Homework 6: GLM and LM, prediction

Nicole Carlson.

## 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:

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.

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.

library(car)

##   
## Attaching package: 'car'

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

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

library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

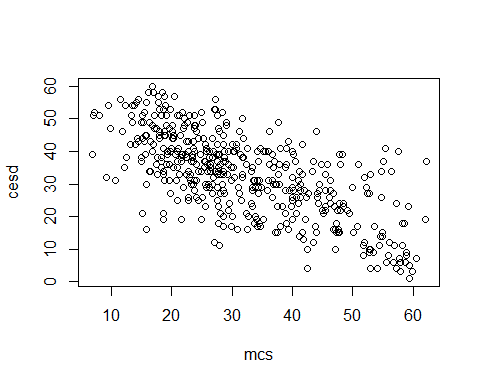
## R package version 5.2.1. https://CRAN.R-project.org/package=stargazer

h1 <- readRDS("h1.rds")  
unajustedCESD <- lm(cesd ~ mcs, data = h1)  
summary(unajustedCESD)

##   
## 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.*

plot(cesd ~ mcs, data = h1)

 Equation for final fitted model: cesd= 53.90 - 0.66 (msc)

Interpretation: The mean cesd score is 53.90 when mcs score is equal to 0. For each increase in mcs of 1 point, the cesd score decreases by 0.66 points. You can see from the plot above that there is a linear relationship beteween these two variables, and most of the cases in this dataset have mcs scores below the ‘normal’ level of 50.

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.

MCS score explains 46% of the variability in the 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.

ajustedCESD <- lm(cesd ~ ., data = h1)  
summary(ajustedCESD)

##   
## Call:  
## lm(formula = cesd ~ ., data = h1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -24.2355 -4.8786 0.3747 4.7882 20.0050   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.31517 3.66021 11.834 < 2e-16 \*\*\*  
## age 0.01319 0.05001 0.264 0.79203   
## female 2.78914 0.89816 3.105 0.00202 \*\*   
## pss\_fr -0.20686 0.09606 -2.153 0.03182 \*   
## homeless 0.53040 0.76499 0.693 0.48846   
## pcs -0.18016 0.03665 -4.916 1.24e-06 \*\*\*  
## mcs -0.45830 0.03393 -13.506 < 2e-16 \*\*\*  
## cesd\_gte16 14.15701 1.44383 9.805 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.883 on 445 degrees of freedom  
## Multiple R-squared: 0.6093, Adjusted R-squared: 0.6032   
## F-statistic: 99.15 on 7 and 445 DF, p-value: < 2.2e-16

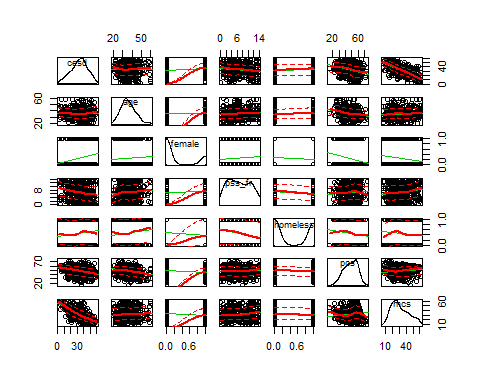
1. Which variables are significant in the model? Write a sentence or two describing the impact of these variables for predicting depression scores (HINT: interpret the coefficient terms).

In this adjusted model of cesd, the significant predictors were: female gender, pss\_fr, pcs, mcs, and cesd\_gte16.

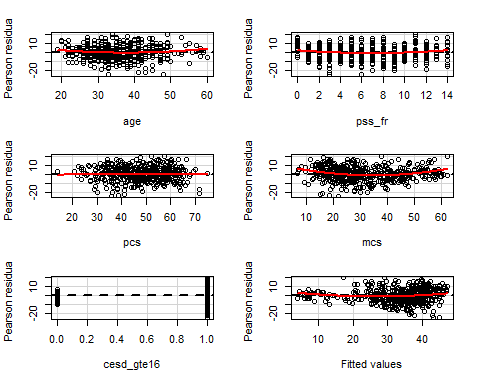
1. Following the example we did in class for the Prestige dataset <https://cdn.rawgit.com/vhertzb/2018week9/2f2ea142/2018week9.html?raw=true>, generate the diagnostic plotss for this model with these 6 predictors (e.g. get the residual plot by variables, the added-variable plots, the Q-Q plot, diagnostic plots). Also run the VIFs to check for multicollinearity issues.

scatterplotMatrix(~ cesd + age + female + pss\_fr + + homeless + pcs + mcs, span =0.7, data = h1)

## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth  
  
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth  
  
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth  
  
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth  
  
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth  
  
## Warning in smoother(x, y, col = col[2], log.x = FALSE, log.y = FALSE,  
## spread = spread, : could not fit smooth

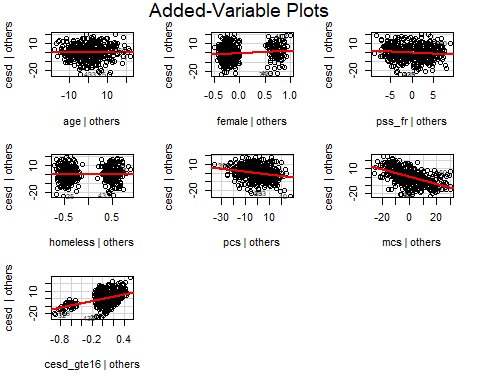


residualPlots(ajustedCESD)

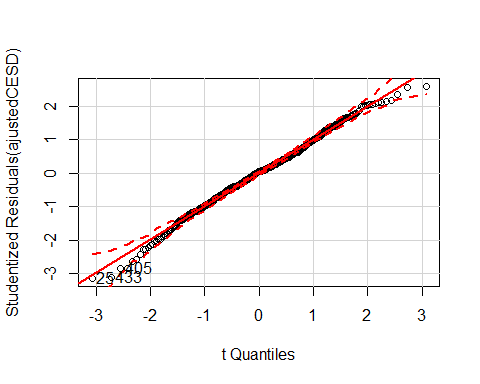


## Test stat Pr(>|t|)  
## age 1.954 0.051  
## pss\_fr 2.511 0.012  
## pcs -0.391 0.696  
## mcs 4.984 0.000  
## cesd\_gte16 -1.504 0.133  
## Tukey test 4.100 0.000

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



qqPlot(ajustedCESD, id.n=3)

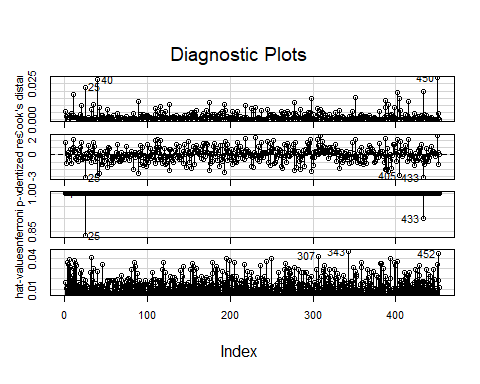


## 25 433 405   
## 1 2 3

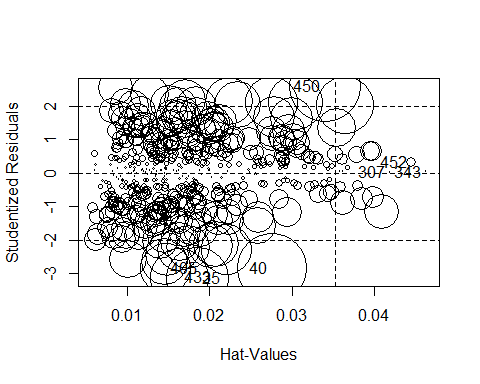
outlierTest(ajustedCESD)

##   
## No Studentized residuals with Bonferonni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferonni p  
## 25 -3.133781 0.00184 0.83354

influenceIndexPlot(ajustedCESD, id.n=3)



influencePlot(ajustedCESD, id.n=3)



## StudRes Hat CookD  
## 25 -3.13378120 0.01852040 2.271392e-02  
## 40 -2.82863028 0.02758366 2.793075e-02  
## 307 0.02864576 0.04198018 4.504813e-06  
## 343 0.04345031 0.04633582 1.149192e-05  
## 405 -2.84590477 0.01460759 1.477224e-02  
## 433 -3.11008039 0.01631893 1.967470e-02  
## 450 2.59355446 0.03406544 2.927615e-02  
## 452 0.31723229 0.04458614 5.882362e-04

ncvTest(ajustedCESD)

## Non-constant Variance Score Test   
## Variance formula: ~ fitted.values   
## Chisquare = 13.47059 Df = 1 p = 0.0002423324

vif(ajustedCESD)

## age female pss\_fr homeless pcs mcs   
## 1.081464 1.060867 1.071092 1.060086 1.135996 1.380623   
## cesd\_gte16   
## 1.386363

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)].

logisticCESD <- glm(cesd\_gte16 ~ mcs, data = h1)  
summary(logisticCESD)

