N741 Spring 2018 - Homework 6

Homework 6 - DUE FRIDAY April 6, 2018

Lacey Gleason

April 5, 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.

### 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:

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.

# 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

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](http://cesd-r.com/) 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.

LinearMod <- lm(formula = cesd ~ mcs, data = h1)  
  
summary(LinearMod)

##   
## 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

1. 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 of the fitted model is cesd = 53.9022 - 0.6647(mcs). The intercept is 53.9022. The slope is -0.6647. For each one unit increase in mcs, there is a 0.6647 unit decrease in CESD score.

1. 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 adjusted R squared is 0.4638. MCS explains 46% of the variation in CESD score.

1. [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.

LinearMod2 <- lm(formula = cesd ~ age + female + pss\_fr + homeless + pcs + mcs, data = h1)  
  
summary(LinearMod2)

##   
## Call:  
## lm(formula = cesd ~ 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 \*\*\*  
## 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 \*\*\*  
## mcs -0.62093 0.03261 -19.042 < 2e-16 \*\*\*  
## ---  
## 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

1. 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 variables in the model that are significant are female, pss\_fr, pcs, and mcs. Being female is associated with a 2.35 increase in CESD score. Every one unit increase in pss\_friend score is associated with a 0.26 decrease in CESD score. Every one unit increase in pcs score is associated with a 0.24 decrease in CESD score. Every one unit increase in mcs score is associated with 0.62 decrease in CESD score.

1. 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.

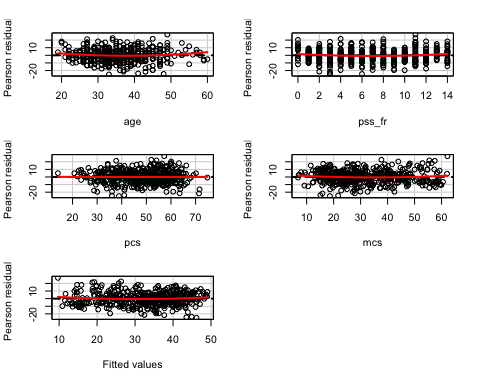
library(car)

##   
## Attaching package: 'car'

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

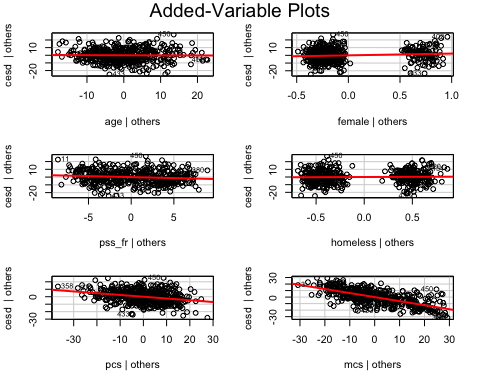
## The following object is masked from 'package:purrr':  
##   
## some

#get the residual plots  
residualPlots(LinearMod2)

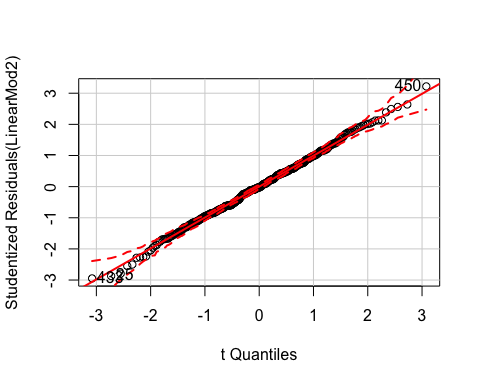


## Test stat Pr(>|t|)  
## age 1.941 0.053  
## pss\_fr 1.964 0.050  
## pcs 0.081 0.936  
## mcs 1.260 0.208  
## Tukey test 1.434 0.152

avPlots(LinearMod2, id.n=2, id.cex=0.7)

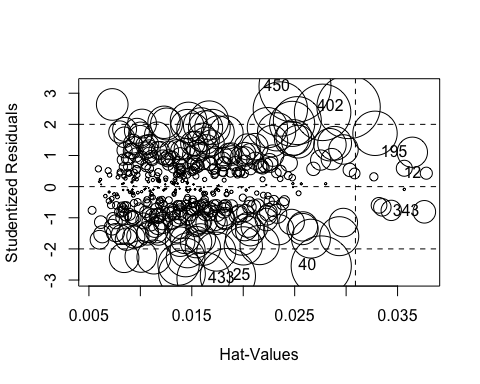


qqPlot(LinearMod2, id.n=3)



