N741 Spring 2018 - Homework 6

Homework 6

Irene Yang

April 6, 2018

## Homework 6 Assignment

# load libraries and dataset  
  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.2.1 ──

## ✔ ggplot2 2.2.1 ✔ purrr 0.2.4  
## ✔ tibble 1.4.2 ✔ dplyr 0.7.4  
## ✔ tidyr 0.8.0 ✔ stringr 1.3.0  
## ✔ readr 1.1.1 ✔ forcats 0.3.0

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

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

## 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.

RegModel.1 <- lm(cesd~mcs, data=h1)

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.*

summary(RegModel.1)

##   
## 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

The regression equation, therefore, is: cesd = 53.9 - (.67)mcs

An intercept of 53.9 indicates that if an individual scored “0” on the mcs (meaning the worst mental health score possible), their cesd score would be 53.9 and strongly indicative of severe depression. The coefficient of -.67 indicates that for every one point increase in mental health quality of life (mcs), the cesd score will drop by .67.

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 R Square of .465 explains how much variance in cesd score is explained by this model, i.e. 46.5% of the variance in cesd score is explained by this model.

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.

LinearModel.1 <- lm(cesd ~ age + female + pss\_fr + homeless + pcs + mcs, data=h1)  
  
summary(LinearModel.1)

##   
## 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

# To generate standardized parameter estimates...  
library(QuantPsyc)

## Loading required package: boot

## Loading required package: MASS

##   
## Attaching package: 'MASS'

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

##   
## Attaching package: 'QuantPsyc'

## The following object is masked from 'package:base':  
##   
## norm

lm.beta(LinearModel.1)

## age female pss\_fr homeless pcs   
## -0.008306522 0.079858084 -0.081623185 0.018561288 -0.203713784   
## mcs   
## -0.637051613

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 following variables were significant in the model: female, pss\_fr, pcs, and mcs. This means that **controlling for all other variables:** 1. Being female is associated with an **increase** in cesd score of 2.35 2. Having a one unit increase in the scale that measures perceived social support from friends (pss\_fr) is associated with a **decrease** in cesd score of .26 3. Having a one unit increase in the pcs (physical component of the SF-36 - a generic indicator of health status) is associated with a **decrease** in cesd score of .24 4. Having a one unit increase in mcs (mental health quality of life) is associated with a **decrese** in cesd score of .62

The standardized estimates suggest that mental health quality of life has the highest impact on cesd scores.

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)

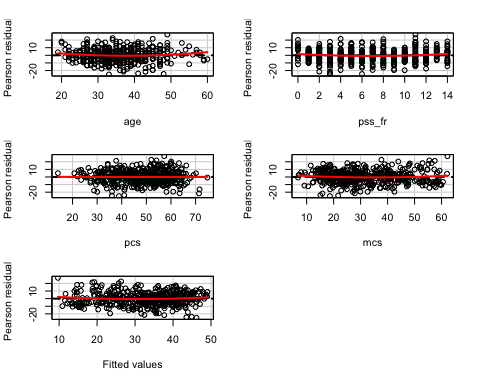
##   
## Attaching package: 'car'

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

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

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

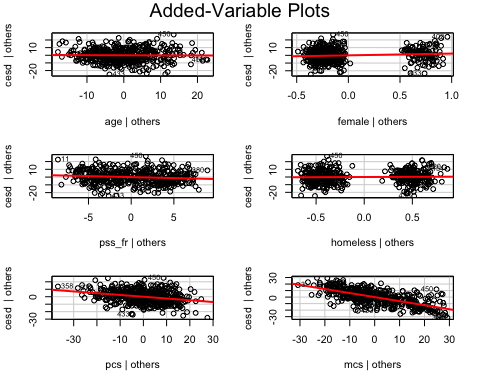
residualPlots(LinearModel.1)



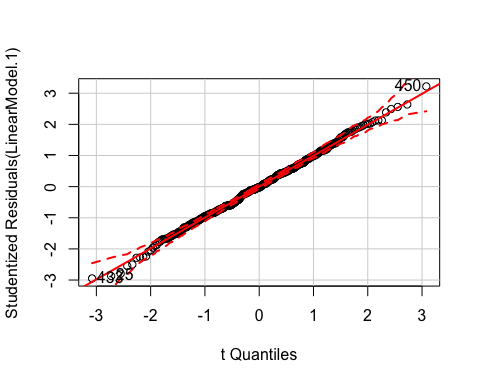
## Test stat Pr(>|t|)  
## age 1.941 0.053  
## pss\_fr 1.964 0.050  
## pcs 0.081 0.936  
## mcs 1.260 0.208  
## Tukey test 1.434 0.152

These plots look good with no evidence of a non-zero trend.

avPlots(LinearModel.1, id.n=2, id.cex=0.7)



