N741 Spring 2018 - Homework 6

Homework 6 - DUE FRIDAY April 6, 2018

Tori Bonisese

April 6, 2018

Git Repo

https://github.com/vbonise/N741Spring2018_Homework6.git

Homework 6

Background and Information on HELP Dataset

For homework 6, you will be working with the **HELP** (Health Evaluation and Linkage to Primary Care) Dataset.

The HELP Dataset:

- You can learn more about the HELP (Health Evaluation and Linkage to Primary Care) dataset at https://nhorton.people.amherst.edu/sasr2/datasets.php. This dataset is also used by Ken Kleinman and Nicholas J. Horton for their book "SAS and R: Data Management, Statistical Analysis, and Graphics" (which is another helpful textbook).
- You can download the datasets from their website https://nhorton.people.amherst.edu/sasr2/datasets.php
- The original publication is referenced at https://www.ncbi.nlm.nih.gov/pubmed/12653820?ordinalpos=17&itool=EntrezSystem2.PEntrez.Pubmed.Pubmed_ResultsPanel.Pubmed_DefaultReportPanel.Pubmed_RVDocSum
- The HELP documentation (including all forms/surveys/instruments used) are located at:
  - https://nhorton.people.amherst.edu/help/
  - specifically the details on all BASELINE assessments are located in this PDF https://nhorton.people.amherst.edu/help/HELP-baseline.pdf
  - with the follow up time points described in the PDF https://nhorton.people.amherst.edu/help/HELP-followup.pdf

Summary of Entire HELP Dataset - Complete Codebook

See complete data descriptions and codebook at
https://melindahiggins2000.github.io/N736Fall2017_HELPdataset/

Variables for Homework 6

```
###load dataset:
load("/Users/victoriabonisese/Downloads/help.Rdata")
```

For Homework 6, you will focus only on these variables from the HELP dataset:

Use these variables from HELP dataset for Homework 06

	Variable Label
age	Age at baseline (in years)
female	Gender of respondent
pss_fr	Perceived Social Support - friends
homeless	One or more nights on the street or shelter in past 6 months
pcs	SF36 Physical Composite Score - Baseline
mcs	SF36 Mental Composite Score - Baseline
cesd	CESD total score - Baseline

Homework 6 Assignment

SETUP Download and run the "loadHELP.R" `R` script (included in this Github repo https://github.com/melindahiggins2000/N741Spring2018_Homework6) to read in the HELP Dataset "helpmkh.sav". This script also pulls out the variables you need and creates the dichotomous variable for depression `cesd_gte16` which you will need for the logistic regression.

After running this R script, you will have a data frame called `h1` you can use to do the rest of your analyses. You can also copy this code into your first R markdown code chunk to get you started on Homework 6.

```
###run script
# use this script to setup the data subset from
# HELP to use for N741 Spring 2018 Homework 6

# load libraries and dataset

library(tidyverse)
library(haven)
helpdata <- haven::read_spss("helpmkh.sav")

# choose variables for Homework 6

h1 <- helpdata %>%
```

```
  select(age, female, pss_fr, homeless,
         pcs, mcs, cesd)

# add dichotomous variable
# to indicate depression for
# people with CESD scores >= 16

h1 <- h1 %>%
  mutate(cesd_gte16 = cesd >= 16)

# change cesd_gte16 LOGIC variable type
# to numeric coded 1=TRUE and 0=FALSE

h1$cesd_gte16 <- as.numeric(h1$cesd_gte16)

# check final data subset h1
summary(h1)
```

```
##       age            female          pss_fr          homeless
##  Min.   :19.00   Min.   :0.0000   Min.   : 0.000   Min.   :0.0000
##  1st Qu.:30.00   1st Qu.:0.0000   1st Qu.: 3.000   1st Qu.:0.0000
##  Median :35.00   Median :0.0000   Median : 7.000   Median :0.0000
##  Mean   :35.65   Mean   :0.2362   Mean   : 6.706   Mean   :0.4614
##  3rd Qu.:40.00   3rd Qu.:0.0000   3rd Qu.:10.000   3rd Qu.:1.0000
##  Max.   :60.00   Max.   :1.0000   Max.   :14.000   Max.   :1.0000
##       pcs              mcs              cesd         cesd_gte16
##  Min.   :14.07   Min.   : 6.763   Min.   : 1.00   Min.   :0.0000
##  1st Qu.:40.38   1st Qu.:21.676   1st Qu.:25.00   1st Qu.:1.0000
##  Median :48.88   Median :28.602   Median :34.00   Median :1.0000
##  Mean   :48.05   Mean   :31.677   Mean   :32.85   Mean   :0.8985
##  3rd Qu.:56.95   3rd Qu.:40.941   3rd Qu.:41.00   3rd Qu.:1.0000
##  Max.   :74.81   Max.   :62.175   Max.   :60.00   Max.   :1.0000
```

