

Homework 6

Jasmine Nakayama

April 6, 2018

Link to repository: (<https://github.com/jynakay/Assignments> (<https://github.com/jynakay/Assignments>))
[<https://github.com/jynakay/Assignments> (<https://github.com/jynakay/Assignments>)]

```
# Load libraries and dataset
library(tidyverse)
library(haven)
library(car)
library(ROCR)
```

```
helpdata <- haven::read_spss("helpmkh.sav")

# choose variable
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
```

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.

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

```
##
## Call:
## lm(formula = cesd ~ mcs, data = h1)
##
## Coefficients:
## (Intercept)          mcs
##      53.9022      -0.6647
```

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).

$$\text{cesd} = 53.9022 - 0.6647 * \text{mcs}$$

For every 1 point increase in MCS score, the CEDS score decreases by 0.6647. Generally, better mental health is associated with lower depression score. Those with an MCS score of 0 will have a CESD of 53.9022.

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`.

The adjusted R^2 is 0.4638, which indicates that `cesd` accounts for 46.38% of the variability in `mcs`, which is fairly good for a simple linear regression model.

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

```
mlr<-lm(cesd~age +female +pss_fr +homeless +pcs +mcs, data=h1)
summary(mlr)
```

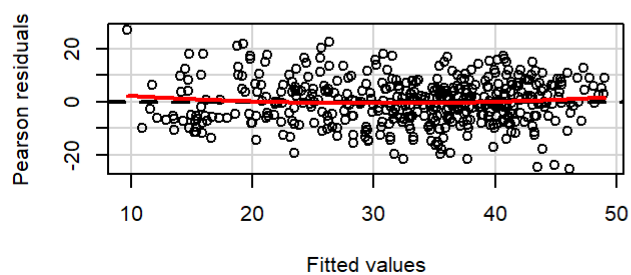
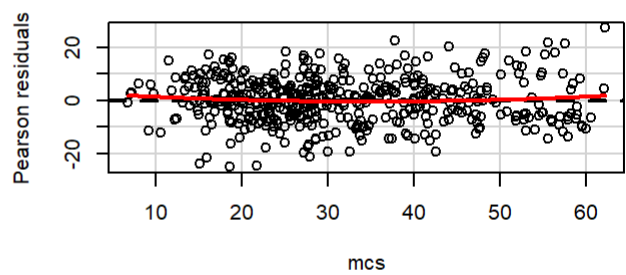
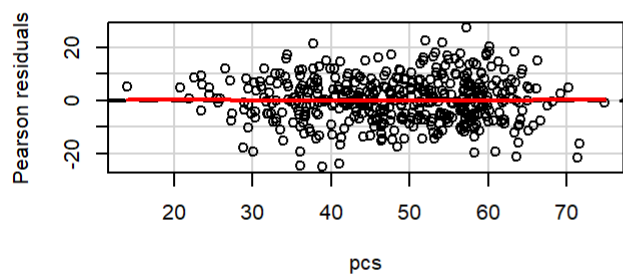
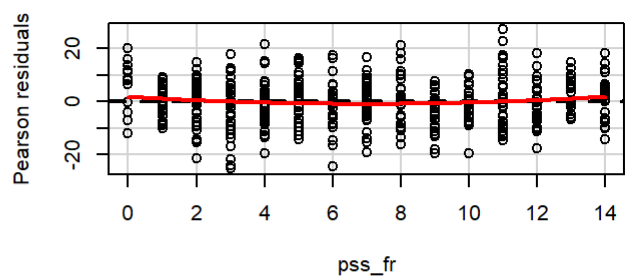
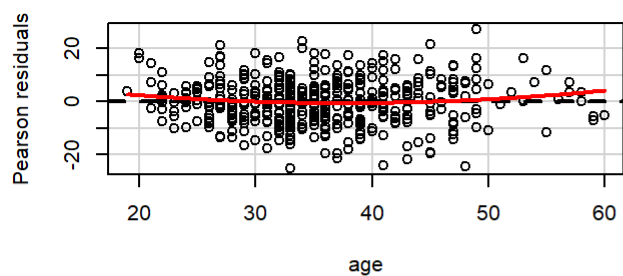
```
##
## Call:
## lm(formula = cesd ~ age + female + pss_fr + homeless + pcs +
##     mcs, data = h1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.1711  -5.9894  -0.2077   5.5706  27.3137
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  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).

When adjusting for the other variables in the model, the following variables are significant in the model: `female`, `pss_fr`, `pcs`, and `mcs`. When adjusting for the other variables, a 1 unit increase in `female` results in a 2.35028 increase in `cesd`, and a 1 unit increase in `pss_fr`, `pcs`, or `mcs` results in a 0.25569, 0.23639, or 0.62093 decrease in `cesd` respectively.

6. 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.

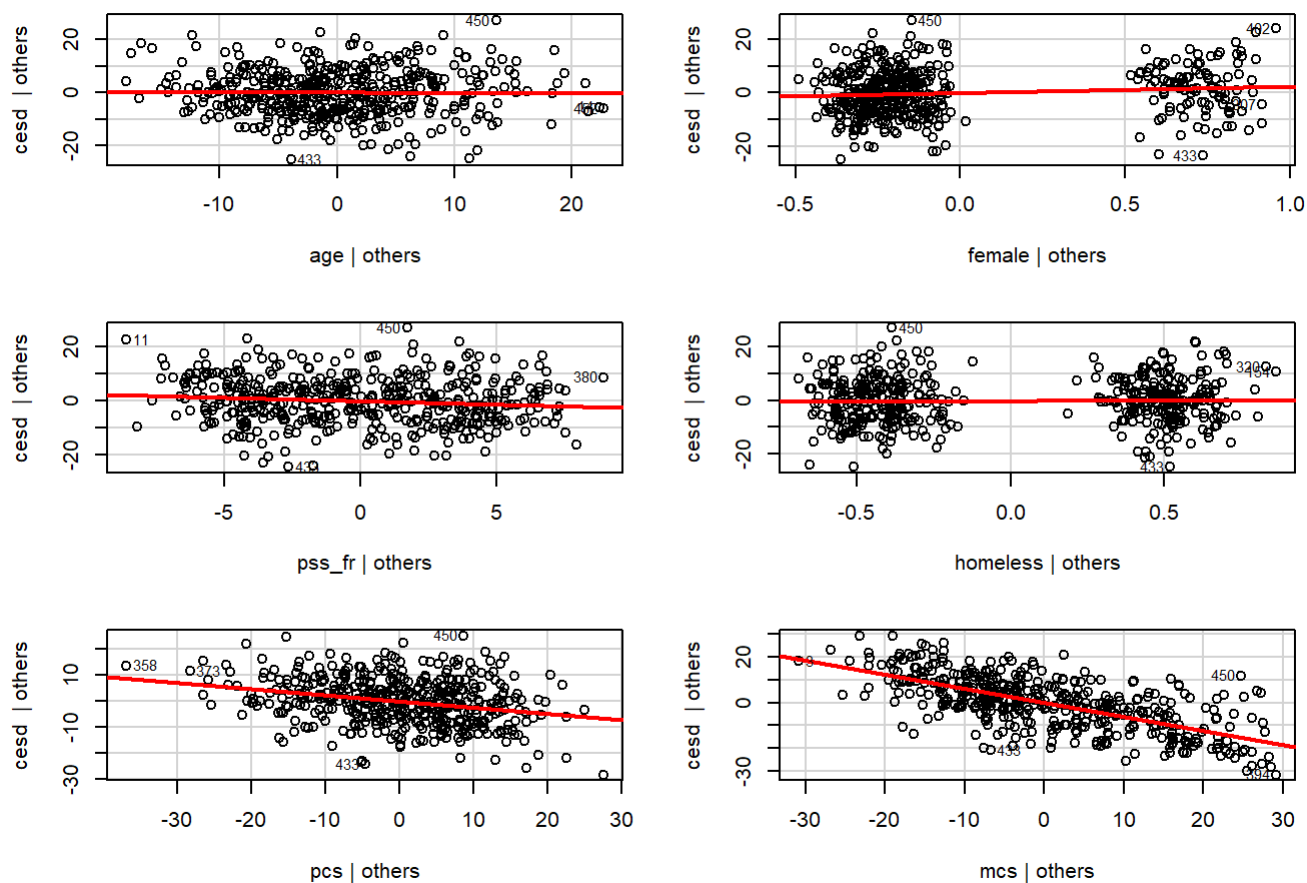
```
residualPlots(mlr)
```



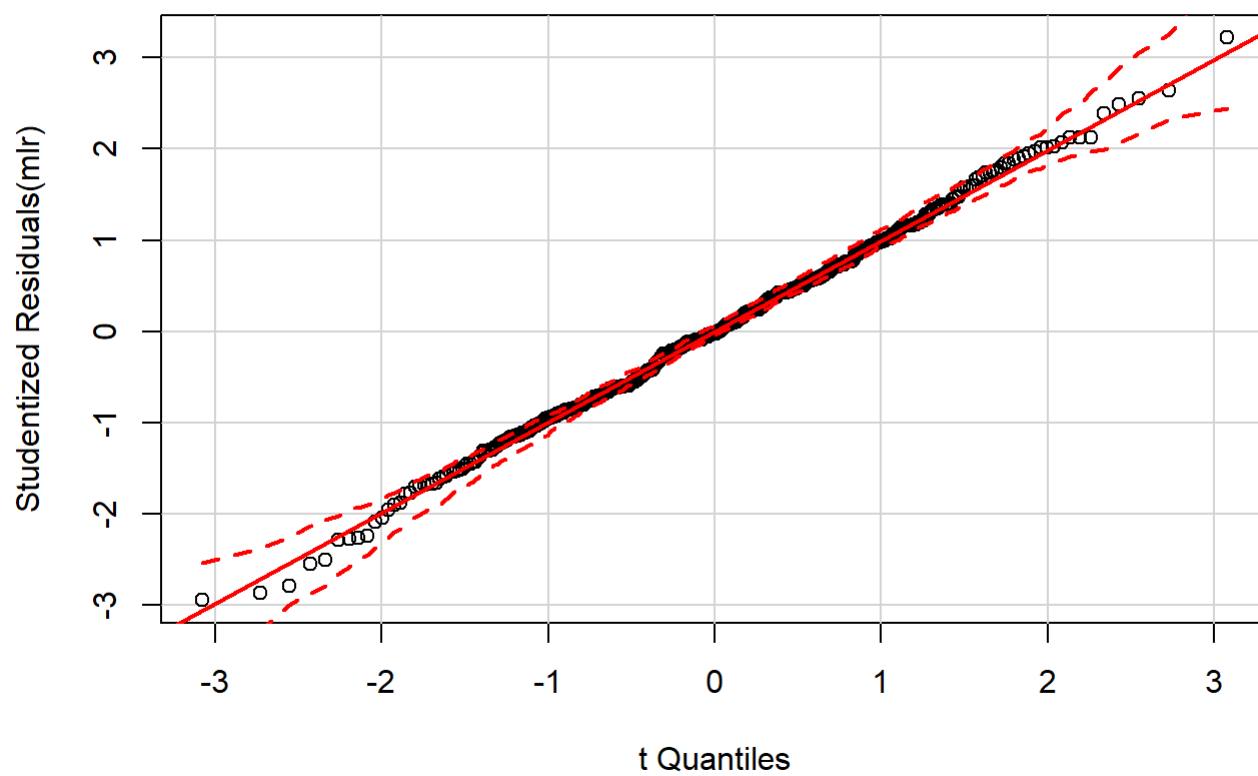
```
##          Test stat Pr(>|t|)
## age          1.941   0.053
## pss_fr        1.964   0.050
## pcs           0.081   0.936
## mcs           1.260   0.208
## Tukey test    1.434   0.152
```

