N741 Spring 2018 - Homework 6

Homework 6 - DUE FRIDAY April 6, 2018

Martha Wetzel

April 4, 2018

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

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

```r
library(tidyverse)
library(haven)
library(stargazer)
library(car)
helpdata <- haven::read_spss("helpmkh.sav")

h1 <- helpdata %>%
  select(age, female, pss_fr, homeless,
         pcs, mcs, cesd)

# create a function to get the label
# label output from the attributes() function
getlabel <- function(x) attributes(x)$label
# getlabel(sub1$age)

library(purrr)
ldf <- purrr::map_df(h1, getlabel) # this is a 1x15 tibble data.frame
# t(ldf) # transpose for easier reading to a 15x1 single column list

# using knitr to get a table of these
# variable names for Rmarkdown
library(knitr)
knitr::kable(t(ldf),
             col.names = c("Variable Label"),
             caption="Use these variables from HELP dataset for Homework 06")
```

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

```r
# 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)
```

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.

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

Question 1

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.

```
cesdm1 <- lm(cesd ~ mcs, h1)
summary(cesdm1)

##
## 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
```

Question 2

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.*

The equation for this model is shown below.

$$CESD = 53.9 - 0.66(mcs)$$

CESD and MCS are inversely related. On average, for a one point increase in mental health quality of life (QoL), depression as measured by the CESD scale decreases by 0.66 points. This intuitively makes sense, as we would expect someone with a high mental health QoL to have less depression. The intercept indicates that if MCS was equal to 0, then CESD would equal 53.9. However, it is important to note that the minimum value in the data set for MCS is 6.7. Thus, the intercept represents a projection out of the data used.

Question 3

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`.

MCS explains 46.5 percent of the variation in CESD.

Question 4

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`

    - Print out the model results with the coefficients and tests and model fit statistics.

```
cesdm2 <- lm(cesd ~ mcs + age + female + pss_fr + homeless + pcs + mcs, h1)
summary(cesdm2)
```

```
##
## 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
```

Question 5

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).

The following variables were significant predictors of CESD score:

- MCS: Similar to the simple model, a one-point increase in the MCS is associated with a decrease in the CESD of 0.62 points on average.

- Female: Compared to a man, the average woman has a higher CESD score by 2.4 points.

- PSS_FR: Higher perceived social support is associated with lower depression. More specifically, a one point increase in the social support scale is associated with a 0.26 point decrease on the CESD scale on average.

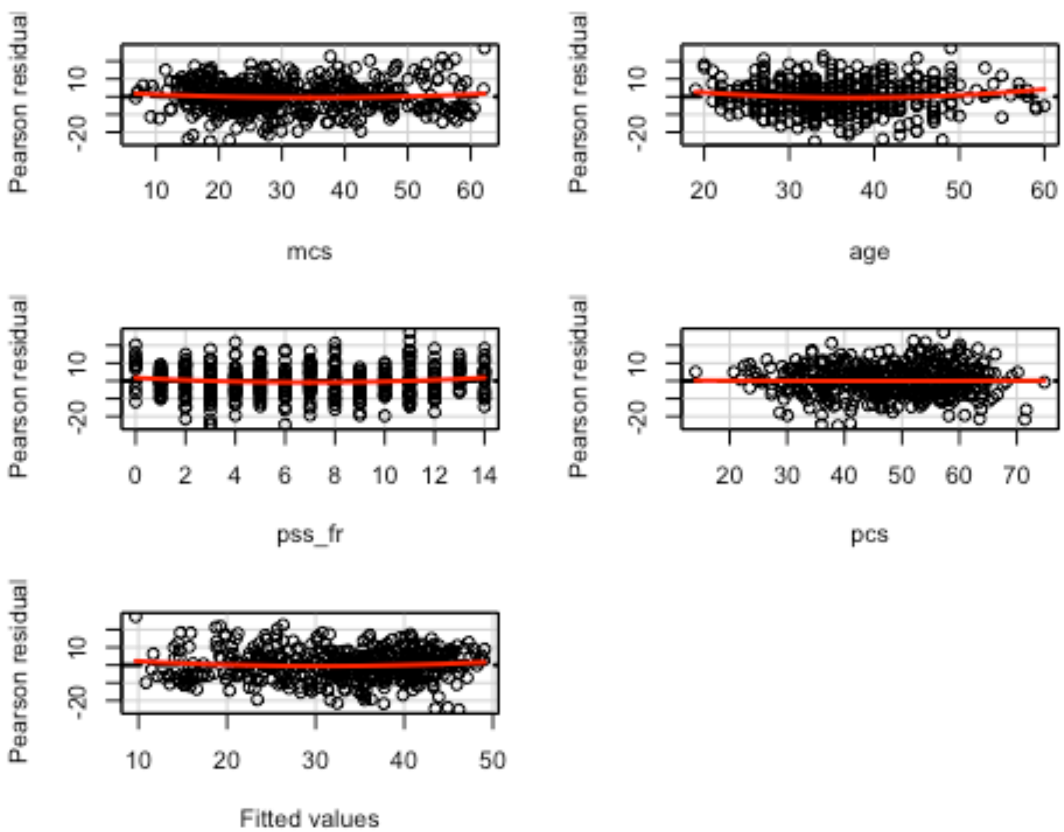
- PCS: Worse physical health is associated with higher depression. A one point increase on the PCS scale is associated with a 0.24 point decrease on the CESD on average.