For Homework 6, you will be looking at depression in these subjects. First, you will be running a model to look at the continuous depression measure - the CESD Center for Epidemiologic Studies Depression Scale which is a measure of depressive symptoms. Also see the APA details on the CESD at http://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/depression-scale.aspx. The CESD can be used to predict actual clinical depression but it is not technically a diagnosis of depression. The CESD scores range from 0 (no depressive symptoms) to 60 (most severe depressive symptoms). You will use the (`cesd`) variable to run a linear regression.

The recommended threshold use to indicate potential clinical depression is for people with scores of 16 or greater. You will then use the variable created using this cutoff (`cesd_gte16`) to perform a similar modeling approach with the variables to predict the probability of clinical depression (using logistic regression).

Homework 6 Tasks

1. [Model 1] Run a simple linear regression (`lm()`) for `cesd` using the `mcs` variable, which is the mental component quality of life score from the SF36.

```
model1 <- lm(cesd ~ mcs, data = h1)
summary(model1)

##
## Call:
## lm(formula = cesd ~ mcs, data = h1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -27.3593  -6.7277  -0.0024   6.2374  24.4239
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 53.90219    1.14723   46.98   <2e-16 ***
## mcs         -0.66467    0.03357  -19.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.164 on 451 degrees of freedom
## Multiple R-squared:  0.465,  Adjusted R-squared:  0.4638
## F-statistic:   392 on 1 and 451 DF,  p-value: < 2.2e-16
```

2. Write the equation of the final fitted model (i.e. what is the intercept and the slope)? Write a sentence describing the model results (interpret the intercept and slope). *NOTE: The `mcs` values range form 0 to 100 where the population norm for "normal mental health quality of life" is considered to be a 50. If you score higher than 50 on the `mcs` you have mental health better than the population and visa versa - if your `mcs` scores are less than 50 then your mental health is considered to be worse than the population norm.*

Answer

Model 1 y = 53.90 + -.66x For every one point increase on the MCS, the CESD score decreases by .66.

3. How much variability in the `cesd` does the `mcs` explain? (what is the R2?) Write a sentence describing how well the `mcs` does in predicting the `cesd`. **The R-squared is about .46 which means that 46% of the variation in CESD is explained by MCS**

4. [Model 2] Run a second linear regression model (`lm()`) for the `cesd` putting in all of the other variables:
   - `age`
   - `female`
   - `pss_fr`
   - `homeless`

- pcs
- mcs

```
model2 <- lm(cesd ~ mcs + age + female + pss_fr + homeless + pcs + mcs, data
= h1)
summary(model2) ##this prints out the results with the test & fit coefficient
s

##
## Call:
## lm(formula = cesd ~ mcs + age + female + pss_fr + homeless +
##     pcs + mcs, data = h1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1711  -5.9894  -0.2077   5.5706  27.3137
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.30046    3.18670  20.492  < 2e-16 ***
## mcs         -0.62093    0.03261 -19.042  < 2e-16 ***
## age         -0.01348    0.05501  -0.245   0.8065
## female       2.35028    0.98810   2.379   0.0178 *
## pss_fr      -0.25569    0.10567  -2.420   0.0159 *
## homeless     0.46545    0.84261   0.552   0.5810
## pcs         -0.23639    0.03987  -5.929  6.1e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.683 on 446 degrees of freedom
## Multiple R-squared:  0.5249, Adjusted R-squared:  0.5185
## F-statistic: 82.14 on 6 and 446 DF,  p-value: < 2.2e-16

+ Print out the model results with the coefficients and tests and model fit s
tatistics.
```

