. Interpret the results. What have you learned about people who self-report being depressed? From this model, controlling for poverty, hours of sleep per night, and self-rated health status, for every 1 unit increase in “age” the predicted value of “Days Mental Health Bad” decreases by -0.028. For every 1 unit increase in the varibale Poverty (which corresponds to a higher ratio of family income to poverty guidelines so less poverty), the predicted value of “Days Mental Health Bad” decreases by -0.363 when controling for all other variables in the model. Increased hours of sleep per night is also associated with a decrease in the predicted “Days Mental Health Bad”.

On the other hand, rating your general health as poor increased your predicted days of bad mental health by 7.961 when controling for all other variables in the model. Similarly, having fair general health increased yoru predicted days of bad mental health by 3.7773 when controling for all other variables in the model.

* regression tree, and

1. Build the classifier.

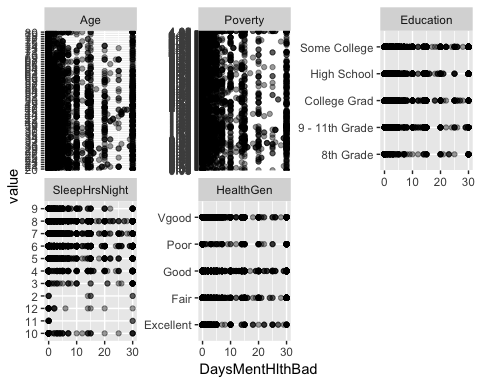
#Make NHANES DATA SET with DaysMentHlthBad  
NHANES3 <- NHANES %>%  
 dplyr::select(DaysMentHlthBad, Age, Poverty, Education, SleepHrsNight, HealthGen) %>%  
 na.omit()  
glimpse(NHANES3)

## Observations: 6,027  
## Variables: 6  
## $ DaysMentHlthBad <int> 15, 15, 15, 10, 3, 3, 3, 0, 0, 0, 0, 7, 0, 0, 0,…  
## $ Age <int> 34, 34, 34, 49, 45, 45, 45, 66, 58, 54, 50, 33, …  
## $ Poverty <dbl> 1.36, 1.36, 1.36, 1.91, 5.00, 5.00, 5.00, 2.20, …  
## $ Education <fct> High School, High School, High School, Some Coll…  
## $ SleepHrsNight <int> 4, 4, 4, 8, 8, 8, 8, 7, 5, 4, 7, 6, 6, 7, 7, 6, …  
## $ HealthGen <fct> Good, Good, Good, Good, Vgood, Vgood, Vgood, Vgo…

summary(NHANES3) #median of sleephrsnight=7

## DaysMentHlthBad Age Poverty Education   
## Min. : 0.000 Min. :20.00 Min. :0.000 8th Grade : 311   
## 1st Qu.: 0.000 1st Qu.:33.00 1st Qu.:1.390 9 - 11th Grade: 705   
## Median : 0.000 Median :47.00 Median :3.000 High School :1275   
## Mean : 4.152 Mean :47.53 Mean :2.988 Some College :1917   
## 3rd Qu.: 4.000 3rd Qu.:60.00 3rd Qu.:5.000 College Grad :1819   
## Max. :30.000 Max. :80.00 Max. :5.000   
## SleepHrsNight HealthGen   
## Min. : 2.000 Excellent: 695   
## 1st Qu.: 6.000 Vgood :1960   
## Median : 7.000 Good :2391   
## Mean : 6.907 Fair : 826   
## 3rd Qu.: 8.000 Poor : 155   
## Max. :12.000

# Plot panels for each covariate  
NHANES3.panel<- melt(NHANES3, id.vars="DaysMentHlthBad")  
ggplot(NHANES3.panel, aes(x=DaysMentHlthBad, y=value)) +  
 geom\_point(alpha=0.4)+  
 scale\_color\_brewer(palette="Set2")+  
 facet\_wrap(~variable, scales="free\_y", ncol=3)

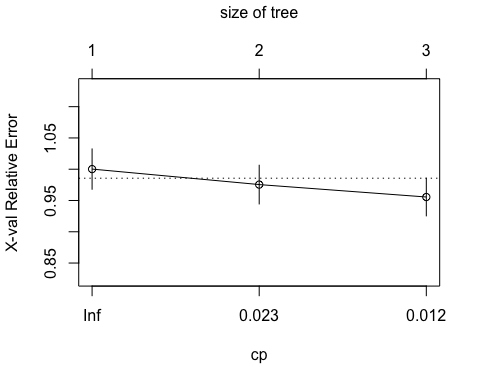


#Not really sure which of these will be good predictors.   
#Let's see how it partitions with all the variables in   
fitall <- rpart(DaysMentHlthBad ~., data=NHANES3)  
#let's look at our fit   
printcp(fitall)

