N741 Spring 2018 - Homework 6

Homework 6

Rosemary Kinuthia

April 2, 2018

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

```
sub1 <- helpdata %>%
  select(age, female, pss_fr, homeless,
         pcs, mcs, cesd)
```

```
# create a function to get the label
# label output from the attributes() function
getlabel <- function(x) attributes(x)$label
# getlabel(sub1$age)
```

```
# load libraries and dataset
library(tidyverse)
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.

```
#linear regression of CESD using MCS
Model1 <- lm(cesd~mcs, data=h1)
summary(Model1)
```

```
##
## 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). *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.*

```
#Y= 53.90-0.66*mcs
#Interpretation: for each unit increase in mcs, cesd decreases by 0.66
```

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`.

```
#R-squared=0.465 indicating that the predictor(mcs) explains 46.5% variability in the outcome(cesd)
```

4. [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.

```
#linear regression for cesd and all other variables
Model2 <- lm(cesd~age + female + pss_fr + homeless + pcs + mcs, data=h1)
summary(Model2)
```

```
##
## 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 significant

#R-squared is 0.5249 indicating that 52.49% of the variabilty in average cesd is due to the independent variables in the model

#From this output we can see that:
#Being female increases the cesd by 2.35, holding all other variables constant
#If pss_fr increases by one unit, the cesd decreases by 0.26 units, holding all other variables constant.
#If pcs increases by one unit, then the cesd decreases by 0.24 units, holding all other variables constant.
##If mcs increases by one unit, then the cesd decreases by 0.62 units, holding all other variables constant.
```

6. Following the example we did in class for the Prestige dataset https://cdn.rawgit.com/vhertzb/2018week9/2f2ea142/2018week9.html?raw=true (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.

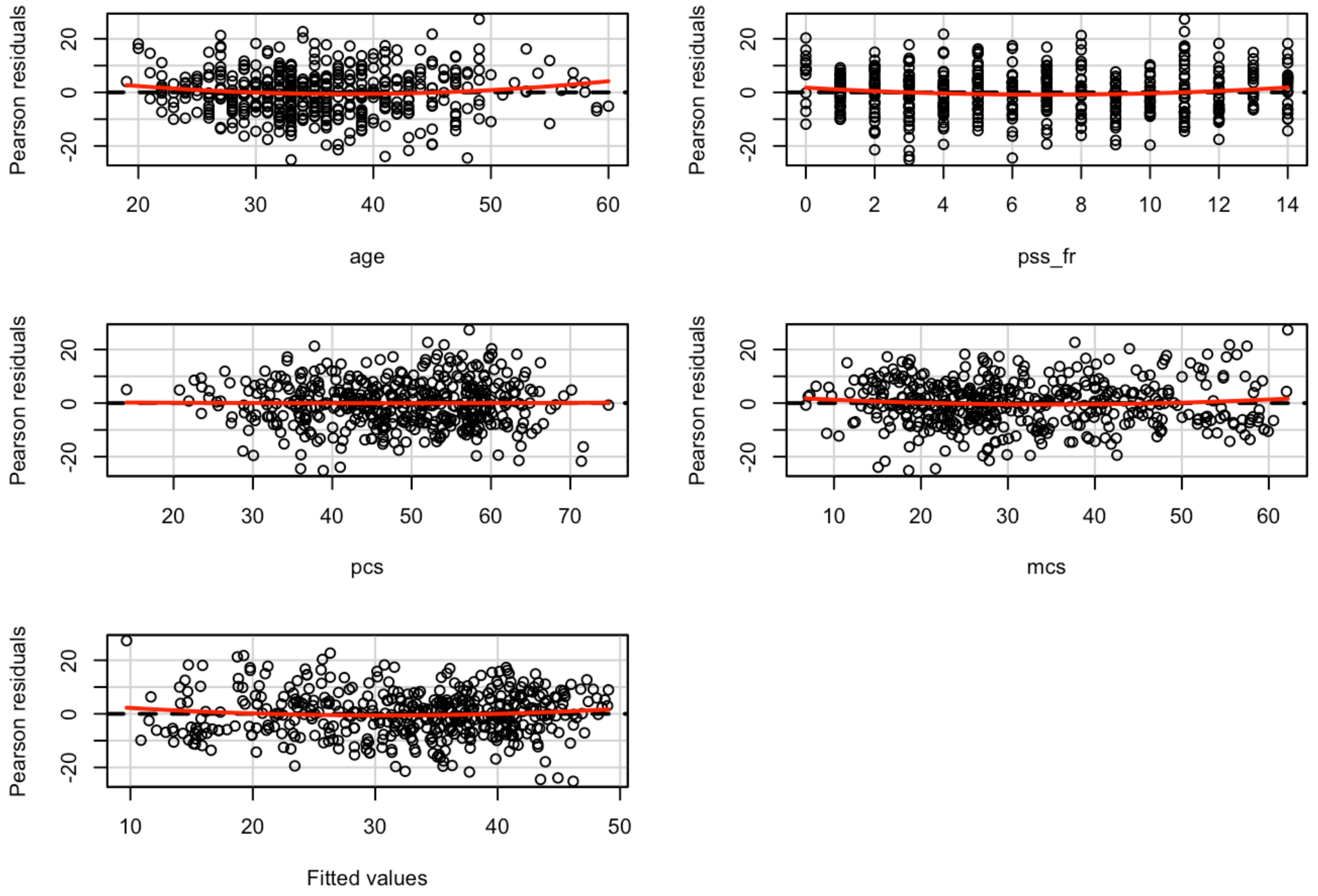
```
library(car)
```

```
##
## Attaching package: 'car'
```