## 433 25 450   
## 1 2 453

influencePlot(LinearMod2, id.n=3)



## StudRes Hat CookD  
## 12 0.4313265 0.03779399 0.001045833  
## 25 -2.8712570 0.01850847 0.021854090  
## 40 -2.5490027 0.02758106 0.026006371  
## 195 1.1039353 0.03643745 0.006580280  
## 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

ncvTest(LinearMod2)

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

vif(LinearMod2)

## age female pss\_fr homeless pcs mcs   
## 1.078264 1.058232 1.068213 1.060007 1.108172 1.050768

1. [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)].

model1 <- glm(cesd\_gte16 ~ mcs, data=h1, family=binomial)  
  
model1

##   
## Call: glm(formula = cesd\_gte16 ~ mcs, family = binomial, data = h1)  
##   
## Coefficients:  
## (Intercept) mcs   
## 9.2691 -0.1716   
##   
## Degrees of Freedom: 452 Total (i.e. Null); 451 Residual  
## Null Deviance: 297.6   
## Residual Deviance: 174.7 AIC: 178.7

summary(model1)

##   
## 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

# coefficients of the model - these are the  
# RAW Betas   
coef(model1)

## (Intercept) mcs   
## 9.2691224 -0.1715576

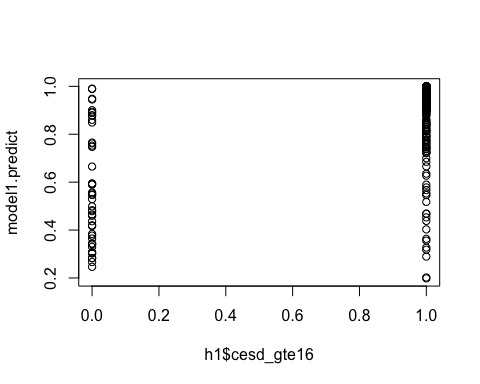
# take the exp to get the odds ratios  
exp(coef(model1))

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

The odds ratio is 0.84. This means that for everyone one unit increase in MCS, the odds of depression are 16% lower.

1. 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.
   * 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).

# look at the predicted probabilities  
# review the help for predict.glm  
model1.predict <- predict(model1, newdata=h1,  
 type="response")  
  
plot(h1$cesd\_gte16, model1.predict)



table(h1$cesd\_gte16, model1.predict > 0.5)

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

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

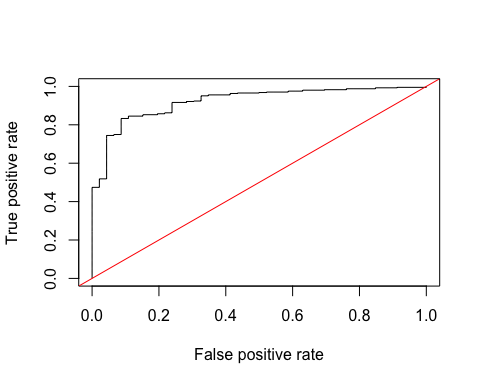
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

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

p <- predict(model1, 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")



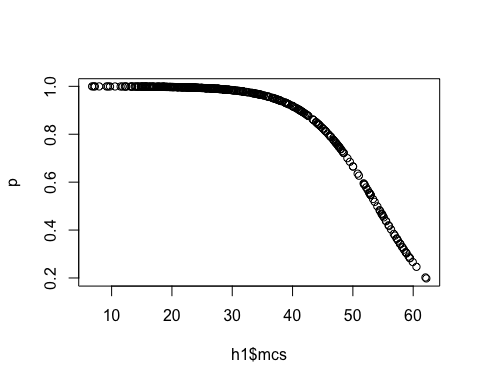
auc <-performance (pr, measure = "auc")  
auc <- auc@y.values[[1]]  
auc

## [1] 0.9221771

Yes, this is a good model for predicting depression because the AUC is 0.92, which means that the classifer is relatively good at separating out positive and negative values for depression.

1. 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” if you’re using Rcmdr to do these analyses.]

plot(h1$mcs, p)



**Link to your Github repo for Homework 6** <https://github.com/lpgleason/N741Spring2018_Homework6.git>