##   
## Regression tree:  
## rpart(formula = DaysMentHlthBad ~ ., data = NHANES3)  
##   
## Variables actually used in tree construction:  
## [1] HealthGen SleepHrsNight  
##   
## Root node error: 380697/6027 = 63.165  
##   
## n= 6027   
##   
## CP nsplit rel error xerror xstd  
## 1 0.033093 0 1.00000 1.00021 0.032193  
## 2 0.015344 1 0.96691 0.97542 0.030913  
## 3 0.010000 2 0.95156 0.95554 0.030094

1. Report its effectiveness on the NHANES dataset.

#Cross-Validation   
plotcp(fitall)

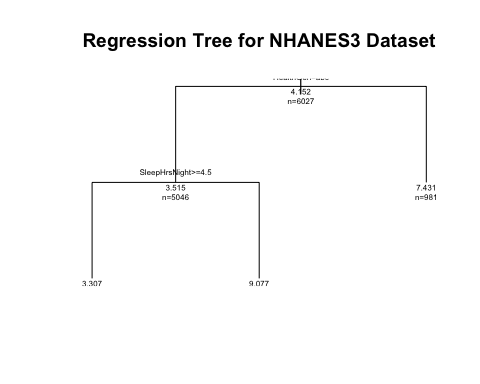


#Summary of Fit   
summary(fitall)

## Call:  
## rpart(formula = DaysMentHlthBad ~ ., data = NHANES3)  
## n= 6027   
##   
## CP nsplit rel error xerror xstd  
## 1 0.03309310 0 1.0000000 1.0002123 0.03219308  
## 2 0.01534431 1 0.9669069 0.9754236 0.03091309  
## 3 0.01000000 2 0.9515626 0.9555369 0.03009374  
##   
## Variable importance  
## HealthGen SleepHrsNight   
## 68 32   
##   
## Node number 1: 6027 observations, complexity param=0.0330931  
## mean=4.152149, MSE=63.16534   
## left son=2 (5046 obs) right son=3 (981 obs)  
## Primary splits:  
## HealthGen splits as LLLRR, improve=0.033093100, (0 missing)  
## SleepHrsNight < 4.5 to the right, improve=0.030070690, (0 missing)  
## Poverty < 1.915 to the right, improve=0.019099250, (0 missing)  
## Education splits as RRRRL, improve=0.007448942, (0 missing)  
## Age < 53.5 to the right, improve=0.006665236, (0 missing)  
##   
## Node number 2: 5046 observations, complexity param=0.01534431  
## mean=3.514665, MSE=50.49196   
## left son=4 (4864 obs) right son=5 (182 obs)  
## Primary splits:  
## SleepHrsNight < 4.5 to the right, improve=0.022927560, (0 missing)  
## Poverty < 3.235 to the right, improve=0.010618260, (0 missing)  
## Age < 53.5 to the right, improve=0.010221350, (0 missing)  
## HealthGen splits as LLR--, improve=0.009734853, (0 missing)  
## Education splits as LRRRL, improve=0.004839813, (0 missing)  
##   
## Node number 3: 981 observations  
## mean=7.431193, MSE=115.5113   
##   
## Node number 4: 4864 observations  
## mean=3.306538, MSE=46.70558   
##   
## Node number 5: 182 observations  
## mean=9.076923, MSE=119.5875

1. Make an appropriate visualization of this model.

#plot the tree   
plot(fitall, uniform = TRUE, compress = FALSE, main = "Regression Tree for NHANES3 Dataset")  
text(fitall, use.n = TRUE, all = TRUE, cex = 0.5)



1. Interpret the results. What have you learned about people who self-report being depressed?

The two most important variables in the NHANES3 dataset that was created for determining “DaysMentHlthBad” are The General Ratings of Health and the number of hours slept per night.

* random forest.

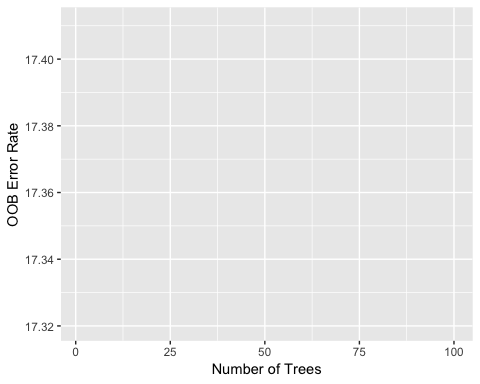
1. Build the classifier.

