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

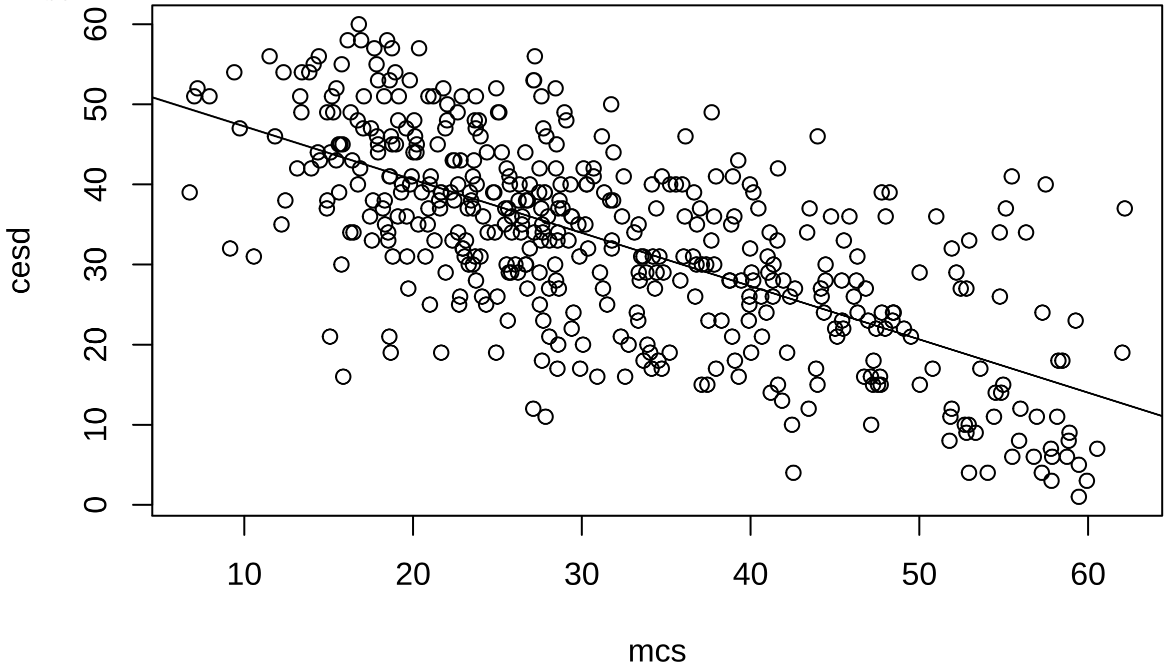
	Variable Label
age	Age at baseline (in years)
female	Gender of respondent
pss_fr	Perceived Social Support - friends
homeless	One or more nights on the street or shelter in past 6 months
pcs	SF36 Physical Composite Score - Baseline
mcs	SF36 Mental Composite Score - Baseline
cesd	CESD total score - Baseline

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.

Anser:




```
##
## 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).

Anser: *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:

