HW 6

Gaea Daniel

11/22/2017

## Assignment

library(tidyverse)

## Loading tidyverse: ggplot2  
## Loading tidyverse: tibble  
## Loading tidyverse: tidyr  
## Loading tidyverse: readr  
## Loading tidyverse: purrr  
## Loading tidyverse: dplyr

## Conflicts with tidy packages ----------------------------------------------

## filter(): dplyr, stats  
## lag(): dplyr, stats

library(haven)  
  
helpdat <- haven::read\_spss("helpmkh.sav")  
  
h1 <- helpdat %>%  
 select(g1b, homeless, age, female, pss\_fr,  
 pcs, mcs, cesd, indtot)  
  
library(psych)

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

psych::corr.test(h1, method="pearson")

## Call:psych::corr.test(x = h1, method = "pearson")  
## Correlation matrix   
## g1b homeless age female pss\_fr pcs mcs cesd indtot  
## g1b 1.00 0.13 -0.06 0.15 -0.09 -0.15 -0.32 0.34 0.17  
## homeless 0.13 1.00 0.09 -0.10 -0.17 -0.10 -0.07 0.09 0.22  
## age -0.06 0.09 1.00 0.04 0.08 -0.23 0.04 0.01 0.03  
## female 0.15 -0.10 0.04 1.00 0.07 -0.16 -0.12 0.18 -0.26  
## pss\_fr -0.09 -0.17 0.08 0.07 1.00 0.08 0.14 -0.18 -0.20  
## pcs -0.15 -0.10 -0.23 -0.16 0.08 1.00 0.11 -0.29 -0.13  
## mcs -0.32 -0.07 0.04 -0.12 0.14 0.11 1.00 -0.68 -0.38  
## cesd 0.34 0.09 0.01 0.18 -0.18 -0.29 -0.68 1.00 0.34  
## indtot 0.17 0.22 0.03 -0.26 -0.20 -0.13 -0.38 0.34 1.00  
## Sample Size   
## [1] 453  
## Probability values (Entries above the diagonal are adjusted for multiple tests.)   
## g1b homeless age female pss\_fr pcs mcs cesd indtot  
## g1b 0.00 0.08 1.00 0.03 0.64 0.03 0.00 0.0 0.01  
## homeless 0.00 0.00 0.70 0.53 0.01 0.55 1.00 0.7 0.00  
## age 0.23 0.07 0.00 1.00 0.80 0.00 1.00 1.0 1.00  
## female 0.00 0.04 0.36 0.00 1.00 0.02 0.18 0.0 0.00  
## pss\_fr 0.05 0.00 0.09 0.15 0.00 0.83 0.06 0.0 0.00  
## pcs 0.00 0.04 0.00 0.00 0.10 0.00 0.28 0.0 0.07  
## mcs 0.00 0.15 0.34 0.01 0.00 0.02 0.00 0.0 0.00  
## cesd 0.00 0.06 0.86 0.00 0.00 0.00 0.00 0.0 0.00  
## indtot 0.00 0.00 0.57 0.00 0.00 0.00 0.00 0.0 0.00  
##   
## To see confidence intervals of the correlations, print with the short=FALSE option

### Complete the following:

#### 1. Consider the continuous variable cesd as a predictor for g1b

##### a. run a logistic regression of the probability of suicidal thoughts (g1b) given their depressive symptoms scores (cesd)

m1 <- glm(g1b ~ cesd, data=h1,  
 family=binomial)  
  
m1

##   
## Call: glm(formula = g1b ~ cesd, family = binomial, data = h1)  
##   
## Coefficients:  
## (Intercept) cesd   
## -3.43335 0.07091   
##   
## Degrees of Freedom: 452 Total (i.e. Null); 451 Residual  
## Null Deviance: 537.5   
## Residual Deviance: 481.4 AIC: 485.4

summary(m1)

##   
## Call:  
## glm(formula = g1b ~ cesd, family = binomial, data = h1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4443 -0.8081 -0.5709 1.0716 2.1883   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.43335 0.40434 -8.491 < 2e-16 \*\*\*  
## cesd 0.07091 0.01049 6.762 1.36e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 537.52 on 452 degrees of freedom  
## Residual deviance: 481.38 on 451 degrees of freedom  
## AIC: 485.38  
##   
## Number of Fisher Scoring iterations: 4

coef(m1)