Question 6

6. Following the example we did in class for the Prestige dataset https://cdn.rawgit.com/vhertzb/2018week9/2f2ea142/2018week9.html?raw=true,

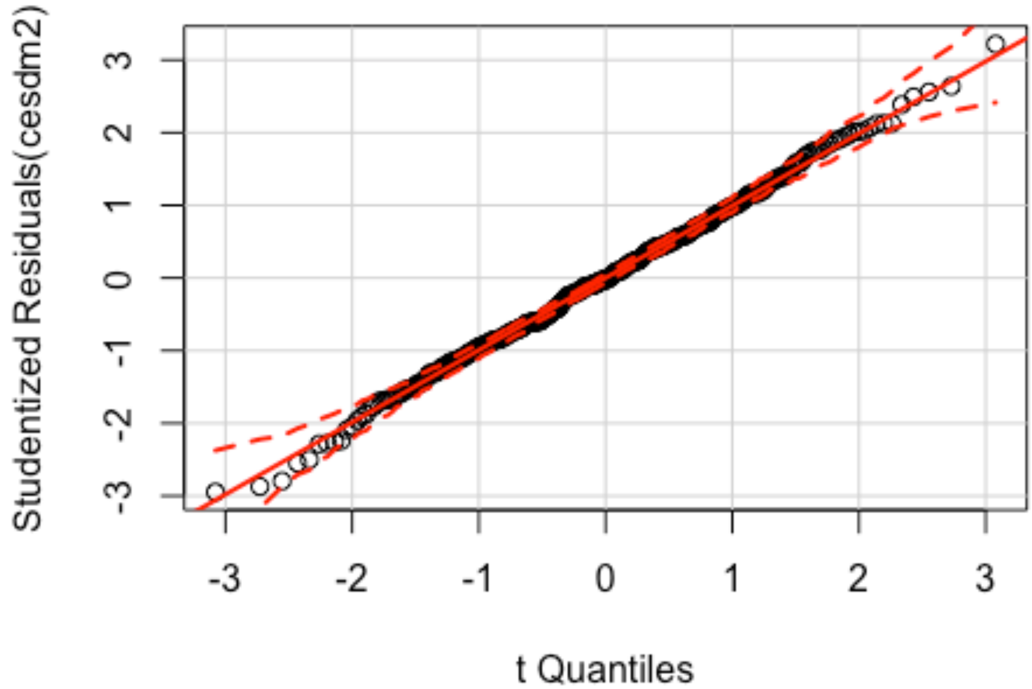
generate the diagnostic plotss for this model with these 6 predictors (e.g. get the residual plot by variables, the added-variable plots, the Q-Q plot, diagnostic plots). Also run the VIFs to check for multicollinearity issues.

```r
residualPlots(cesdm2)
```