```
avPlots(mlr, id.n=2, id.cex=0.7)
```

Added-Variable Plots



```
qqPlot(mlr)
```

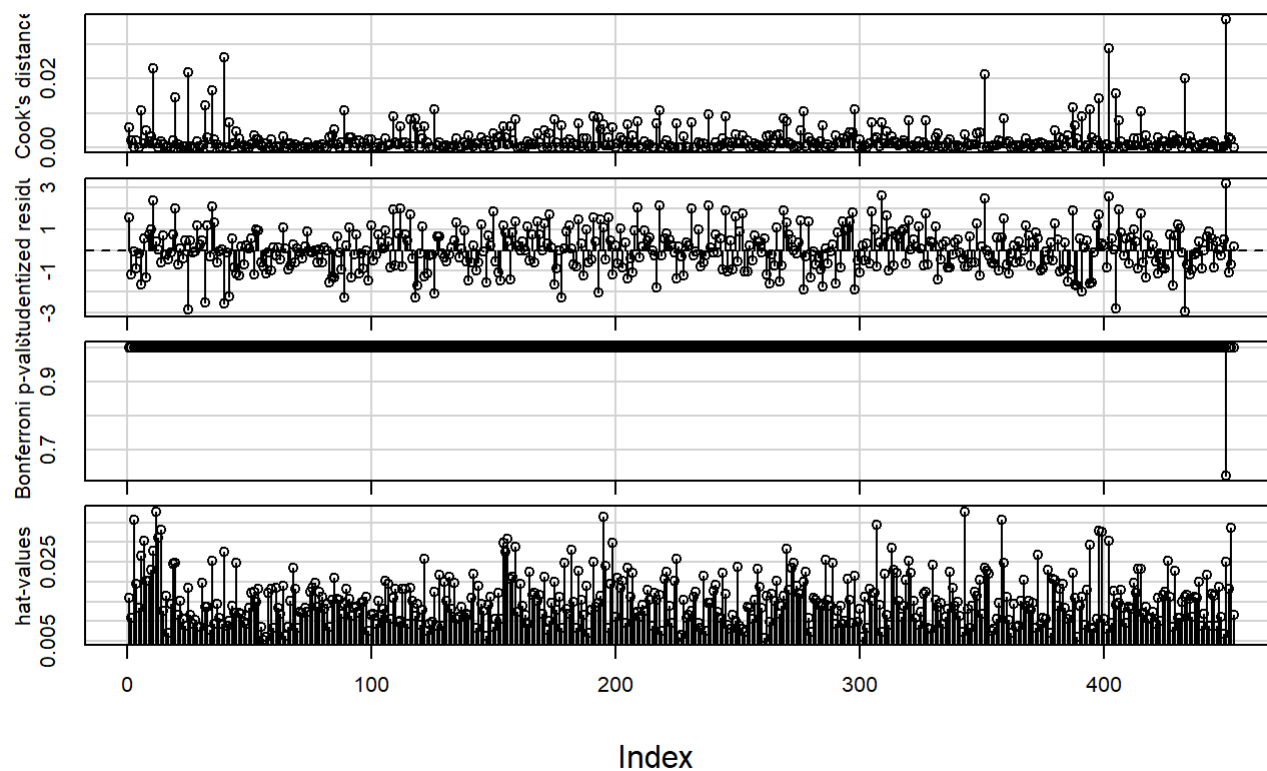


```
outlierTest(mlr)
```

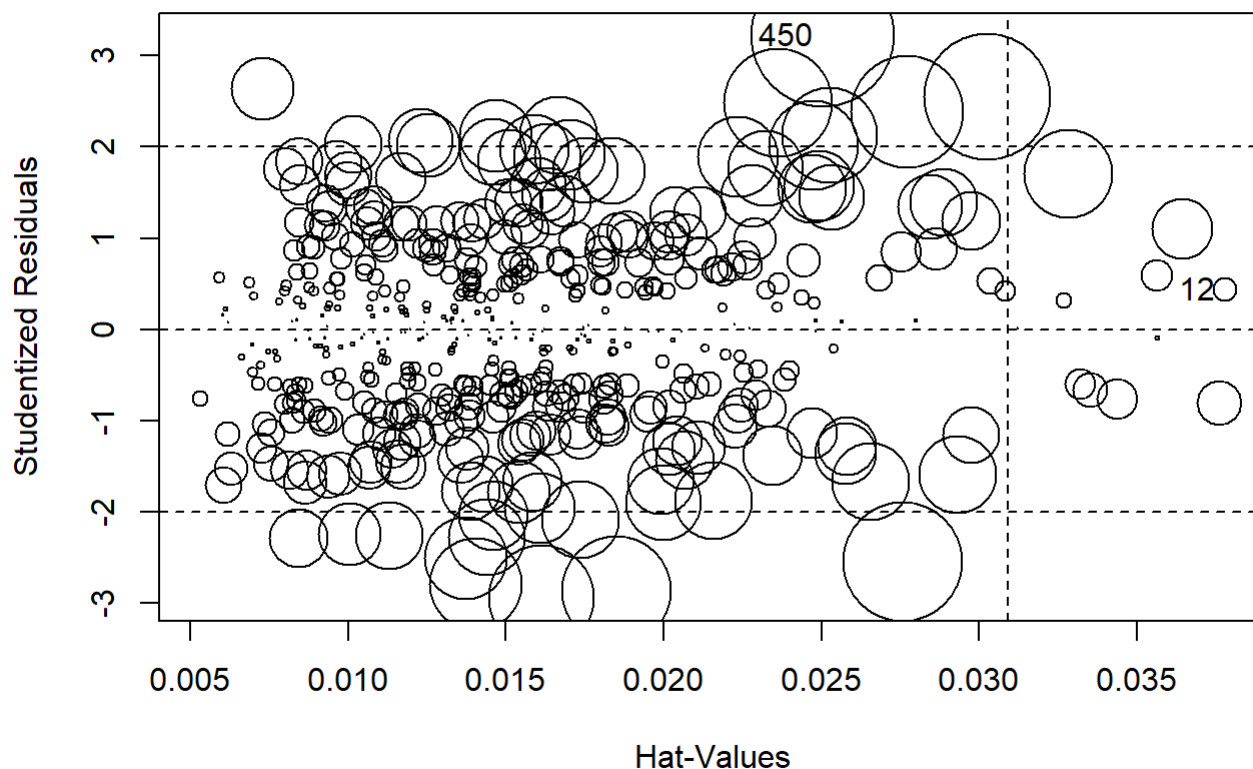
```
##  
## No Studentized residuals with Bonferonni p < 0.05  
## Largest |rstudent|:  
##      rstudent unadjusted p-value Bonferonni p  
## 450 3.218868      0.0013811      0.62564
```