```r
#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**                 2.289**
##                               (0.988)                 (0.980)
##
## pss_fr                       -0.256**                -0.267**
##                               (0.106)                 (0.104)
##
## homeless                       0.465
##                               (0.843)
##
## pcs                          -0.236***               -0.236***
##                               (0.040)                 (0.039)
##
## mcs                          -0.621***               -0.622***
##                               (0.033)                 (0.032)
##
## Constant                     65.300***               65.154***
##                               (3.187)                 (2.154)
##
## --------------------------------------------------------------------
## Observations                    453                     453
## R2                             0.525                   0.525
## Adjusted R2                    0.519                   0.520
## Residual Std. Error      8.683 (df = 446)        8.667 (df = 448)
## F Statistic         82.135*** (df = 6; 446) 123.574*** (df = 4; 448)
## ====================================================================
```

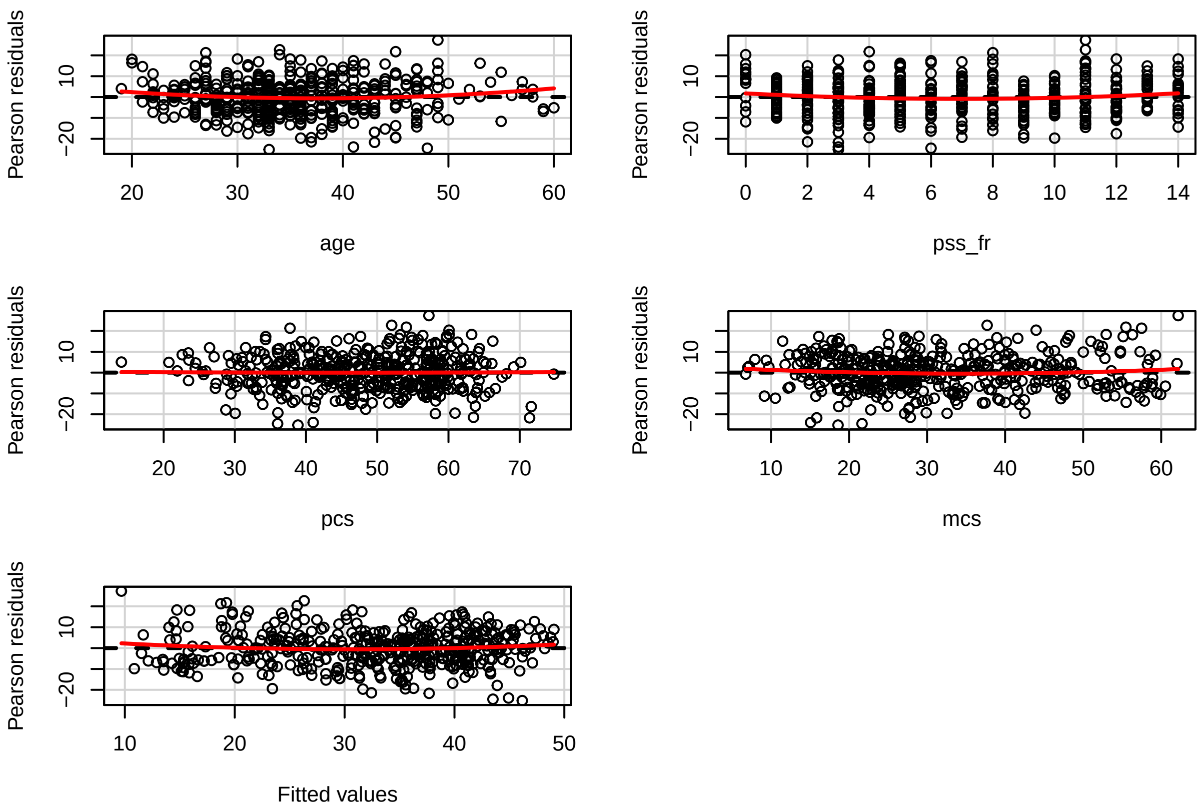
```
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

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 unites, and 1 unit increase on the social support scale decreases* `cesd` *by 0.27 units. Overall, this model accounts fo 52% of the variability in* `cesd` ($R^2 = 0.52, p =< 0.001$).