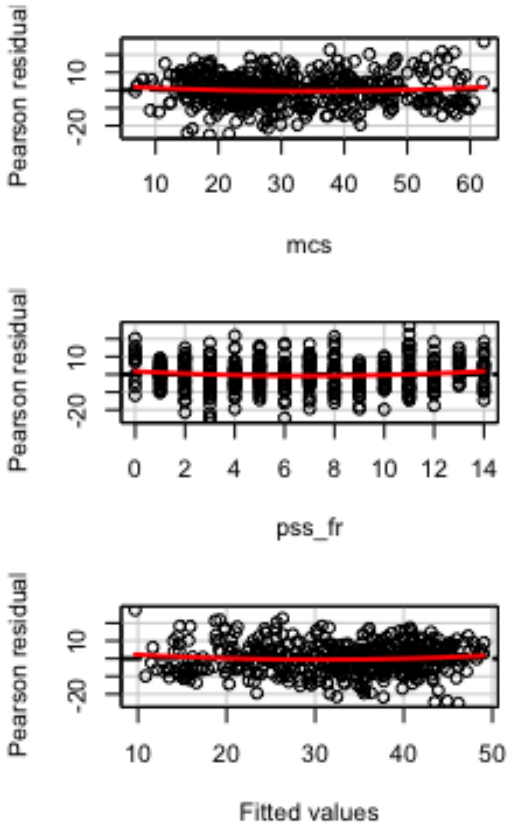
5. Which variables are significant in the model? Write a sentence or two describing the impact of these variables for predicting depression scores (HINT: interpret the coefficient terms).

MCS and PCS are significant at the 0.001 level and female and pss_fr are significant at the 0.1 level. This means that for every unit incease of MCS, CESD decreases by 0.63, for every unit increase of age, CESD decreases by 0.013, females are associated with a CESD 2.35 points higher than makes and for every unit increase of PSS_FR, CESD decreases by 0.26.

6. Following the example we did in class for the Prestige dataset https://cdn.rawgit.com/vhertzb/2018week9/2f2ea142/2018week9.html?raw=true, generate the diagnostic plots for this model with these 6 predictors (e.g. get the

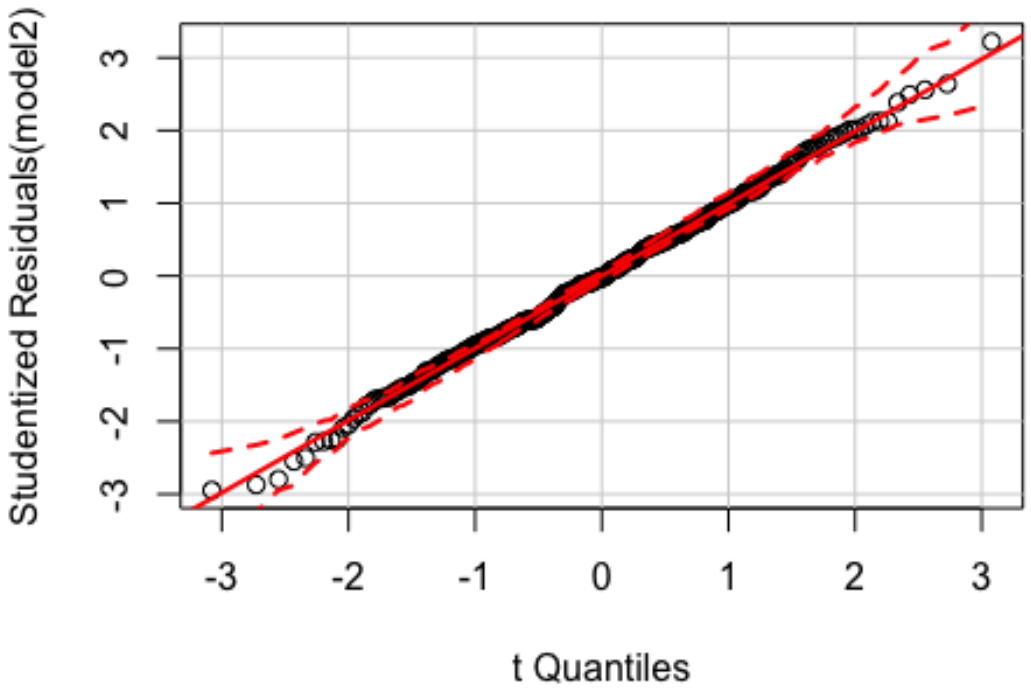
residual plot by variables, the added-variable plots, the Q-Q plot, diagnostic plots). Also run the VIFs to check for multicollinearity issues.

```
library(car)

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

#residual plots
residualPlots(model2)
```

