

NRSG 741 Homework 6

Tommy FLynn

April 8th, 2018

GitHub Repository: https://github.com/tommyflynn/N741_Homework/tree/master/Flynn_HW_06

For homework 6, we use the **HELP** (Health Evaluation and Linkage to Primary Care) Dataset.

Table 1: Variable Labels for Homework 6, and Table 2: First 6 Observations

Only on the following variables from the HELP dataset are used for this assignment:

Table 1: Use these variables from HELP dataset for Homework 06

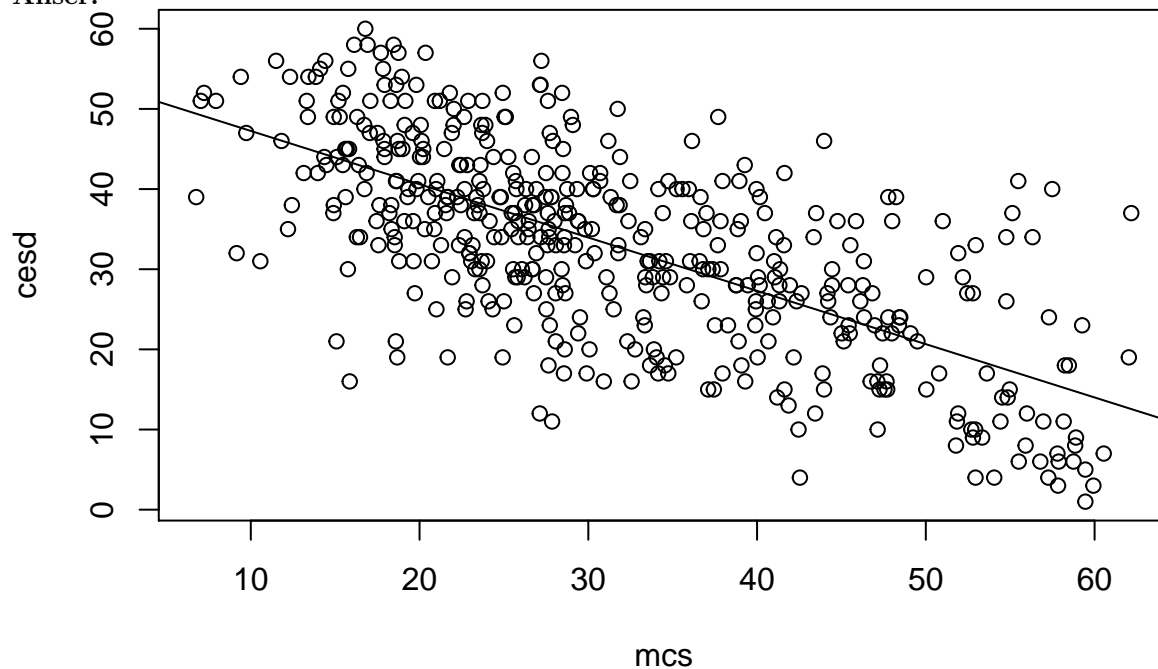
	Variable Label
age	Age at baseline (in years)
female	Gender of respondent
pss_fr	Perceived Social Support - friends
homeless	One or more nights on the street or shelter in past 6 months
pcs	SF36 Physical Composite Score - Baseline
mcs	SF36 Mental Composite Score - Baseline
cesd	CESD total score - Baseline

Table 2: First six rows of the new HELP subset

age	female	pss_fr	homeless	pcs	mcs	cesd	cesd_gte16
37	0	0	0	58.41369	25.111990	49	1
37	0	1	1	36.03694	26.670307	30	1
26	0	13	0	74.80633	6.762923	39	1
39	1	11	0	61.93168	43.967880	15	0
32	0	10	1	37.34558	21.675755	39	1
47	1	5	0	46.47521	55.508991	6	0

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.

Answer:



```
##
## 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
```

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).

Answer: For each unit increase in *mcs*, the *cesd* score decreases by 0.665 units.

$$cesd = 53.902 - (0.665)mcs$$

3. How much variability in the *cesd* does the *mcs* explain? (what is the R^2 ?) Write a sentence describing how well the *mcs* does in predicting the *cesd*.

Answer: 47% of the variability in *cesd* is explained by *mcs* ($R^2 = 0.47$).