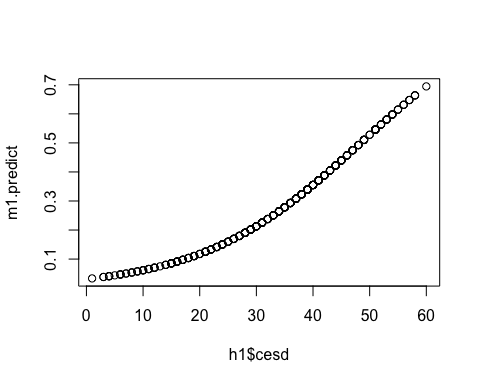
## (Intercept) cesd   
## -3.43334975 0.07091097

exp(coef(m1))

## (Intercept) cesd   
## 0.03227863 1.07348565

##### b. make a plot of the the predicted probability of suicidal thoughts (g1b) by the depressive symptoms scores (cesd)

m1.predict <- predict(m1, newdata=h1,  
 type="response")  
  
plot(h1$cesd, m1.predict)



##### c. what value of the cesd leads to a probability of suicidal thoughts => 0.5? (hint: use the plot you just made)

47.5

#confusion matrix  
table(h1$g1b, m1.predict > 0.5)

##   
## FALSE TRUE  
## 0 299 27  
## 1 103 24

library(gmodels)  
CrossTable(h1$g1b, m1.predict > 0.5)

##   
##   
## Cell Contents  
## |-------------------------|  
## | N |  
## | Chi-square contribution |  
## | N / Row Total |  
## | N / Col Total |  
## | N / Table Total |  
## |-------------------------|  
##   
##   
## Total Observations in Table: 453   
##   
##   
## | m1.predict > 0.5   
## h1$g1b | FALSE | TRUE | Row Total |   
## -------------|-----------|-----------|-----------|  
## 0 | 299 | 27 | 326 |   
## | 0.325 | 2.565 | |   
## | 0.917 | 0.083 | 0.720 |   
## | 0.744 | 0.529 | |   
## | 0.660 | 0.060 | |   
## -------------|-----------|-----------|-----------|  
## 1 | 103 | 24 | 127 |   
## | 0.835 | 6.583 | |   
## | 0.811 | 0.189 | 0.280 |   
## | 0.256 | 0.471 | |   
## | 0.227 | 0.053 | |   
## -------------|-----------|-----------|-----------|  
## Column Total | 402 | 51 | 453 |   
## | 0.887 | 0.113 | |   
## -------------|-----------|-----------|-----------|  
##   
##

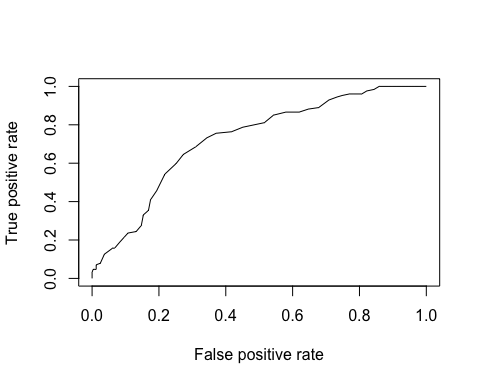
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

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

p <- predict(m1, newdata=h1,   
 type="response")  
pr <- prediction(p, as.numeric(h1$g1b))  
prf <- performance(pr, measure = "tpr", x.measure = "fpr")  
plot(prf)



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

## [1] 0.7227429

#To access R-commander  
library(Rcmdr)

## Loading required package: splines

## Loading required package: RcmdrMisc

## Loading required package: car

##   
## Attaching package: 'car'

## The following object is masked from 'package:psych':  
##   
## logit

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

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

## 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

library(splines)  
library(RcmdrMisc)  
library(car)

#### 2. Using variable selection methods, develop a logistic regression model for the probability of suicidal thoughts (g1b) considering all of these variables for possible inclusion: age, female, pss\_fr, homeless, pcs, mcs, cesd, indtot

##### a. present the final model results

names(h1) <- make.names(names(h1))  
GLM.1 <- glm(g1b ~ cesd, family=binomial(logit), data=h1)  
summary(GLM.1)

