Flynn_HW_06

TommY

4/7/2018

Abstract

Homework 6 was due on 04/06/2018, the same day as my F-31 NRSA application. I wasn't able to complete both. My apologies for being late on this assignment. I wanted to get it done anyway, so I don't get behind. Thanks.

Find the associated GitHub Repository Here: 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.

Variables for Homework 6

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

Homework 6 Assignment

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 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.

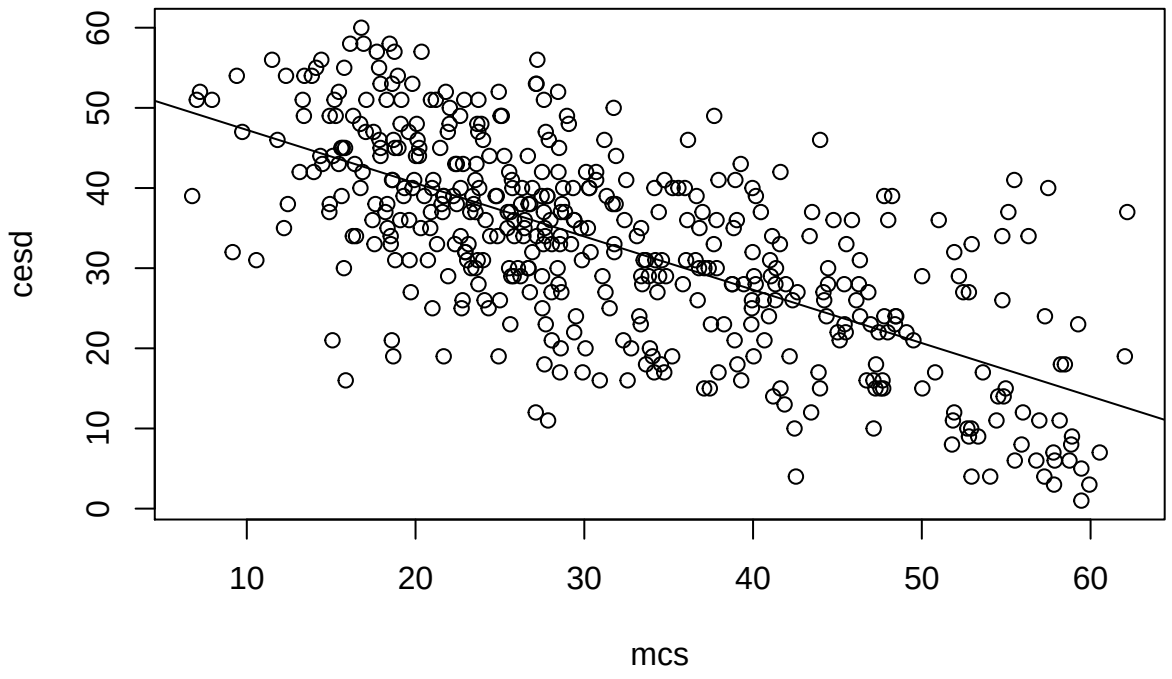
The recommended threshold used to indicate potential clinical depression is for people with scores of 16 or greater. You will then use the variable created using this cutoff (`cesd_gte16`) to perform a similar modeling approach with the variables to predict the probability of clinical depression (using logistic regression).

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.

```
reg1 <- lm(cesd ~ mcs, data = h1)
summary(reg1)
```

```
##
## 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
```

```
plot(cesd ~ mcs, data=h1)
abline(a=53.902, b=-0.665)
```



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).

$$cesd = 53.902 - (0.665)mcs \tag{1}$$

For each unit increase in `mcs`, the `cesd` score decreases by 0.665 units.

3. 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`.

46% of the variability in `cesd` is due to `mcs` (R2=0.47).

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

```
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)
```

```
##
## 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
```

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).

`Female`, `pss_fr`, `pcs` and `mcs` are all significantly associated with `cesd`. On average, women score higher on the `cesd` by 2.34 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 unites, and 1 unit increase on the social support scale decreased `cesd` by 0.26 units.

6. 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.

7. [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)].

8. 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 "confusi
```

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