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

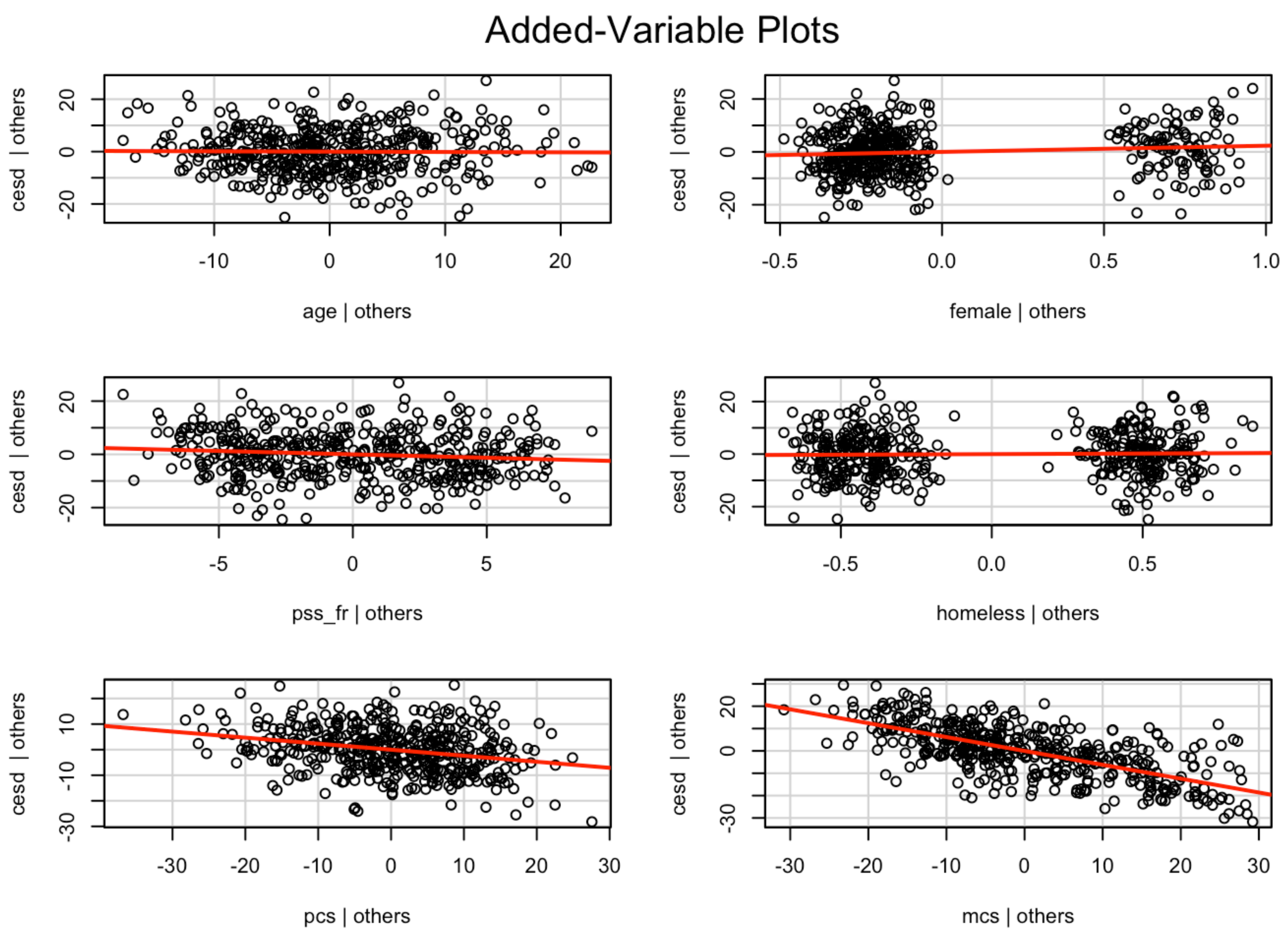
```
## The following object is masked from 'package:purrr':
##
##     some
```

```
#residual plot
residualPlots(Model2)
```