# Q-Q plot  
qqPlot(LinearModel.1, id.n=3)



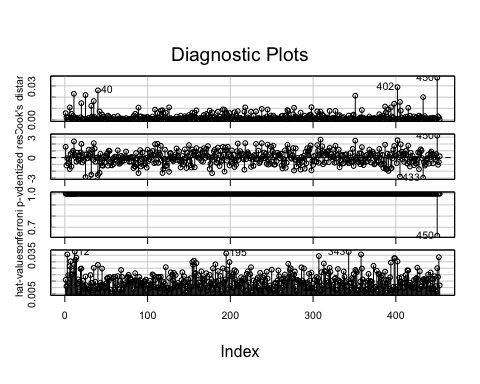
## 433 25 450   
## 1 2 453

# Check for outliers  
outlierTest(LinearModel.1)

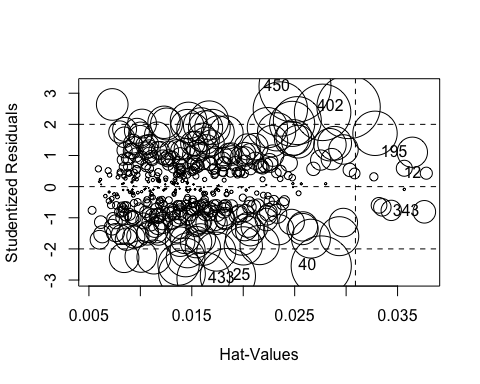
##   
## No Studentized residuals with Bonferonni p < 0.05  
## Largest |rstudent|:  
## rstudent unadjusted p-value Bonferonni p  
## 450 3.218868 0.0013811 0.62564

No studentized residuals with Bonferonni p < 0.05

# Identify influential points  
influenceIndexPlot(LinearModel.1, id.n=3)



# Create an influence plot  
influencePlot(LinearModel.1, id.n=3)



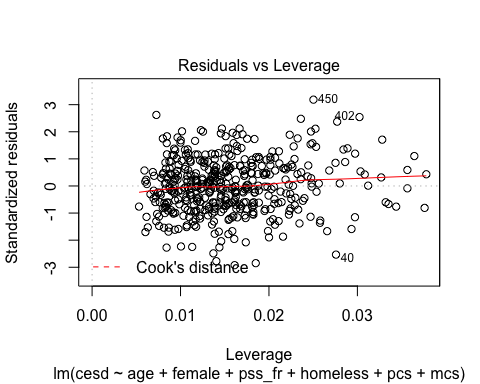
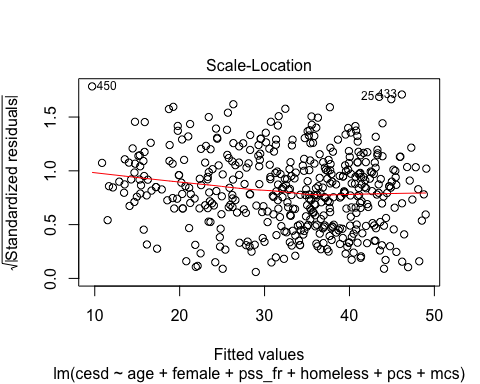
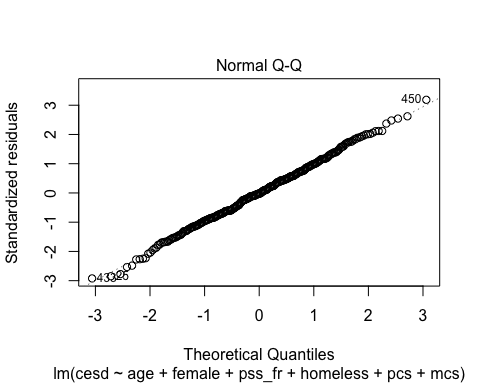
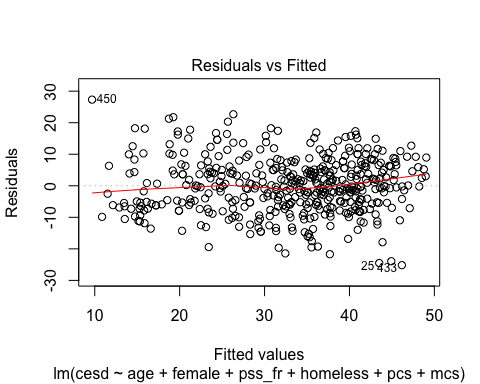
## StudRes Hat CookD  
## 12 0.4313265 0.03779399 0.001045833  
## 25 -2.8712570 0.01850847 0.021854090  
## 40 -2.5490027 0.02758106 0.026006371  
## 195 1.1039353 0.03643745 0.006580280  
## 343 -0.8084322 0.03760068 0.003650624  
## 402 2.5591353 0.03023968 0.028815823  
## 433 -2.9474775 0.01612078 0.019990575  
## 450 3.2188680 0.02502996 0.037218269

# Test for heteroskedasticity  
ncvTest(LinearModel.1)

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

Since p-value < .05, we reject the null hypothesis, i.e. there IS evidence of non-constant variance of the residuals. The residual plots (above), however, looked pretty good. I will try a different visualization.

plot(LinearModel.1)



There is a very slight curvature to the line, but I think overall this looks pretty good, so I will proceed without rebuilding the model.

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)].