##   
## Call:  
## glm(formula = g1b ~ cesd, family = binomial(logit), data = h1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.4443 -0.8081 -0.5709 1.0716 2.1883   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.43335 0.40434 -8.491 < 2e-16 \*\*\*  
## cesd 0.07091 0.01049 6.762 1.36e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 537.52 on 452 degrees of freedom  
## Residual deviance: 481.38 on 451 degrees of freedom  
## AIC: 485.38  
##   
## Number of Fisher Scoring iterations: 4

exp(coef(GLM.1)) # Exponentiated coefficients ("odds ratios")

## (Intercept) cesd   
## 0.03227863 1.07348565

GLM.2 <- glm(g1b ~ age + female + pss\_fr + homeless + pcs + mcs   
 + cesd + indtot, family=binomial(logit), data=h1)  
summary(GLM.2)

##   
## Call:  
## glm(formula = g1b ~ age + female + pss\_fr + homeless + pcs +   
## mcs + cesd + indtot, family = binomial(logit), data = h1)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6584 -0.7946 -0.5152 0.9239 2.3715   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -0.880372 1.594090 -0.552 0.58076   
## age -0.028233 0.015547 -1.816 0.06938 .   
## female 0.696841 0.288047 2.419 0.01555 \*   
## pss\_fr -0.005106 0.030422 -0.168 0.86670   
## homeless 0.529576 0.240341 2.203 0.02756 \*   
## pcs -0.016347 0.011765 -1.389 0.16469   
## mcs -0.037111 0.013825 -2.684 0.00727 \*\*  
## cesd 0.035462 0.014040 2.526 0.01154 \*   
## indtot 0.031003 0.022454 1.381 0.16735   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 537.52 on 452 degrees of freedom  
## Residual deviance: 455.35 on 444 degrees of freedom  
## AIC: 473.35  
##   
## Number of Fisher Scoring iterations: 5

exp(coef(GLM.2)) # Exponentiated coefficients ("odds ratios")

## (Intercept) age female pss\_fr homeless pcs   
## 0.4146287 0.9721622 2.0074022 0.9949065 1.6982120 0.9837858   
## mcs cesd indtot   
## 0.9635692 1.0360981 1.0314890

stepwise(GLM.2, direction='forward', criterion='AIC')

##   
## Direction: forward  
## Criterion: AIC   
##   
## Start: AIC=539.52  
## g1b ~ 1  
##   
## Df Deviance AIC  
## + cesd 1 481.38 485.38  
## + mcs 1 487.23 491.23  
## + indtot 1 522.59 526.59  
## + pcs 1 527.69 531.69  
## + female 1 527.73 531.73  
## + homeless 1 529.61 533.61  
## + pss\_fr 1 533.76 537.76  
## <none> 537.52 539.52  
## + age 1 536.07 540.07  
##   
## Step: AIC=485.38  
## g1b ~ cesd  
##   
## Df Deviance AIC  
## + mcs 1 472.66 478.66  
## + homeless 1 476.00 482.00  
## + female 1 478.14 484.14  
## + indtot 1 478.39 484.39  
## + age 1 479.05 485.05  
## <none> 481.38 485.38  
## + pcs 1 480.06 486.06  
## + pss\_fr 1 480.88 486.88  
##   
## Step: AIC=478.66  
## g1b ~ cesd + mcs  
##   
## Df Deviance AIC  
## + homeless 1 467.01 475.01  
## + female 1 469.20 477.20  
## + pcs 1 469.94 477.94  
## <none> 472.66 478.66  
## + age 1 470.90 478.90  
## + indtot 1 471.28 479.28  
## + pss\_fr 1 472.12 480.12  
##   
## Step: AIC=475.01  
## g1b ~ cesd + mcs + homeless  
##   
## Df Deviance AIC  
## + female 1 462.11 472.11  
## + pcs 1 464.94 474.94  
## <none> 467.01 475.01  
## + age 1 465.03 475.03  
## + indtot 1 466.57 476.57  
## + pss\_fr 1 466.93 476.93  
##   
## Step: AIC=472.11  
## g1b ~ cesd + mcs + homeless + female  
##   
## Df Deviance AIC  
## + indtot 1 459.69 471.69  
## + age 1 459.85 471.85  
## <none> 462.11 472.11  
## + pcs 1 460.83 472.83  
## + pss\_fr 1 461.89 473.89  
##   
## Step: AIC=471.69  
## g1b ~ cesd + mcs + homeless + female + indtot  
##   
## Df Deviance AIC  
## + age 1 457.32 471.32  
## <none> 459.69 471.69  
## + pcs 1 458.79 472.79  
## + pss\_fr 1 459.61 473.61  
##   
## Step: AIC=471.32  
## g1b ~ cesd + mcs + homeless + female + indtot + age  
##   
## Df Deviance AIC  
## <none> 457.32 471.32  
## + pcs 1 455.38 471.38  
## + pss\_fr 1 457.28 473.28