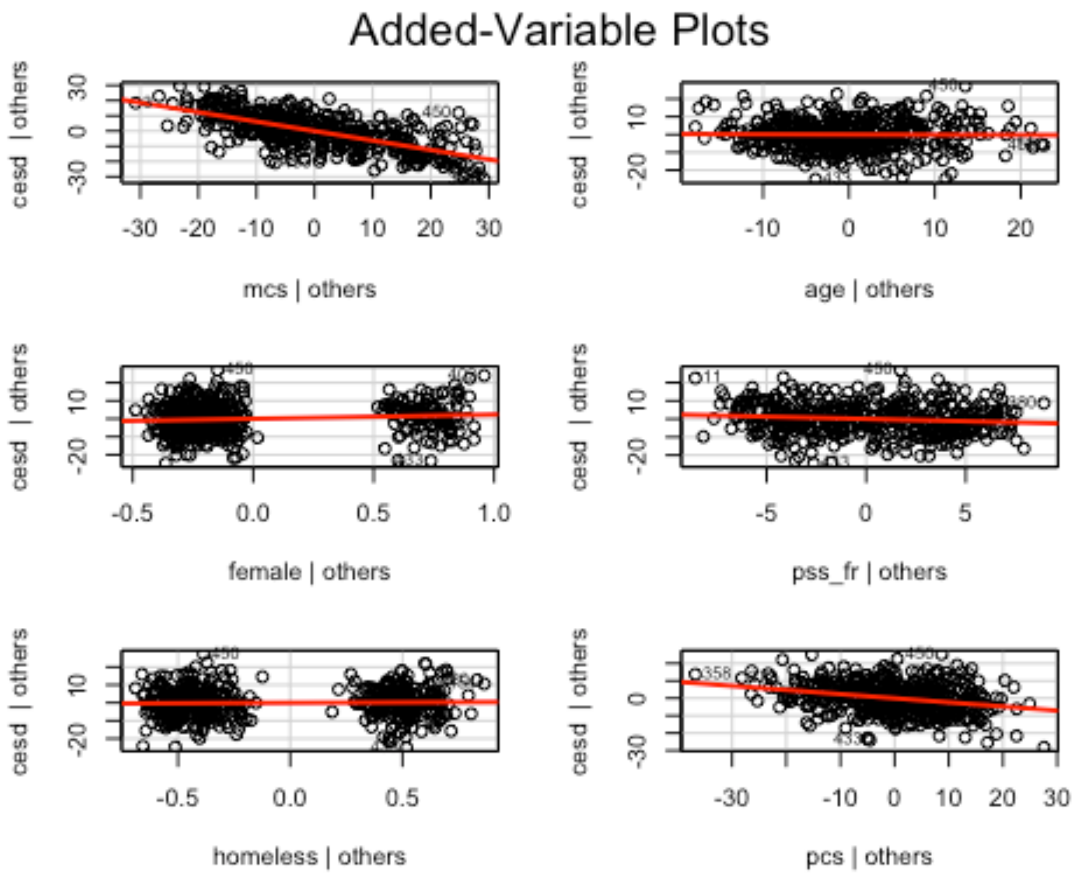


```
##            Test stat Pr(>|t|)
## mcs            1.260    0.208
## age            1.941    0.053
## pss_fr         1.964    0.050
## pcs            0.081    0.936
## Tukey test     1.434    0.152
```

```r
qqPlot(cesdm2)
```




```
avPlots(cesdm2, id.n=2, id.cex=0.7)
```




```
ncvTest(cesdm2)

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

vif(cesdm2)

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

The residual and QQ plots show no issues. There is some evidence of heteroskedasticity, meaning OLS may not be the most efficient model for this data. Some evidence of influential outliers is present; however, data descriptive stats did not show any invalid values. There is no evidence of collinearity.

Question 7

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)].

```
cesdlr1 <- glm(cesd_gte16 ~ mcs, data=h1,
          family=binomial)
summary(cesdlr1)
```

```
##
## 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
```

```
# take the exp to get the odds ratios
exp(coef(cesdlr1))
```

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

A one point increase in MCS score is associated with a 16% decrease in the odds of meeting the criteria for depression.

Question 8

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).

```
cesdpred <- predict(cesdlr1, newdata=h1,
                    type="response")
```

```r
table(h1$cesd_gte16, cesdpred > 0.5)

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

t1 <- table(cesdpred > 0.5, h1$cesd_gte16)
t1 # this shows homeless as the 0/1, and then model predictions are TRUE/FALS
E. This shows false negatives and false positives.