NHANES.df <- as.data.frame(NHANES)  
set.seed(123456789)  
# Random Forest for the NHANES dataset  
fitallrf2 <- rfsrc(DaysMentHlthBad~., data=NHANES.df, ntree = 100, tree.err=TRUE, na.action = c("na.impute"))  
# view the results  
fitallrf2

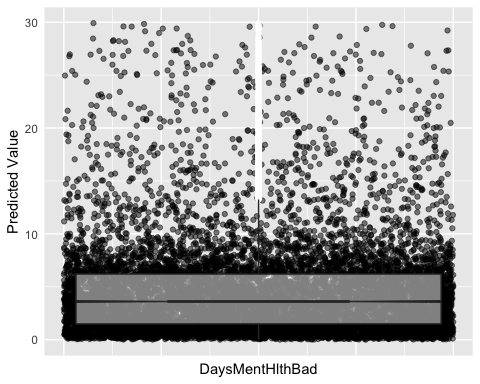
## Sample size: 10000  
## Was data imputed: yes  
## Number of trees: 100  
## Forest terminal node size: 5  
## Average no. of terminal nodes: 1211.82  
## No. of variables tried at each split: 26  
## Total no. of variables: 76  
## Resampling used to grow trees: swr  
## Resample size used to grow trees: 10000  
## Analysis: RF-R  
## Family: regr  
## Splitting rule: mse \*random\*  
## Number of random split points: 10  
## % variance explained: 71.7  
## Error rate: 17.37

1. Report its effectiveness on the NHANES dataset AND (C) Make an appropriate visualization of this model. I think?

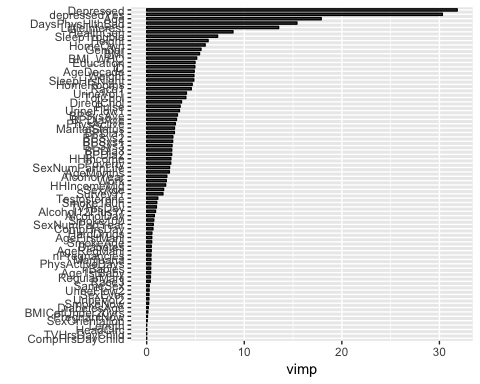
# Plot the OOB errors against the growth of the forest  
gg\_e2 <- gg\_error(fitallrf2) #only one tree reported an error value and it was tree '100' and the rate was very small.   
plot(gg\_e2)



# Plot the predicted depressedYes values  
plot(gg\_rfsrc(fitallrf2), alpha = 0.5)



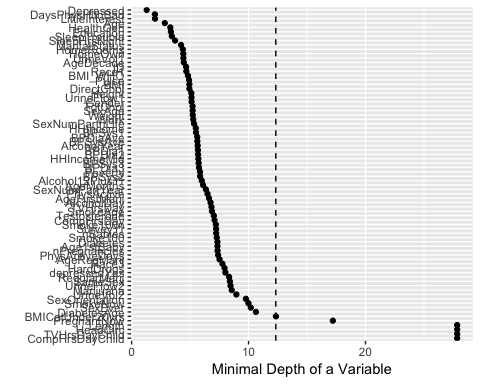
#Plot VIMP rankings of independent variables   
plot(gg\_vimp(fitallrf2))



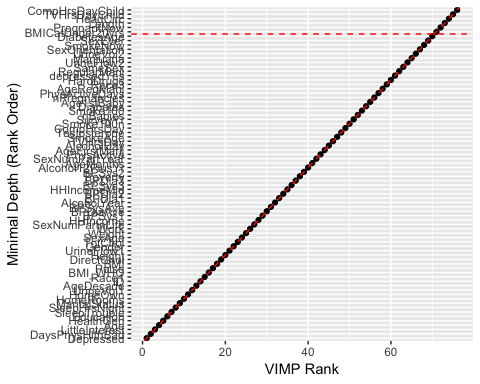
#minimal depth   
varsel\_depressedYes2 <- var.select(fitallrf2)