##   
## Call: glm(formula = g1b ~ cesd + mcs + homeless + female + indtot +   
## age, family = binomial(logit), data = h1)  
##   
## Coefficients:  
## (Intercept) cesd mcs homeless female   
## -2.32289 0.04067 -0.03343 0.56484 0.76358   
## indtot age   
## 0.03442 -0.02291   
##   
## Degrees of Freedom: 452 Total (i.e. Null); 446 Residual  
## Null Deviance: 537.5   
## Residual Deviance: 457.3 AIC: 471.3

library(olsrr)

##   
## Attaching package: 'olsrr'

## The following object is masked from 'package:datasets':  
##   
## rivers

model <- lm(g1b ~ cesd + mcs + homeless + female + indtot + age, data = h1)  
ols\_all\_subset(model)

## # A tibble: 63 x 6  
## Index N Predictors `R-Square` `Adj. R-Square` `Mallow's Cp`  
## <int> <int> <chr> <chr> <chr> <chr>  
## 1 1 1 cesd 0.11335 0.11138 19.64296  
## 2 2 1 mcs 0.09963 0.09764 26.89202  
## 3 3 1 indtot 0.02922 0.02707 64.10963  
## 4 4 1 female 0.02263 0.02047 67.59007  
## 5 5 1 homeless 0.01747 0.01529 70.32012  
## 6 6 1 age 0.00315 0.00094 77.88691  
## 7 7 2 cesd mcs 0.12719 0.12331 14.32527  
## 8 8 2 cesd homeless 0.12401 0.12011 16.00929  
## 9 9 2 cesd female 0.12171 0.11780 17.22525  
## 10 10 2 cesd indtot 0.11713 0.11321 19.64345  
## # ... with 53 more rows

#To determine how often the classifer is correct  
print((24+299)/(24+299+27+103))

## [1] 0.7130243

#To determine how often the classifer is incorrect  
print((27+103)/(24+299+27+103))

## [1] 0.2869757

#To determine how often the prediction is correct when the actual value is positive  
print(24/(103+24))

## [1] 0.1889764

#To determine how often the prediction is correct when the actual value is negative  
print(299/(299+27))

## [1] 0.9171779

#### b. write a few sentences describing your results including:

##### i. model fit

Stepwise model selection was used in R Commander. "Forward" was the chosen direction and "AIC" was the chosen criterion. As variables were added to the model, the AIC deceased. The model the included the variables age, female, homeless, mcs, cesd, and indtot yielded the highest adjusted r^2 (0.14484) and lowest AIC (471.3).

##### ii. model classification table results - remember to report the threshold used for the classification table - you can change it from 0.5 if you think a different threshold might work better

When the threshold used for the classification table is 0.5, the classifer is correct 71.3% of the time. When the actual value is positive, the prediction is correct 18.9% of the time. When the actual value is negative, the prediction is correct 91.7% of the time.

##### iii. odds ratios for each significant predictor in the model

For a one unit decrease in experiencing serious thoughts of suicide in the last 30 days (g1b), the odds of the total score of the CESD (cesd) increase by a factor of 0.041, the odds of the total score of the SF36 mental composite score (mcs) decrease by 0.033, the odds of one or more nights on the street or in a shelter in the past six months (homeless) increase by 0.565, the odds of being a female (female) increase by 0.764, the odds of the total score of the InDue (indtot) increase by 0.034, and the odds of age in years (age) decrease by 0.023.