4. [Model 2] Run a second linear regression model (`lm()`) for the *cesd* putting in all of the other variables:

```
#Use lm() to regress all variables on cesd
model1 <- lm(cesd ~ age + female + pss_fr + homeless + pcs + mcs, data=h1)
#Print out the model results with the coefficients and tests and model fit statistics.
#summary(model1)
model2 <- lm(cesd ~ female + pss_fr + pcs + mcs, data=h1)
#summary(model2)
stargazer(model1, model2, title="Comparison of 2 Regression Outputs",
           type = "text", align=TRUE)
```

```
##
## Comparison of 2 Regression Outputs
## =====
##                               Dependent variable:
##                               -----
##                               cesd
##                               (1)          (2)
## -----
## age                          -0.013
##                               (0.055)
##
## female                       2.350**
##                               (0.988)
##                               2.289**
##                               (0.980)
##
## pss_fr                      -0.256**
##                               (0.106)
##                               -0.267**
##                               (0.104)
##
## homeless                     0.465
##                               (0.843)
##
## pcs                         -0.236***
##                               (0.040)
##                               -0.236***
##                               (0.039)
##
## mcs                        -0.621***
##                               (0.033)
##                               -0.622***
##                               (0.032)
##
## Constant                    65.300***
##                               (3.187)
##                               65.154***
##                               (2.154)
## -----
## Observations                 453
## R2                           0.525
## Adjusted R2                  0.519
## Residual Std. Error    8.683 (df = 446)
## F Statistic            82.135*** (df = 6; 446)
##                               8.667 (df = 448)
##                               123.574*** (df = 4; 448)
## =====
```

Note:

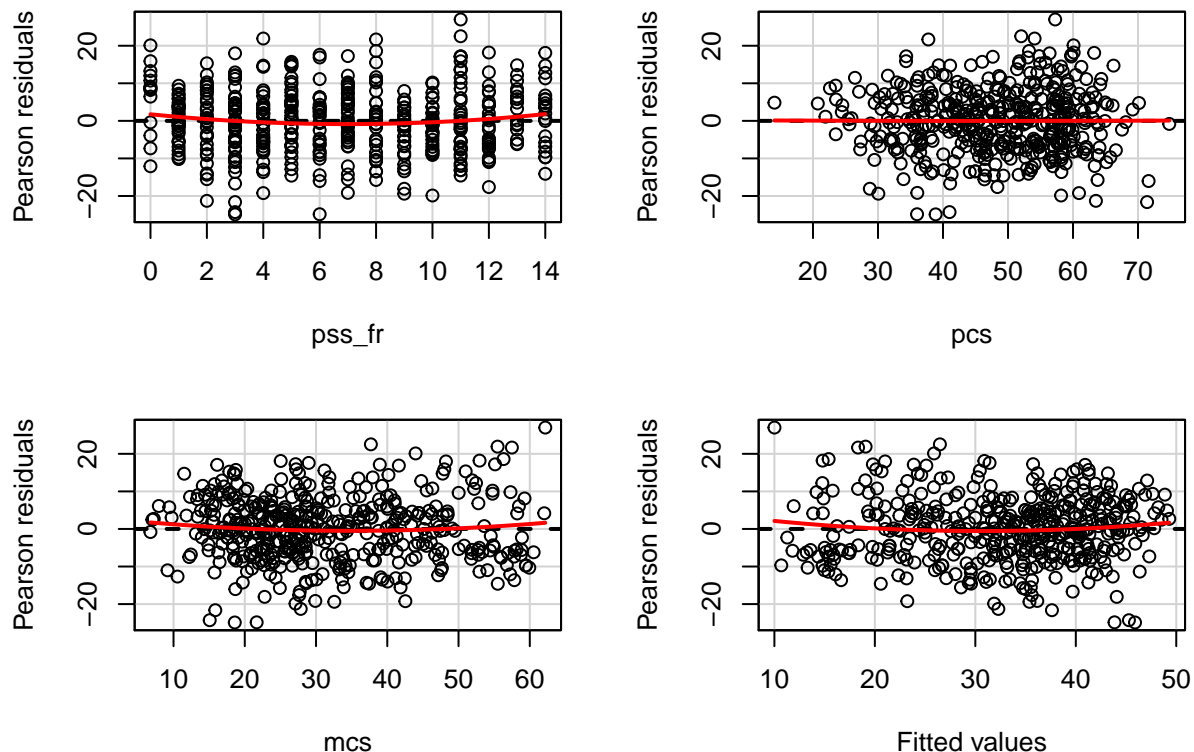
*p<0.1; **p<0.05; ***p<0.01

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).

Answer: Female, pss_fr, pcs and mcs are all significantly associated with cesd. Based on the model with only significant predictors, on average women score higher on the cesd by 2.29 points, every unit increase on the physical composite score decreases the cesd score by 0.24, a unit increase on the mental composite score decreases cesd by 0.62 units, and 1 unit increase on the social support scale decreases cesd by 0.27 units. Overall, this model accounts for 52% of the variability in cesd ($R^2 = 0.52, p = < 0.001$).

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

```
residualPlots(model2)
```



```
##          Test stat Pr(>|t|)
## pss_fr      1.958   0.051
## pcs         0.038   0.970
## mcs         1.232   0.219
## Tukey test   1.392   0.164
```

```
#lot(resid ~ age + female + pss_fr + homeless + pcs + mcs, data=h1)
```

7. [Model 3] Repeat Model 1 above, except this time run a logistic regression (`glm()`) to predict CESD scores \Rightarrow 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)].

8. Use the `predict()` function like we did in class to predict CESD \Rightarrow 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 \Rightarrow 16 (indicating depression)? (make the "confusion matrix")

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

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" if you're using `Rcmdr` to do these analyses.]