##   
## Call:  
## glm(formula = cesd\_gte16 ~ mcs, data = h1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -0.95274 -0.05516 0.03689 0.15255 0.46512   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.2760656 0.0326857 39.04 <2e-16 \*\*\*  
## mcs -0.0119208 0.0009564 -12.46 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for gaussian family taken to be 0.06816063)  
##   
## Null deviance: 41.329 on 452 degrees of freedom  
## Residual deviance: 30.740 on 451 degrees of freedom  
## AIC: 72.847  
##   
## Number of Fisher Scoring iterations: 2

exp(cbind(OR = coef(logisticCESD), confint(logisticCESD)))

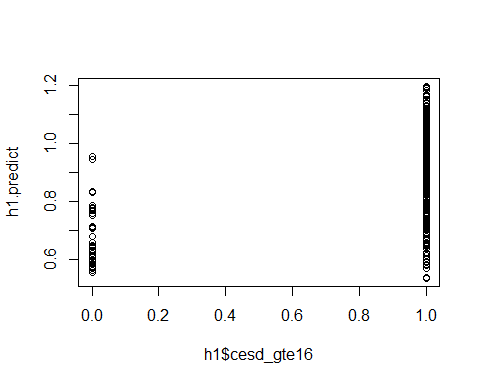
## Waiting for profiling to be done...

## OR 2.5 % 97.5 %  
## (Intercept) 3.582517 3.3602081 3.8195336  
## mcs 0.988150 0.9862994 0.9900041

In this dataset, mcs is a significant predictor of cesd score of at least 16. The odds ratio of mcs is 0.98, meaning that for every 1 point increase in a person’s mcs score, their odds of having a cesd score of at least 16 is decreased by 2%–in other words, their risk of depression (as diagnosed by a score of at least 16 on the CESD) is slightly decreased by having a higher mcs score.

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. **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).

h1.predict <- predict(logisticCESD, newdata=h1,  
 type="response")  
plot(h1$cesd\_gte16, h1.predict)



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

##   
## TRUE  
## 0 46  
## 1 407

Based on this confusion matrix results, the model was pretty good at predicting a cesd score of greater than or equal to 16. Prediction was that the cesd result would be in the depression range for only 46 cases that were actually not depressed (cesd < 16), while 407 of the guesses were correct.

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)

## Warning: package 'ROCR' was built under R version 3.4.4

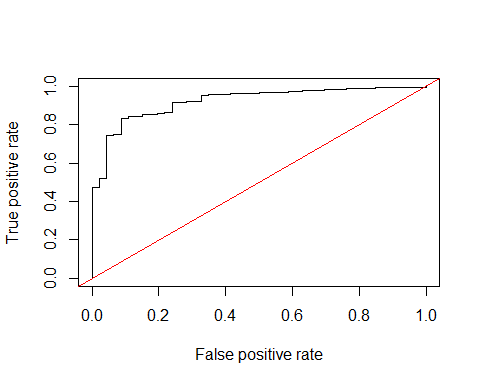
## Loading required package: gplots

## Warning: package 'gplots' was built under R version 3.4.4

##   
## Attaching package: 'gplots'

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

p <- predict(logisticCESD, 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

This model, with only one predictor (msc score) did very well at predicting cesd less than 16; 92.3% of true positives were predicted by msc score.

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

library(Rcmdr)

## Loading required package: splines

## Loading required package: RcmdrMisc

## Loading required package: sandwich

## Loading required package: effects

## Loading required package: carData

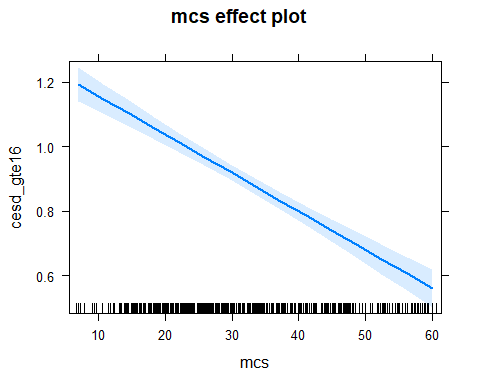
##   
## Attaching package: 'carData'

## The following objects are masked from 'package:car':  
##   
## Guyer, UN, Vocab

## lattice theme set by effectsTheme()  
## See ?effectsTheme for details.

## The Commander GUI is launched only in interactive sessions

plot(allEffects(logisticCESD))

 Here is the effects plot for this logistic regression. Based on the tight confidence interval and grouping of mcs scores around a cesd\_gte16 value of 1 on the plotted line, this plot demonstrates thata mcs scores (especially those between 17 and 33, where there is a clumping of points on the x-axis) appear to be predictive of depression.

My work can be found at github repo: <https://github.com/nicolecarlson/N741Spring2018_Homework7>

**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**