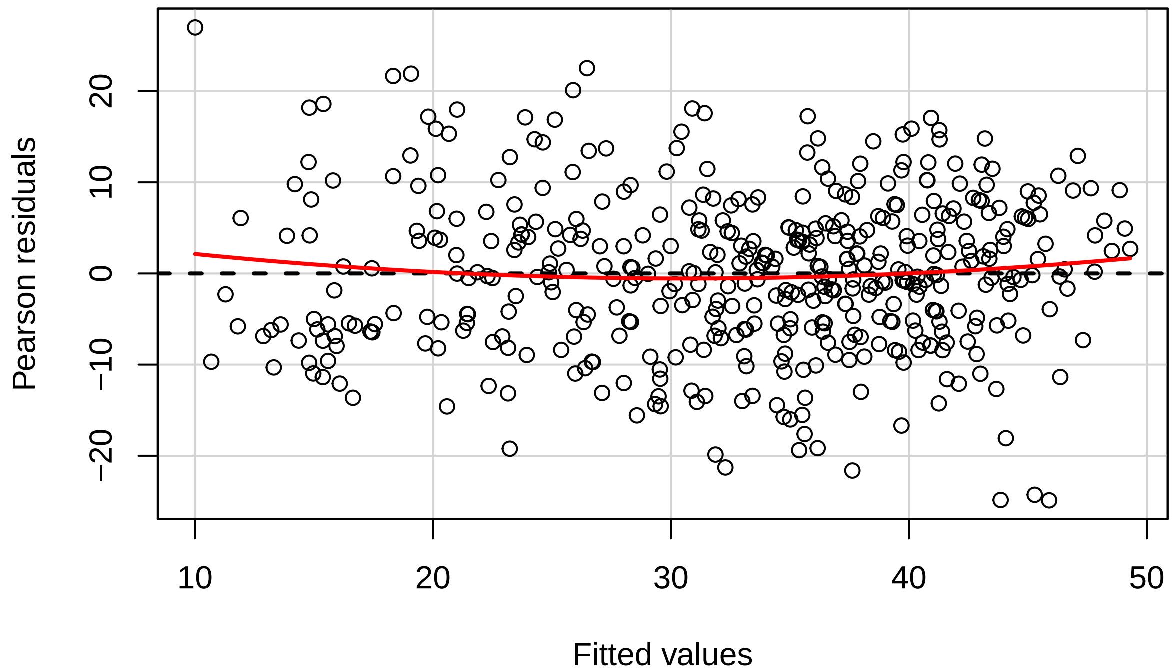
6. Generate the diagnostic plotss for this model with these 6 predictors (e.g. get the residual plot by variables, the added-variable plots, the Q-Q plot, diagnostic plots). Also run the VIFs to check for multicollinearity issues.

```r
#residual plot on models 1 & 2
residualPlots(model1)
```

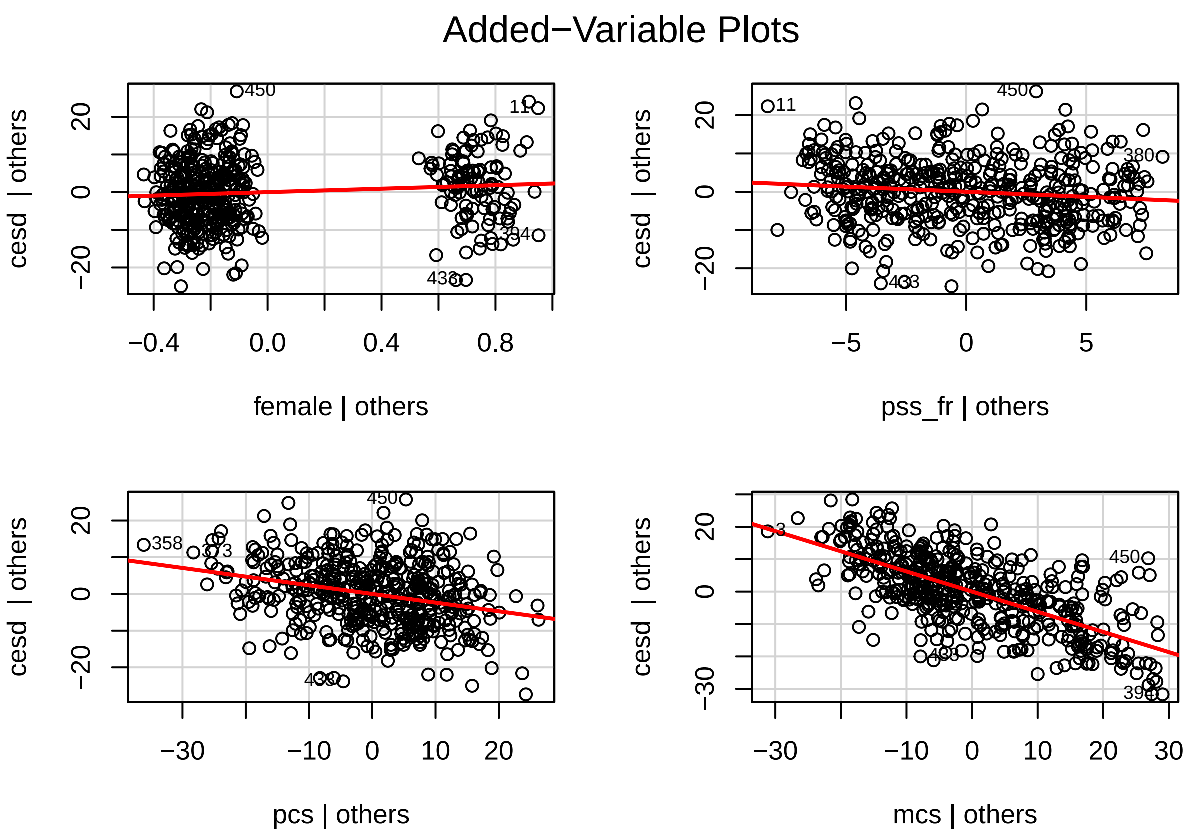



```
##              Test stat Pr(>|t|)
## age              1.941    0.053
## pss_fr           1.964    0.050
## pcs              0.081    0.936
## mcs              1.260    0.208
## Tukey test       1.434    0.152
```