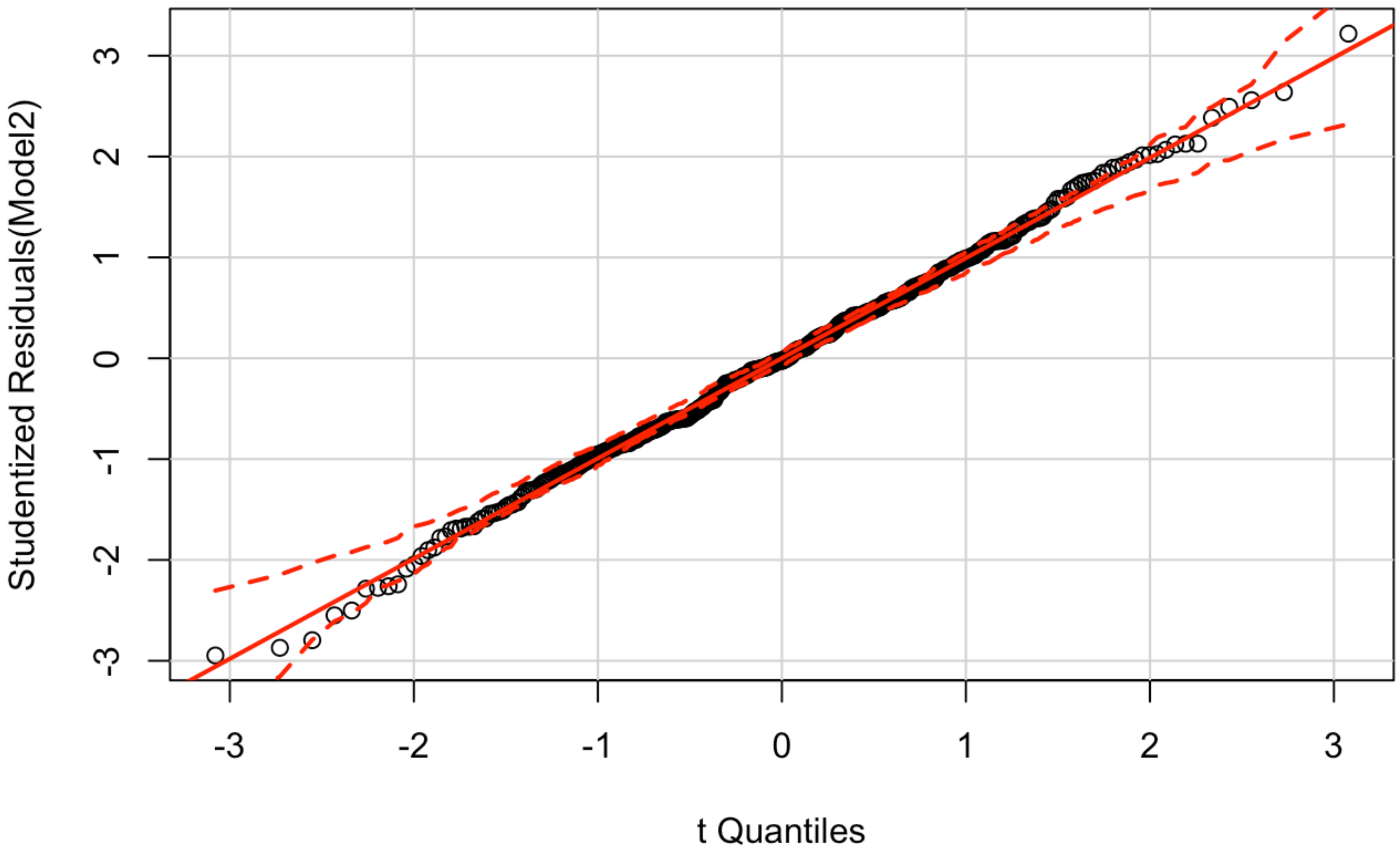



```
##            Test stat Pr(>|t|)
## age            1.941    0.053
## pss_fr         1.964    0.050
## pcs            0.081    0.936
## mcs            1.260    0.208
## Tukey test     1.434    0.152
```

```
#added variable plots
avPlots(Model2)
```