```
influenceIndexPlot(mlr)
```

Diagnostic Plots



```
influencePlot(mlr)
```



```
##      StudRes      Hat      CookD
## 12  0.4313265 0.03779399 0.001045833
## 450 3.2188680 0.02502996 0.037218269
```

```
vif(mlr)
```

```
##      age  female  pss_fr homeless  pcs  mcs
## 1.078264 1.058232 1.068213 1.060007 1.108172 1.050768
```

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.

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



```
##
## 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
```

```
exp(coef(glm)) # Exponentiated coefficients ("odds ratios")
```

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

The odds of CESD score greater than or equal to 16 is 0.8423518 higher for each increase in MCS.

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.

```
+ How well did the model correctly predict CESD scores  $\Rightarrow$  16 (indicating depression)? (make the
"confusion matrix" and look at the true positives and true negatives versus the false positives
and false negatives).
```

```
glm.predict <- predict(glm, newdata=h1,
                      type="response")
table(h1$cesd_gte16, glm.predict > 0.5)
```

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

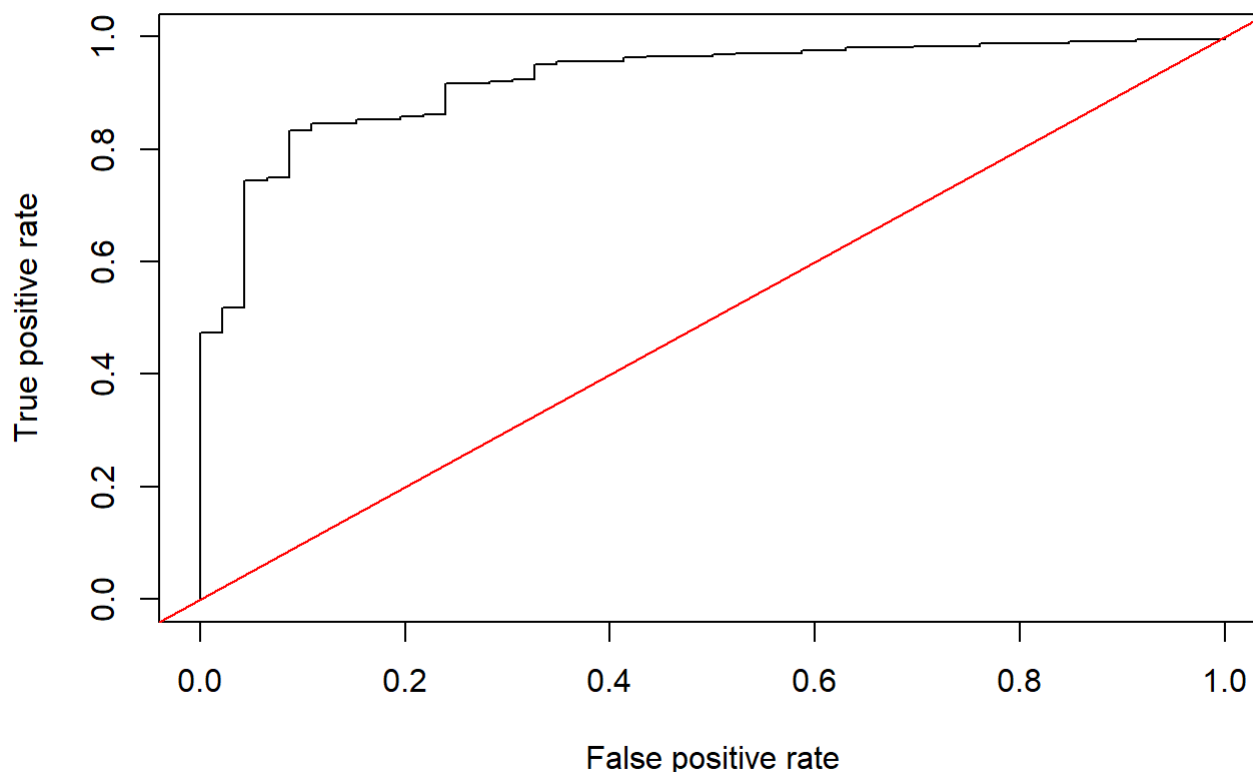
```
t1 <- table(glm.predict > 0.5, h1$cesd_gte16)
t1
```

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

The model was able to predict 395 of the 419 true cases CESD scores ≥ 16 .

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