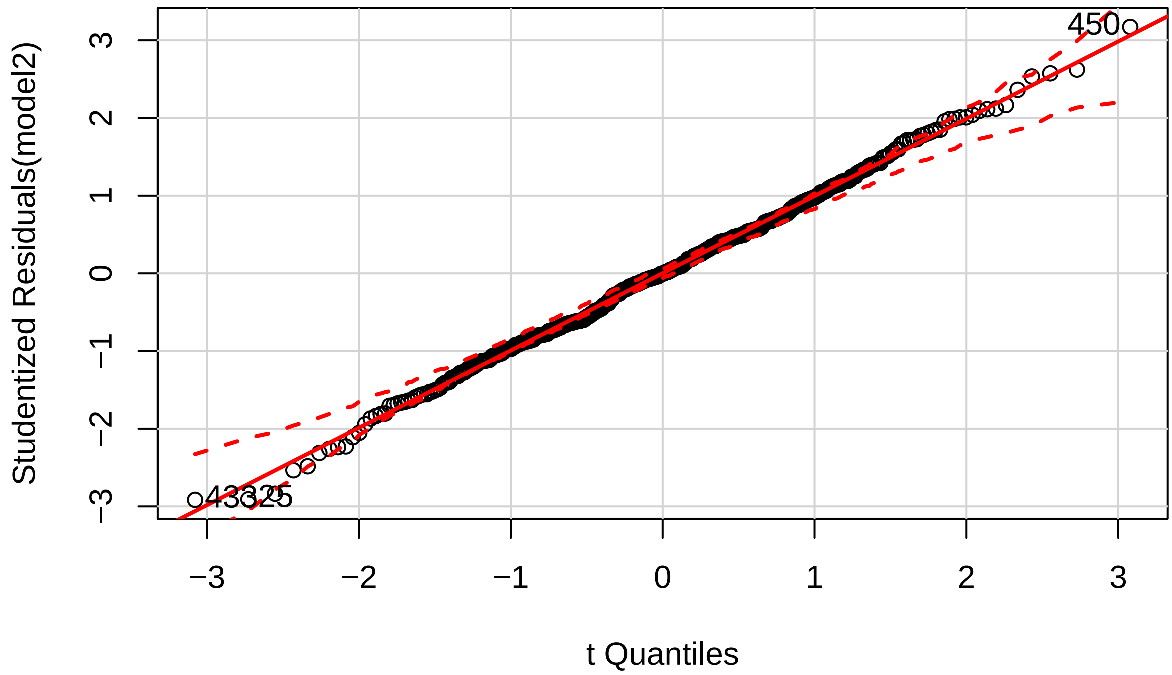
```r
residualPlot(model2)
```




```r
#Added Variable plots for model 2
avPlots(model2, id.n=2, id.cex=0.7)
```




```r
#Q-Q plot for model 2
qqPlot(model2, id.n=3)
```