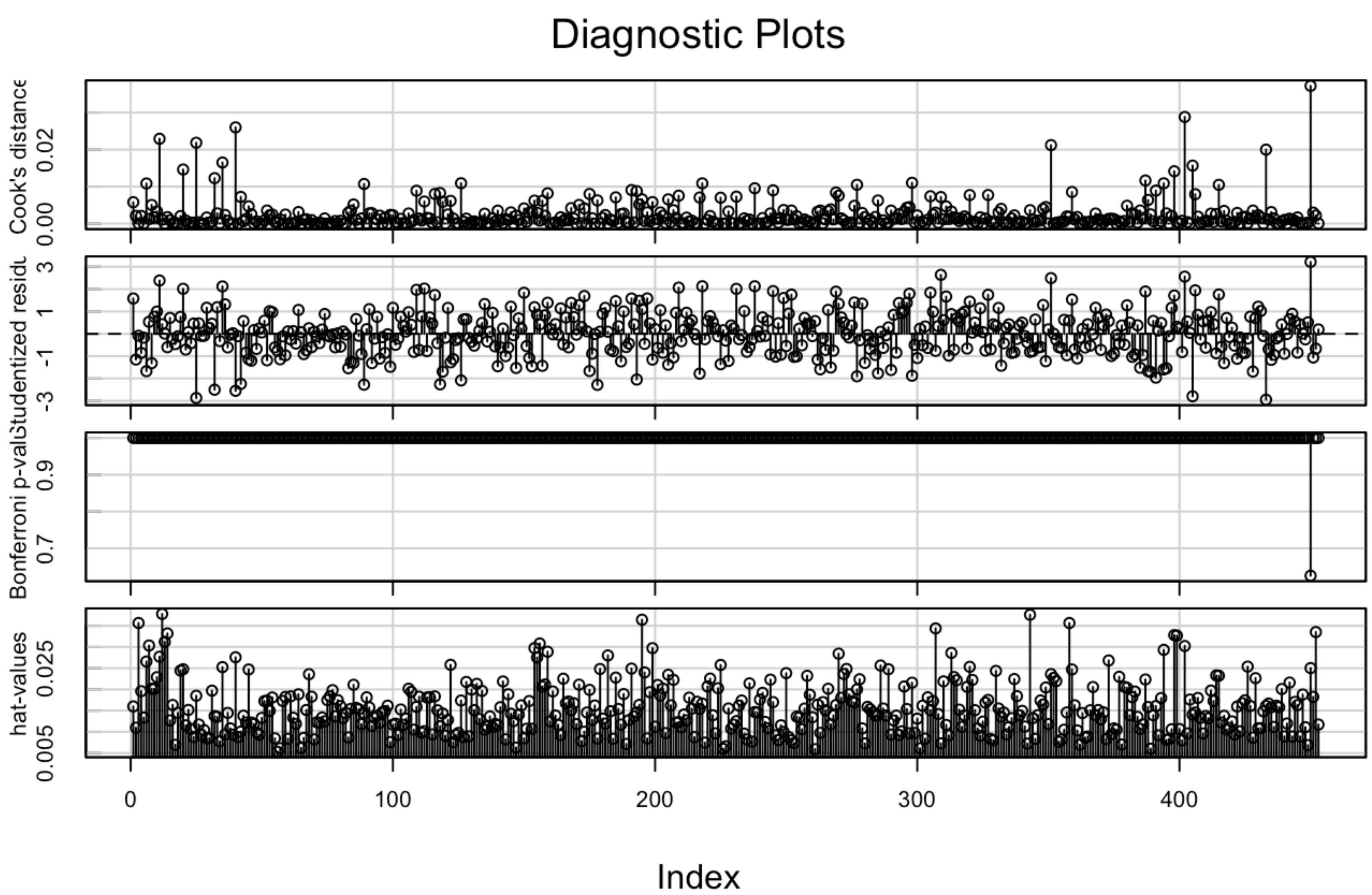

```
#qq plot
qqPlot(Model2)
```