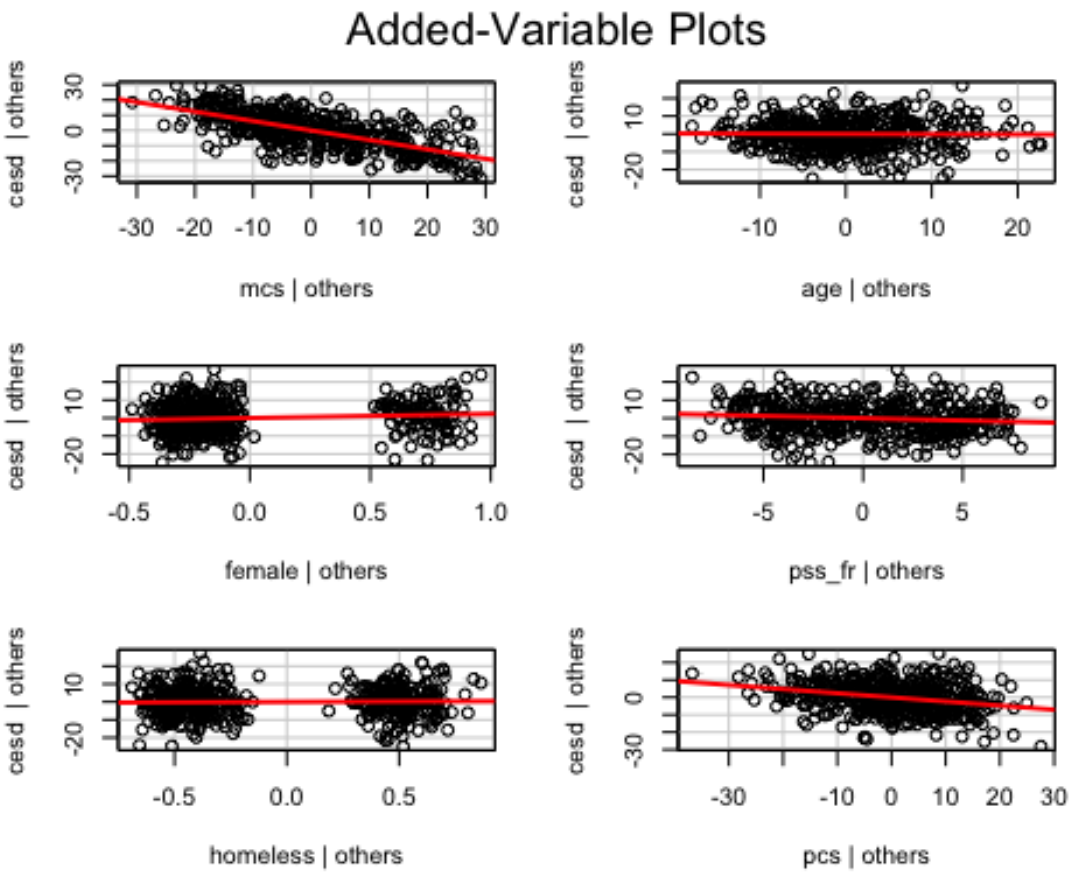



```
##            Test stat Pr(>|t|)
## mcs            1.260    0.208
## age            1.941    0.053
## pss_fr         1.964    0.050
## pcs            0.081    0.936
## Tukey test     1.434    0.152
```