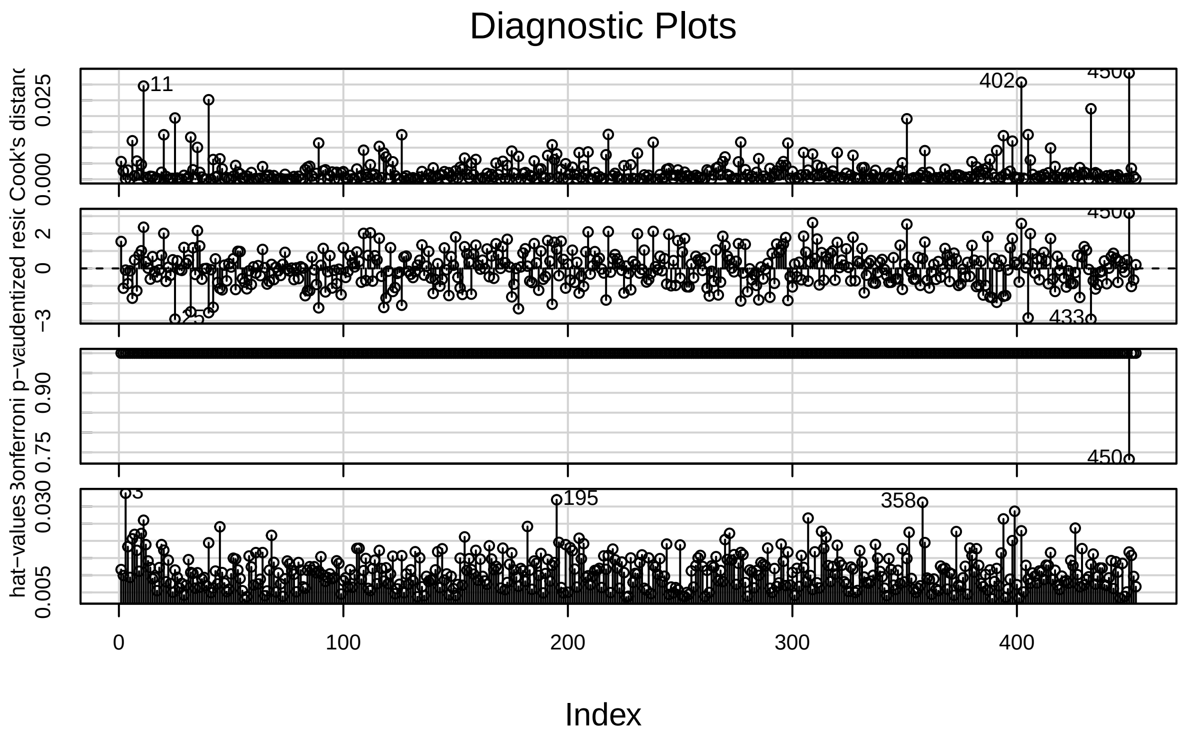



```
## 433  25 450
##   1   2 453
```

```
#Any Outliers?
outlierTest(model2)
```

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferonni p
## 450 3.172271          0.0016167      0.73238
```

```
#Highly influential observations? Diagnostic plots:
influenceIndexPlot(model2, id.n=3)
```




```r
#Now use VIFs to check for multicolinearity (GVIF > 4 = colinearity)
vif(model2)
```

```
##   female   pss_fr      pcs      mcs
## 1.045607 1.032659 1.040147 1.043754
```

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)].

```r
logit1 <- glm(cesd_gte16 ~ mcs, data=h1, family=binomial)
summary(logit1)
```

```
##
## Call:
## glm(formula = cesd_gte16 ~ mcs, family = binomial, 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
```

```r
exp(coef(logit1))
```

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

Answer: $cesd.gte16 = 9.27 - 0.17(mcs)$ $(OR : 0.84, p = 0)$

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.