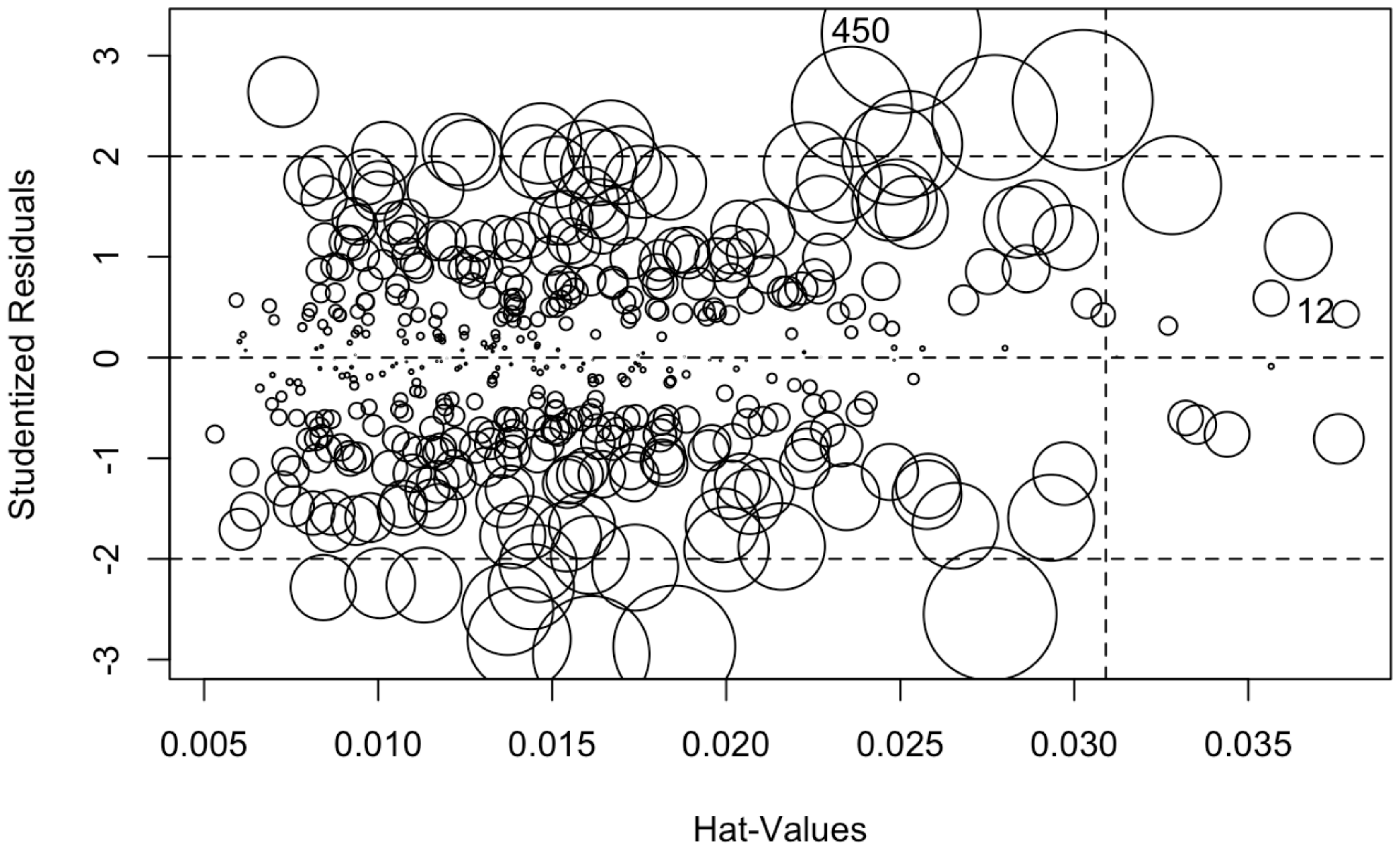

```
#run Bonferroni test for outliers
outlierTest(Model2)
```