GLM.1 <- glm(cesd\_gte16 ~ mcs, family=binomial(logit), data=h1)  
  
summary(GLM.1)

##   
## Call:  
## glm(formula = cesd\_gte16 ~ mcs, family = binomial(logit), 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

Since the raw coefficient of -.1716 has a p-value of less than 0, we can determine that menatl health quality of life (mcs) significantly contributes to the ability of the model to predict CESD scores > 16. To calculate the odds ratio, we convert the raw beta with the following code.

coef(GLM.1)

## (Intercept) mcs   
## 9.2691224 -0.1715576

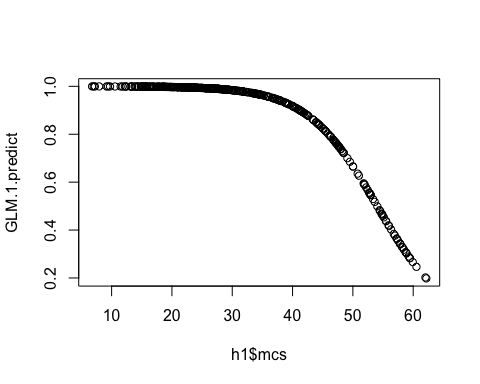
exp(coef(GLM.1))

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

The OR is < 1 (i.e. 0.84). This indicates that the higher a person’s mental health quality of life (mcs), the less likely they will be to have a cesd scores of > 16. For every point increase in mcs, the odds of them having a cesd score > 16 decreases by .84. Or the OR can be inverted and interpreted as: for every point decrease in mcs, the odds of having cesd > 16 (or potential for clinical depression) is increased by 1.19, or 19%.

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).

# Preticted probabilities  
GLM.1.predict <- predict(GLM.1, newdata=h1, type="response")  
  
# Plot of continuous predictor for these probabilities  
plot(h1$mcs, GLM.1.predict)



# Confusion matrix   
t1 <- table(GLM.1.predict > 0.5, h1$cesd\_gte16)  
t1

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

# Sensitivity  
tpr <- t1[2,2]/(t1[2,2]+t1[1,2])  
tpr

## [1] 0.970516

#Specificity  
tnr <- t1[1,1]/(t1[1,1]+t1[2,1])  
tnr

## [1] 0.4782609

Using a cutoff probability of 0.5, the confusion matrix indicates that: a) There were 395 true positives. That is, there are 395 “correct” predictions where the model correctly predicted a cesd score >/= 16. b) There were 22 true negatives. That is, there are 22 preditions that the individual scored < 16, when this was actually the case. c) There were 24 false positives. That is, there are 24 cases that were incorrectly predicted a cesd score >/=16, when they actually were < 16. d) There were 12 false negatives, or 12 cases that were incorrectly predicted to have a cesd score < 16, when they were actually > 16.

The sensitivity (true positive) rate is 0.97 and the specificity (true negative) rate is 0.48. The sensitivity of this model is excellent, i.e. its ability to predict a cesd >/= 16, however, the specificity is only fair, meaning that less than half of the time the model is able to correctly identify people who had a cesd </= 16.

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

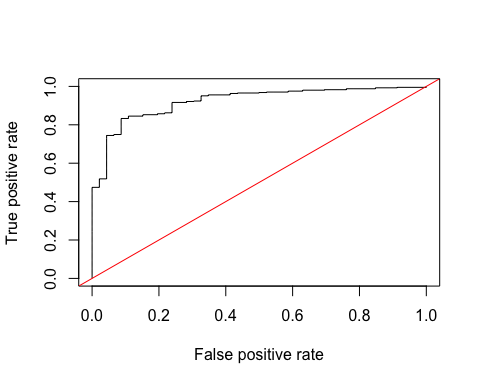
library(ROCR)

## Loading required package: gplots

##   
## Attaching package: 'gplots'

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

p <- predict(GLM.1, 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

The ROC curve is very close to the upper left hand corner, confirming, what we already saw in previous analyses, i.e. that it is a very good model for prediting depression (cesd >/= 16). We see this also with the auc value which is > .9 indicating excellent predictive ability.

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(effects)

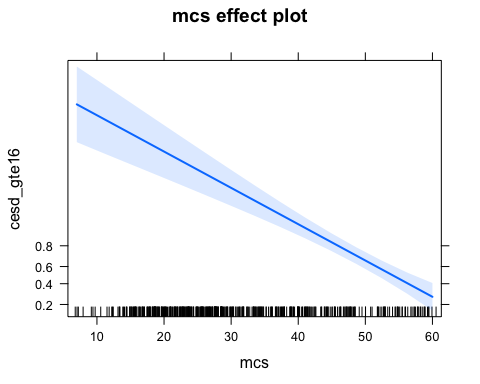
## Loading required package: carData

##   
## Attaching package: 'carData'

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

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

plot(allEffects(GLM.1))



This plot confirms that people with high mental health quality of life are less likely to have cesd scores indicative of depression. MCS does appear to be a good predictor of depression, however, the predictive ability of the tool appears to drop off at higher levels, i.e. the model doesn’t work as well for people with excellent mental health. This, however, makes sense since this is the only predictor in the model!