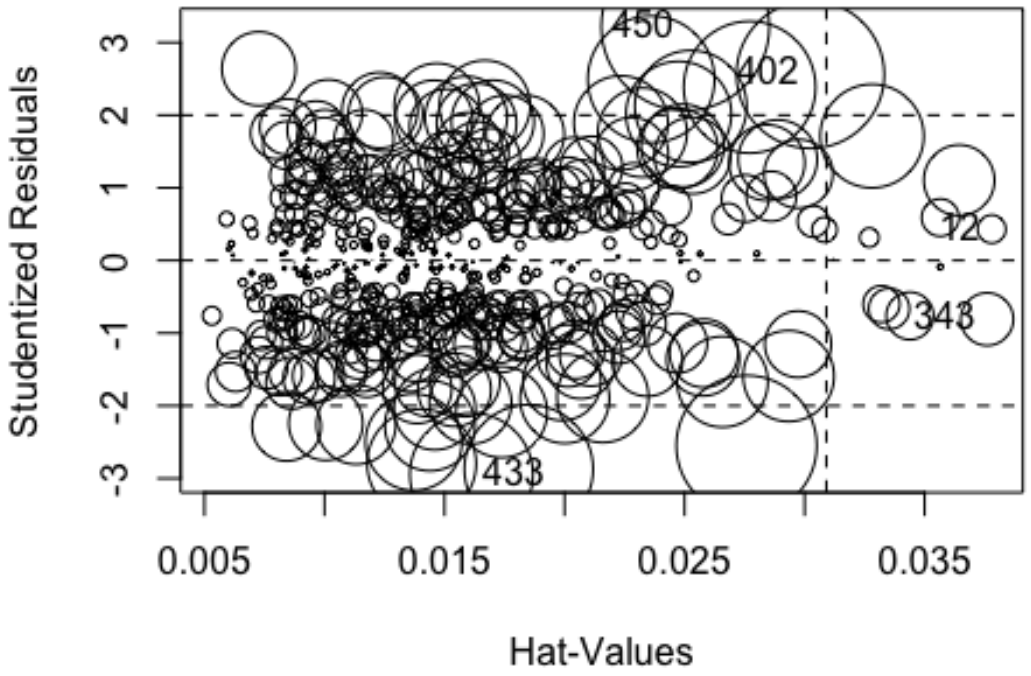
```
##qqPlot
qqPlot(model2)
```




```
##added variable plots
avPlots(model2, id.n=2, id.cex=.08)
```

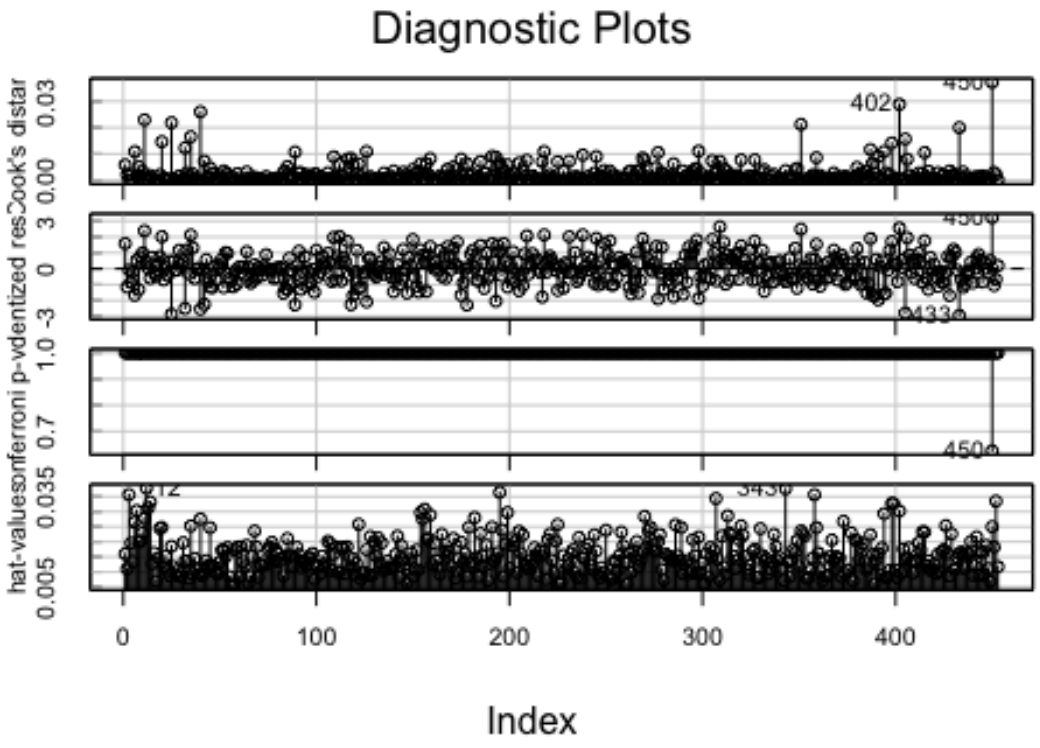



```
#Diagnotic Plots
##identify highly influential points
influencePlot(model2, id.n=2)
```




```
##         StudRes         Hat       CookD
## 12   0.4313265 0.03779399 0.001045833
## 343 -0.8084322 0.03760068 0.003650624
## 402  2.5591353 0.03023968 0.028815823
## 433 -2.9474775 0.01612078 0.019990575
## 450  3.2188680 0.02502996 0.037218269
```

```
influenceIndexPlot(model2, id.n=2)
```




```
##heteroskedasticity
ncvTest(model2)

## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 4.857132    Df = 1     p = 0.02753206

##multicollinearity
vif(model2)

##      mcs      age   female   pss_fr homeless      pcs
## 1.050768 1.078264 1.058232 1.068213 1.060007 1.108172
```