```
+ How well did the model correctly predict CESD scores => 16 (indicating depression)? (make the "confusi
```

```r
logit1.predict <- predict(logit1, newdata=h1,
                          type="response")

# plot the continuous predictor
# for these predicted probabilities
#plot(h1$mcs, logit1.predict)
#table(h1$cesd_gte16, logit1.predict > 0.5)
#t1 <- table(logit1.predict > 0.5, h1$cesd_gte16)
#t1
library(gmodels)
CrossTable(h1$cesd_gte16, logit1.predict > 0.5)
```

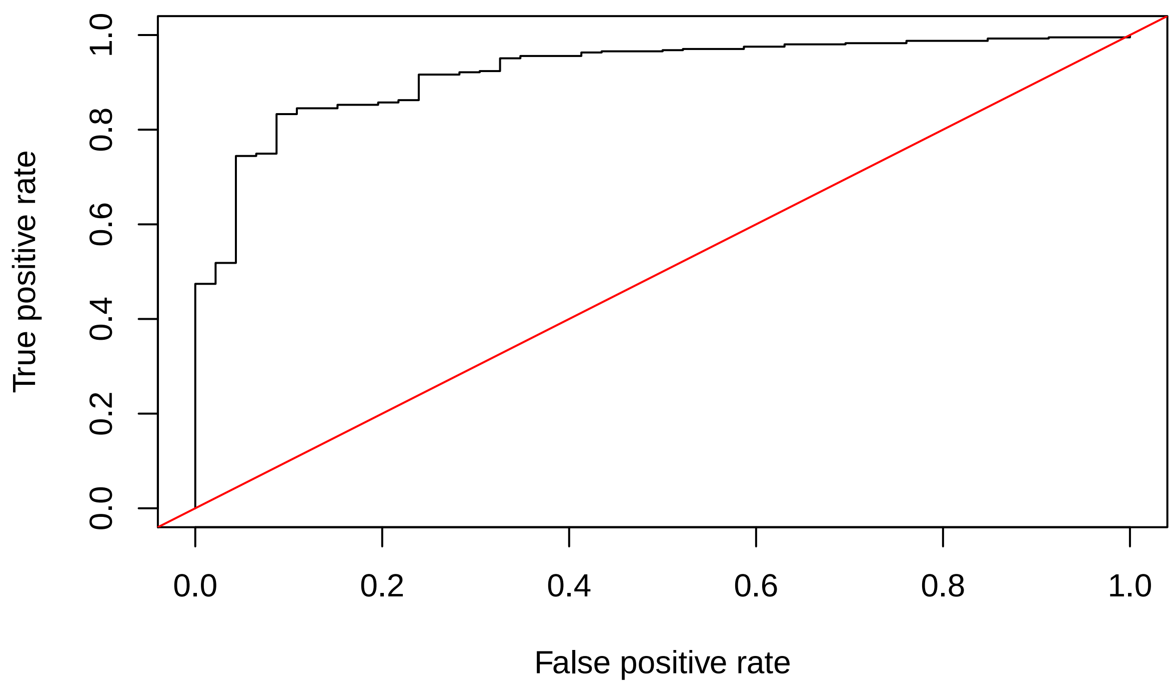
```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  453
##
##
##               | logit1.predict > 0.5
## h1$cesd_gte16 |     FALSE |      TRUE | Row Total |
## --------------|-----------|-----------|-----------|
##             0 |        22 |        24 |        46 |
##               |    99.639 |     8.085 |           |
##               |     0.478 |     0.522 |     0.102 |
##               |     0.647 |     0.057 |           |
##               |     0.049 |     0.053 |           |
```

```
## --------------|-----------|-----------|-----------|
##              1 |        12 |       395 |       407 |
##                |    11.261 |     0.914 |           |
##                |     0.029 |     0.971 |     0.898 |
##                |     0.353 |     0.943 |           |
##                |     0.026 |     0.872 |           |
## --------------|-----------|-----------|-----------|
##   Column Total |        34 |       419 |       453 |
##                |     0.075 |     0.925 |           |
## --------------|-----------|-----------|-----------|
##
##
```

Answer: *The model actually did very well, it correctly predicted 22 cesd scores <16 and 395 scores >= 16. It incorrectly predicted 12 true as false, and 24 true as negative.*

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

```
library(ROCR)
p <- predict(logit1, 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
```

Answer: *The area under the curve us 0.922, which is great!*

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.]