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

```
#identify highly influential points
influenceIndexPlot(Model2)
```




```
#Influence plot
influencePlot(Model2)
```




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

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 sfinal fitted model and explain the coefficients. [**REMEMBER** to compute the Odds Ratios after you get the raw coefficient (betas)].

```
#logistic regression using cesd_gte16 as outcome
Model3 <- glm(cesd_gte16~mcs, data=h1, family=binomial)
```

```
# look at the model results
Model3
```

```
##
## Call:  glm(formula = cesd_gte16 ~ mcs, family = binomial, data = h1)
##
## Coefficients:
## (Intercept)          mcs
##      9.2691      -0.1716
##
## Degrees of Freedom: 452 Total (i.e. Null);  451 Residual
## Null Deviance:       297.6
## Residual Deviance: 174.7     AIC: 178.7
```

```
# summary of the model results
summary(Model3)
```

```
##
## 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
```

```
# coefficients of the model - these are the
# RAW Betas
coef(Model3)
```

```
## (Intercept)         mcs
##   9.2691224  -0.1715576
```

```
#take the exp to get the odds ratios
exp(coef(Model3))
```

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

```
#Interpretation: for each unit increase in mcs, cesd decreases by 0.17
```

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 (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).

```
# look at the predicted probabilities
Model3.predict <- predict(Model3, newdata=h1,
                      type="response")
```

```
# Look at the tradeoffs for at threshold
# of 0.5
# confusion matrix
table(h1$cesd_gte16, Model3.predict > 0.5)
```

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

```
#The model did a good job of predicting CESD scores => 16. It correclty predicted 395 of all the true cases
```