7. [Model 3] Repeat Model 1 above, except this time run a logistic regression (`glm()`) to predict CESD scores => 16 (using the `cesd_gte16` as the outcome) as a function of `mcs` scores. Show a summary of the final fitted model and explain the coefficients. [**REMEMBER** to compute the Odds Ratios after you get the raw coefficient (betas)].

```
model3 <- glm(cesd_gte16 ~mcs, data = h1, family=binomial)
summary(model3)

##
## Call:
## glm(formula = cesd_gte16 ~ mcs, family = binomial, data = h1)
##
```

```
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.04167   0.06727   0.13027   0.29676   1.79914
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   9.2691     1.0621   8.727  < 2e-16 ***
## mcs          -0.1716     0.0219  -7.835 4.68e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 297.59  on 452  degrees of freedom
## Residual deviance: 174.73  on 451  degrees of freedom
## AIC: 178.73
##
## Number of Fisher Scoring iterations: 7
```

```
exp(coef((model3)))
```

```
##  (Intercept)          mcs
## 1.060544e+04 8.423518e-01
```

For every one point increase in MCS, people are 16% less likely to be depressed.

8. Use the `predict()` function like we did in class to predict CESD => 16 and compare it back to the original data. For now, use a cutoff probability of 0.5 - if the probability is > 0.5 consider this to be true and false otherwise. Like we did in class. **REMEMBER** See the R code for the class example at https://github.com/melindahiggins2000/N741_lecture11_27March2018/blob/master/lesson11_logreg_Rcode.R
   - How well did the model correctly predict CESD scores => 16 (indicating depression)? (make the "confusion matrix" and look at the true positives and true negatives versus the false positives and false negatives).

```
m3.predict <- predict(model3, newdata=h1, type = "response") ##why wont this
work?
summary(m3.predict)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1982  0.9042  0.9874  0.8985  0.9961  0.9997
```

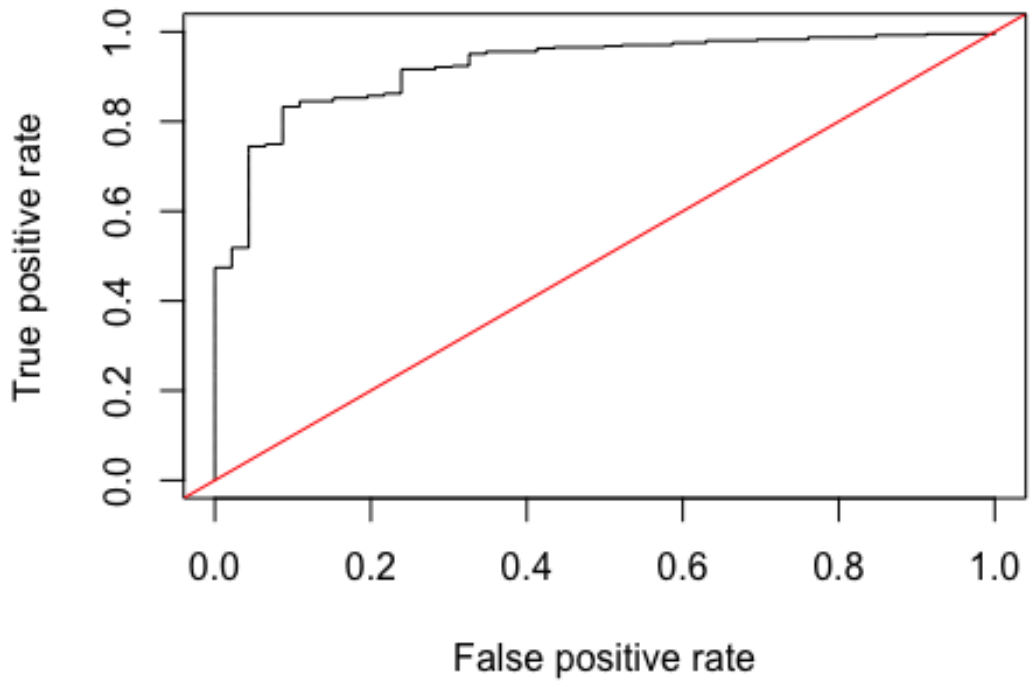
```
##confusion matrix
table(h1$cesd_gte16, m3.predict >.5)
```

```
##
##     FALSE TRUE
##   0    22   24
##   1    12  395
```