```
p <- predict(glm, 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")
```



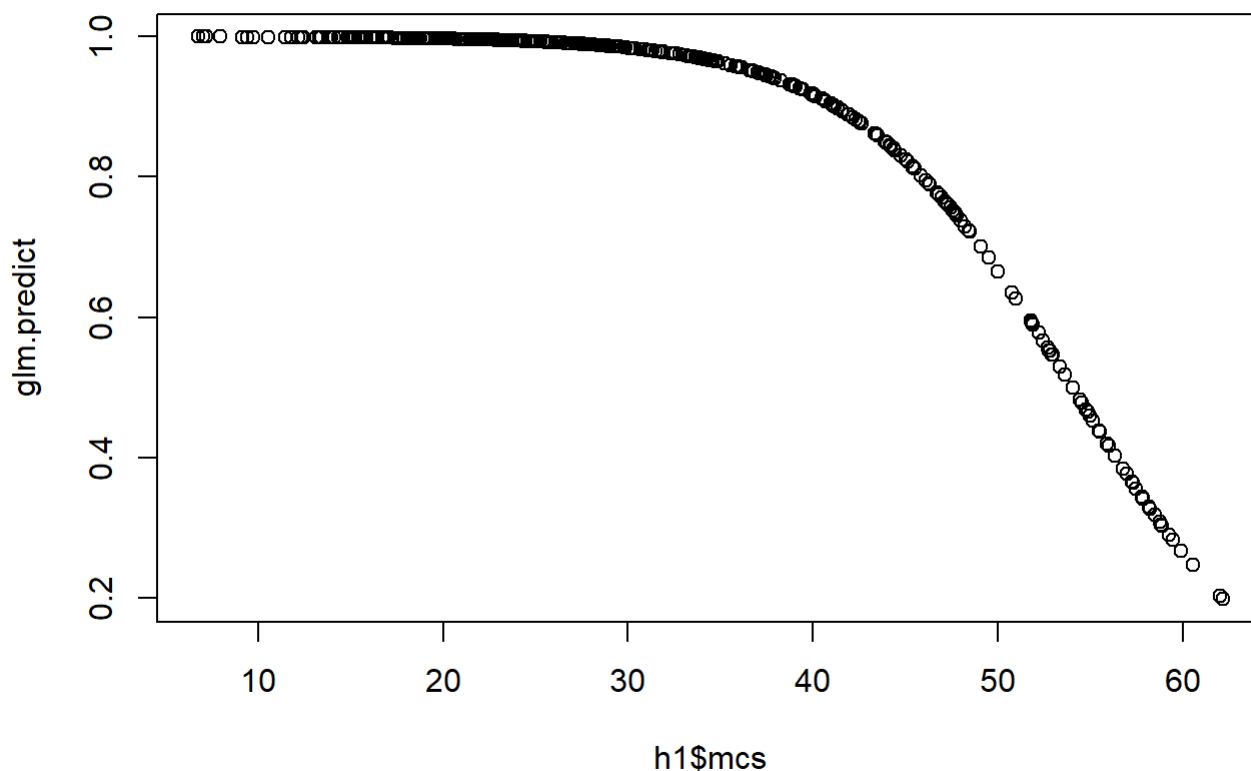
```
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

```
## [1] 0.9221771
```

AUC of 0.922 is great, so this is a good model to predict depression.

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?

```
plot(h1$mcs, glm.predict)
```



MCS less than 30 seems to indicate depression. With MCS scores greater than 30, higher scores indicate less depression.