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

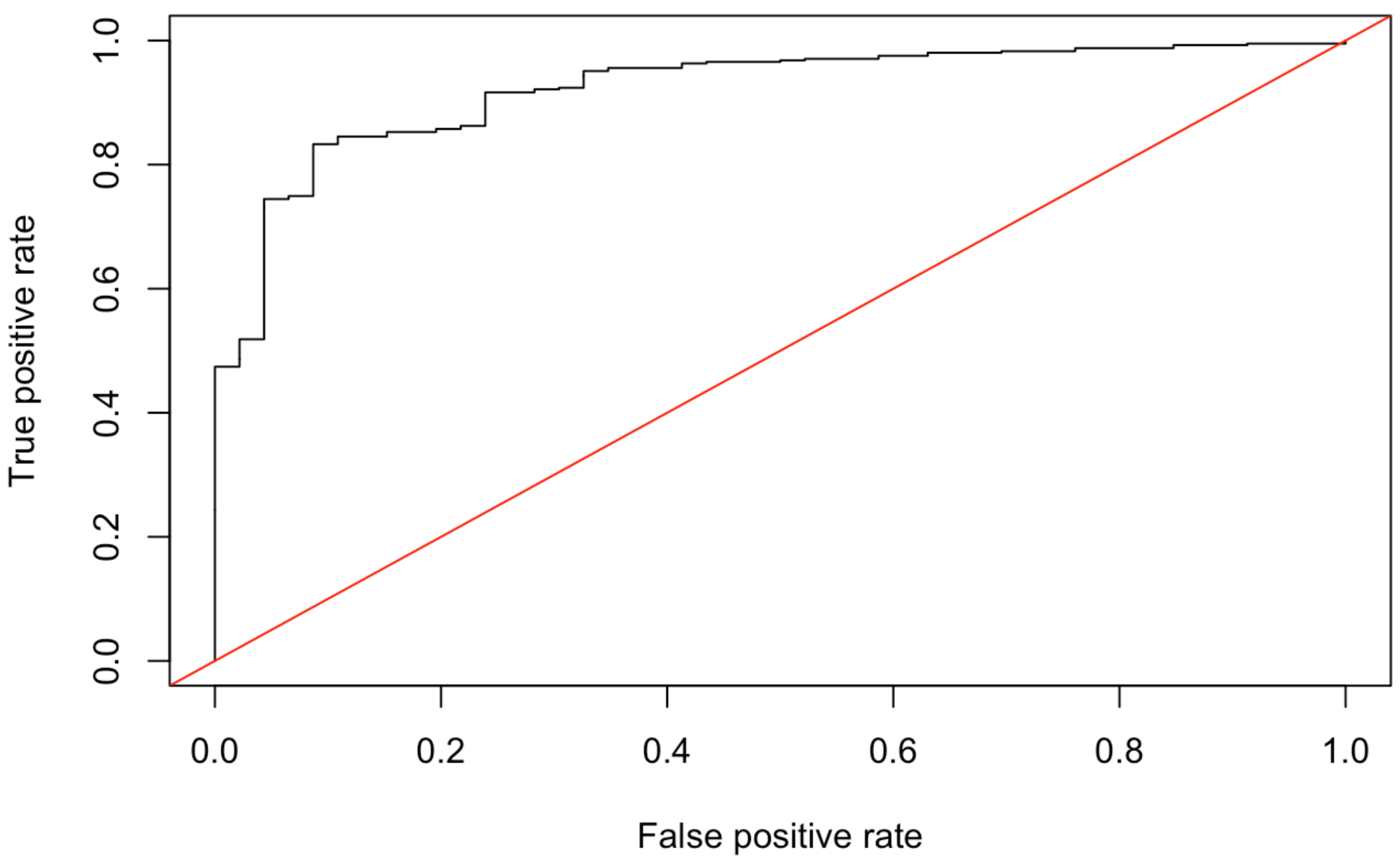
```
#ROC curve plot
library(ROCR)
```

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

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

```
p <- predict(Model3, 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")
```