## minimal depth variable selection ...  
##   
##   
## -----------------------------------------------------------  
## family : regr   
## var. selection : Minimal Depth   
## conservativeness : medium   
## x-weighting used? : TRUE   
## dimension : 76   
## sample size : 10000   
## ntree : 100   
## nsplit : 10   
## mtry : 26   
## nodesize : 5   
## refitted forest : FALSE   
## model size : 70   
## depth threshold : 12.3313   
## PE (true OOB) : 17.3655   
##   
##   
## Top variables:  
## depth vimp  
## Depressed 1.29 NA  
## DaysPhysHlthBad 1.99 NA  
## LittleInterest 2.00 NA  
## Age 2.84 NA  
## HealthGen 3.31 NA  
## Education 3.36 NA  
## SleepTrouble 3.44 NA  
## SleepHrsNight 3.72 NA  
## MaritalStatus 4.23 NA  
## HomeRooms 4.38 NA  
## HomeOwn 4.42 NA  
## UrineVol1 4.43 NA  
## AgeDecade 4.46 NA  
## ID 4.66 NA  
## Race1 4.68 NA  
## BMI\_WHO 4.82 NA  
## Pulse 4.91 NA  
## BMI 4.94 NA  
## DirectChol 4.94 NA  
## Height 5.11 NA  
## UrineFlow1 5.12 NA  
## Gender 5.13 NA  
## TotChol 5.21 NA  
## SexAge 5.22 NA  
## Weight 5.23 NA  
## Work 5.27 NA  
## SexNumPartnLife 5.34 NA  
## HHIncome 5.49 NA  
## BPSys1 5.51 NA  
## BPDiaAve 5.61 NA  
## BPSysAve 5.62 NA  
## AlcoholYear 5.65 NA  
## BPDia1 5.67 NA  
## BPDia2 5.68 NA  
## HHIncomeMid 5.71 NA  
## BPSys3 5.73 NA  
## BPDia3 5.76 NA  
## Poverty 5.85 NA  
## BPSys2 5.86 NA  
## Alcohol12PlusYr 6.00 NA  
## AgeMonths 6.11 NA  
## SexNumPartYear 6.35 NA  
## PhysActive 6.49 NA  
## AgeFirstMarij 6.62 NA  
## AlcoholDay 6.75 NA  
## TVHrsDay 6.83 NA  
## SmokeAge 6.87 NA  
## Testosterone 7.03 NA  
## CompHrsDay 7.09 NA  
## Smoke100n 7.20 NA  
## SurveyYr 7.23 NA  
## nBabies 7.23 NA  
## Smoke100 7.28 NA  
## Diabetes 7.34 NA  
## Age1stBaby 7.34 NA  
## nPregnancies 7.35 NA  
## PhysActiveDays 7.40 NA  
## AgeRegMarij 7.52 NA  
## Race3 7.77 NA  
## HardDrugs 7.95 NA  
## depressedYes 8.02 NA  
## RegularMarij 8.34 NA  
## SameSex 8.40 NA  
## UrineFlow2 8.46 NA  
## Marijuana 8.56 NA  
## UrineVol2 8.96 NA  
## SexOrientation 9.77 NA  
## SmokeNow 9.99 NA  
## SexEver 10.19 NA  
## DiabetesAge 10.63 NA  
## -----------------------------------------------------------

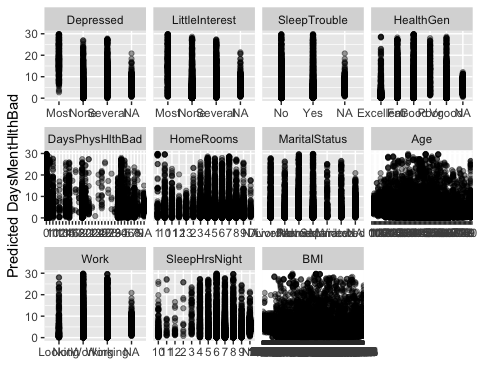
# Save the gg\_minimal\_depth object for later use  
gg\_md2 <- gg\_minimal\_depth(varsel\_depressedYes2)  
# Plot the object  
plot(gg\_md2)



# Plot minimal depth v VIMP  
gg\_mdVIMP2 <- gg\_minimal\_vimp(gg\_md2)  
plot(gg\_mdVIMP2) #Going off the diagonal line, the measurements are in agreement. #Create the variable dependence object from the random forest



gg\_v2 <- gg\_variable(fitallrf2)  
# Use the top ranked minimal depth variables only, plotted in minimal depth rank order  
xvar2 <- gg\_md2$topvars  
# Plot the variable list in a single panel plot  
plot(gg\_v2, xvar = xvar, panel = TRUE, alpha = 0.4) +  
 labs(y="Predicted DaysMentHlthBad", x="")



#plot the random forest plot

1. Interpret the results. What have you learned about people who self-report being depressed?

Looking at the minimal depth variables again (those closer to the node), Depressed, DaysPhysHlthBad, Little Interest, and Age are all <3. These variables being important in the prediction of DaysMentHlthBad makes since clinically as if you are depressed you are probaly going to have more bad mental health days; the interaction between physical and mental health is well established and I’m not surprised to see daysphyshlthbad being important; little interest is also a symptom of depressiona and is not surprising to be imporant in prediction DaysMentHlthBad; and finally age is also inlcuded in predicting DaysMentHlthBad.

And answer parts A, B, C, and D again for each model.

**NOTE: depressedYes and DaysMentHlthBad are correlated but were 2 separate questions and are not perfectly aligned. The amount of missing data NA's are different between the 2 variables.** To learn more about the variables in the dataset, run help(NHANES, package = "NHANES").