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

tpr <- t1[2,2]/(t1[2,2]+t1[1,2]) # true positive rate
tpr #senstivity

## [1] 0.970516

tnr <- t1[1,1]/(t1[1,1]+t1[2,1]) # true negative rate
tnr #specificity

## [1] 0.4782609
```

The sensitivity, or true positive rate, is very high at 97 percent. The specificity, or true negative rate, is 48 percent.

Question 9

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

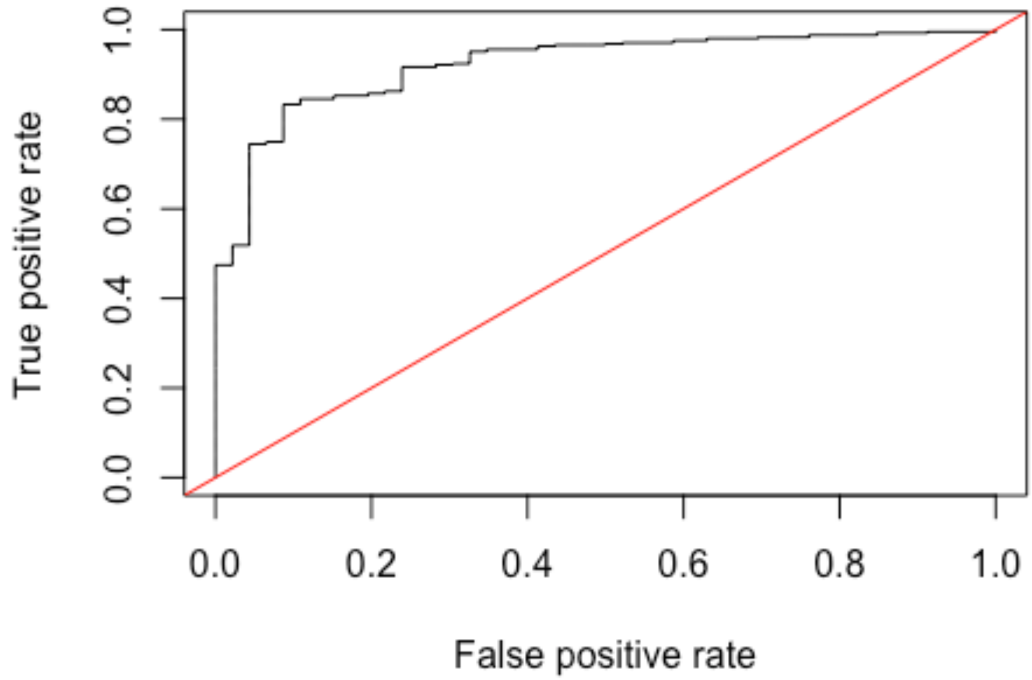
```r
library(ROCR)

## Loading required package: gplots

##
## Attaching package: 'gplots'

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

# Make the ROC curve plot
pr <- prediction(cesdpred, as.numeric(h1$cesd_gte16)) # this is a function to
create "prediction objects" which must be in a standardized format
prf <- performance(pr, measure = "tpr", x.measure = "fpr") #tpr is true posit
ive rate, fpr is false positive rate
plot(prf)
abline(a=0, b=1, col="red")
```



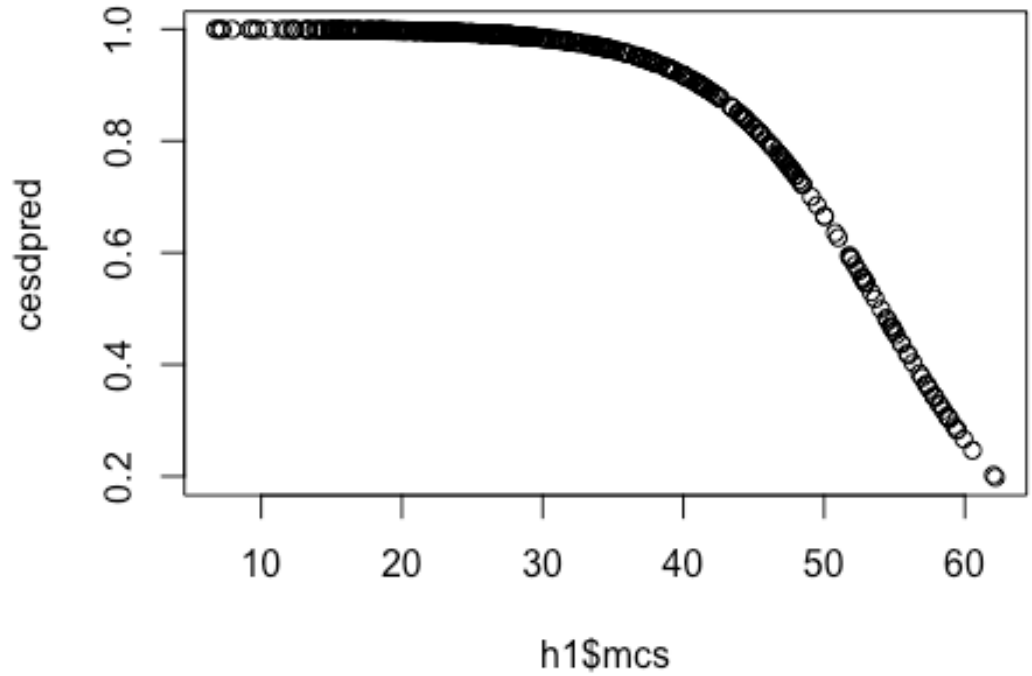

```
# Calculate AUC
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
```

The AUC is 0.922. This is a good model for predicting depression because the AUC is over 0.9.

Question 10

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, cesdpred)
```



This plot shows that there is a high predicted probability of depression when mental health scores are low and a lower probability of depression as mental health scores improve. This makes intuitive sense.

This code can be found on Github