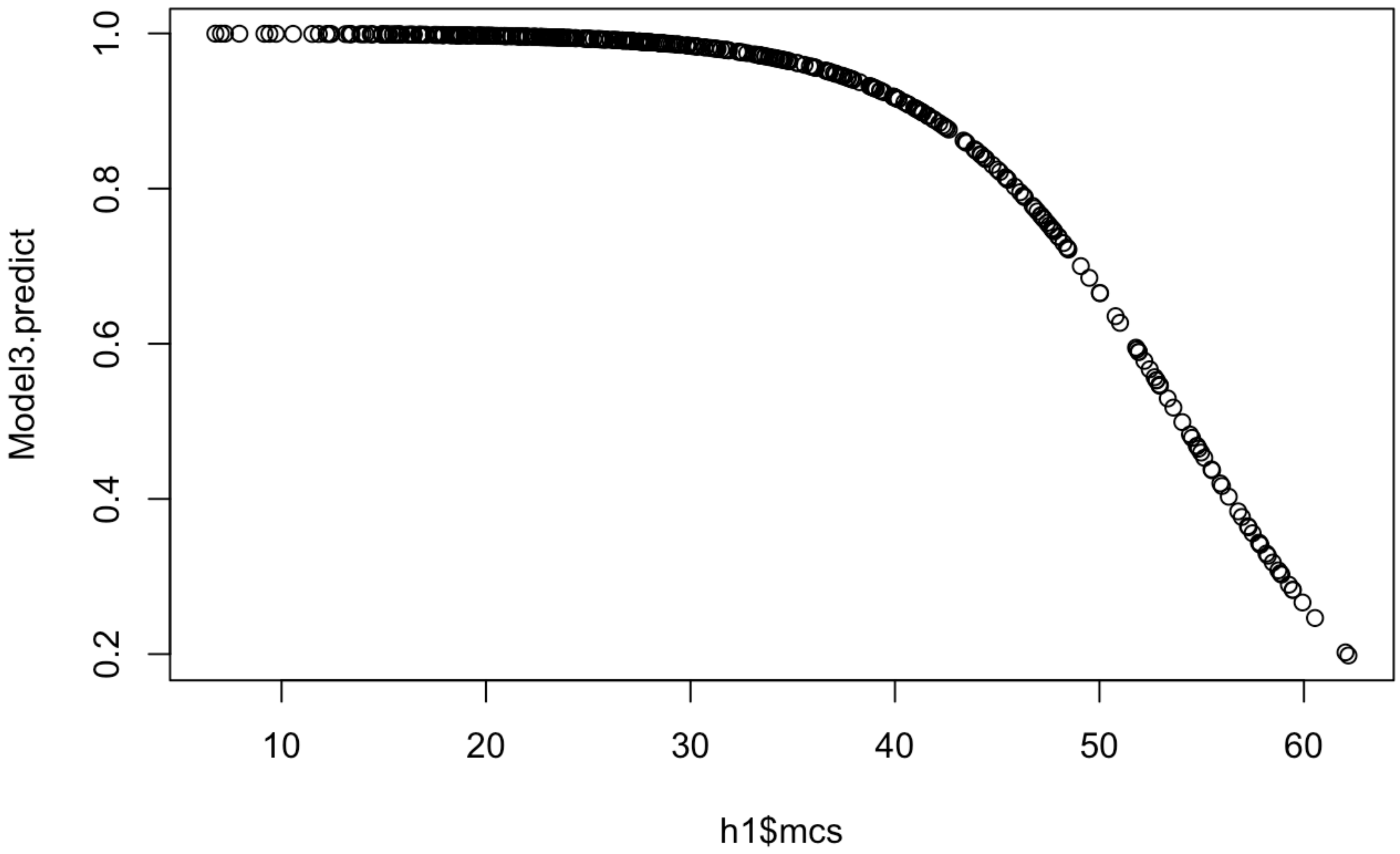

```
#Compute AUC
auc <- performance(pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

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

```
# The AUC is 0.922 indicating that this model is great for predicting 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? [**FYI** This plot is also called an "effect plot" is you're using `Rcmdr` to do these analyses.]

```
# plot the continuous predictor
# for these predicted probabilities
plot(h1$mcs, Model3.predict)
```




```
#Based on this plot, it does appear that mcs is a good predictor of depression.
```

The github repository for this assignment can be accessed via this link (https://github.com/RosemaryKinuthia/N741Spring2018_Homework6.g (https://github.com/RosemaryKinuthia/N741Spring2018_Homework6.g [https://github.com/RosemaryKinuthia/N741Spring2018_Homework6.g (https://github.com/RosemaryKinuthia/N741Spring2018_Homework6.g