The model did a good job, only incorrectly predicted 36 people out of 453 total; 12 people were depressed but were not predicted to be depressed and 24 weren't depressed but were predicted to be depressed.

9. Make an ROC curve plot and compute the AUC and explain if this is a good model for predicting depression or not

```
library(ROCR)

## Loading required package: gplots

##
## Attaching package: 'gplots'

## The following object is masked from 'package:stats':
##
##     lowess

p <- predict(model3, newdata=h1, type="response")
pr <- prediction(p, as.numeric(h1$cesd_gte16))
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
abline(a=0, b=1, col = "red")
```

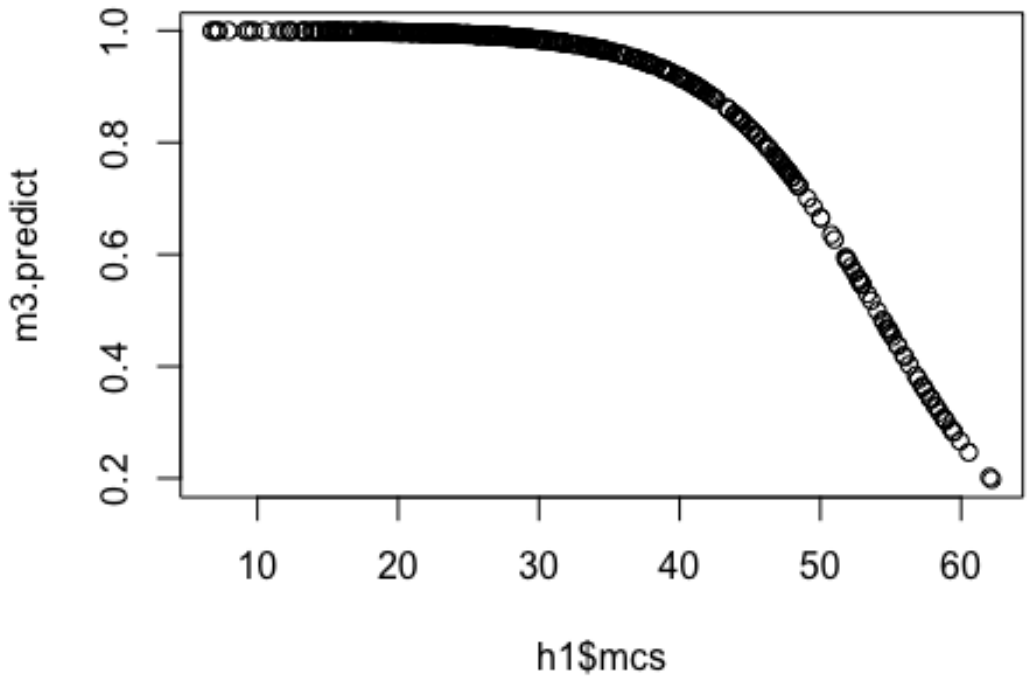


```
##area under the curve
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc

## [1] 0.9221771
```

This is a good model for prediciting because an auc of .5 is a 50/50 prediction (think coin toss) where as an auc of .9 is much stronger.

10. Make a plot showing the probability curve - put the `mcs` values on the X-axis and the probability of depression on the Y-axis. Based on this plot, do you think the `mcs` is a good predictor of depression? [**FYI** This plot is also called an "effect plot" is you're using `Rcmdr` to do these analyses.]

```
plot(h1$mcs, m3.predict)
```



MCS is a fairly good predictor of depression as higher scores on the MCS indicate lower liklihood of depression, and this is exhibited in the model/plot

Use R markdown to complete your homework and show all of your code and output in your final report - Turn in a PDF of your report to Canvas. Include a link to your Github repo for Homework 6
