N741 Spring 2018 - Homework 7

Homework 7 - DUE WED April 11, 2018

Melinda Higgins

April 4, 2018

Homework 7

Background and Information on HELP Dataset

For homework 7, you will be working with the **HELP** (Health Evaluation and Linkage to Primary Care) Dataset. See complete details posted in Homework 6.

Variables for Homework 7

For Homework 7, you will focus on these variables from the HELP dataset:

Table 1: Use these variables from HELP dataset for Homework 07

	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
cesd_gte16	Indicator of Depression

Homework 7 Assignment

SETUP Download and run the "loadHELP.R" `R` script (included in this Github repo https://github.com/melindahiggins2000/N741Spring2018_Homework6) to read in the HELP Dataset "helpmkh.sav". This script also pulls out the variables you need and creates the dichotomous variable for depression `cesd_gte16`.

After running this R script, you will have a data frame called `h1` you can use to do the rest of your analyses. You can also copy this code into your first R markdown code chunk to get you started on Homework 7.

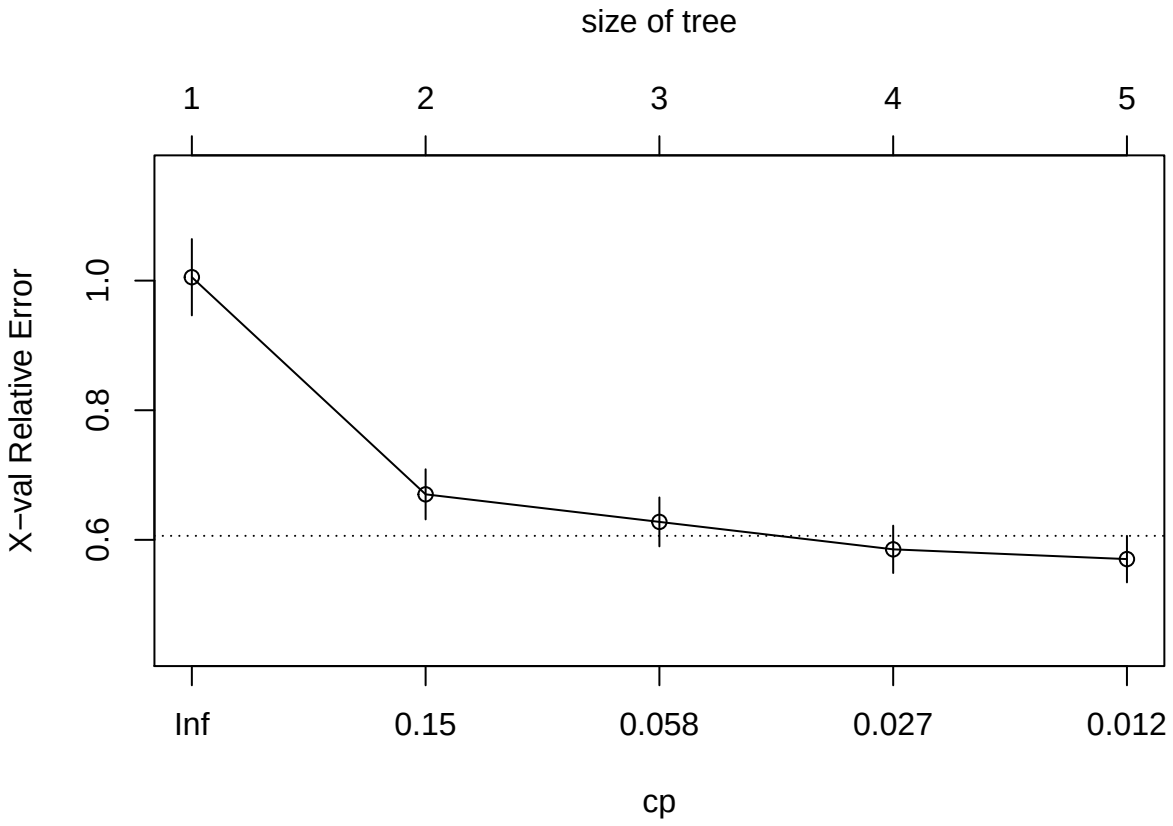
For Homework 7, you will be looking at perceived social support from friends (`pss_fr`) in these subjects.

```
library(rpart)
library(partykit)
# library(RColorBrewer)
library(reshape2)
# library(NHANES)
# library(mosaic)
# library(faraway) # skip this one
```

```r
fitcesd <- rpart::rpart(cesd ~ mcs, data = h1)
rpart::printcp(fitcesd) # Display the results
```

```
##
## Regression tree:
## rpart::rpart(formula = cesd ~ mcs, data = h1)
##
## Variables actually used in tree construction:
## [1] mcs
##
## Root node error: 70788/453 = 156.27
##
## n= 453
##
##         CP nsplit rel error  xerror     xstd
## 1 0.340353      0   1.00000 1.00526 0.058848
## 2 0.063092      1   0.65965 0.67023 0.038664
## 3 0.053626      2   0.59655 0.62769 0.037733
## 4 0.013872      3   0.54293 0.58533 0.036536
## 5 0.010000      4   0.52906 0.57026 0.035890
```

```r
rpart::plotcp(fitcesd) # Visualize cross-validation results
```

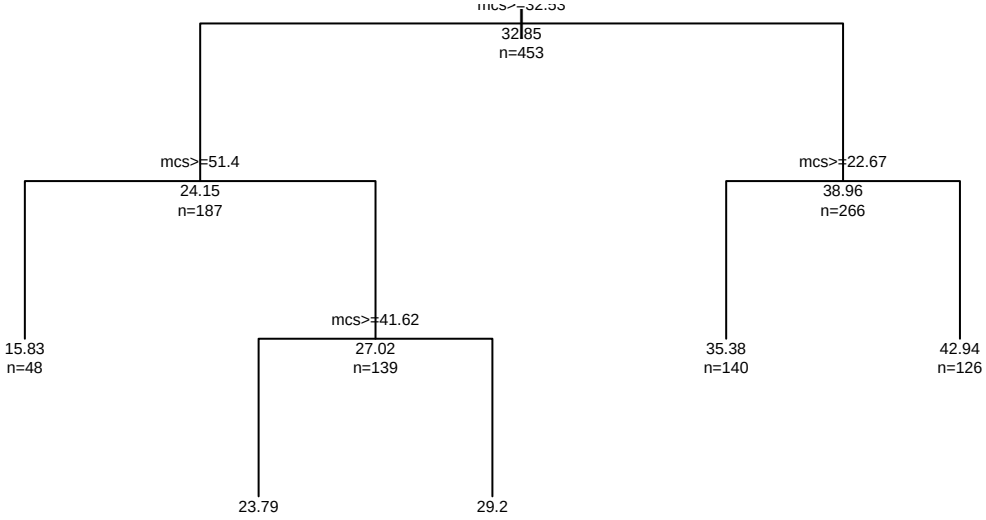


```r
summary(fitcesd) # Detailed summary of fit
```

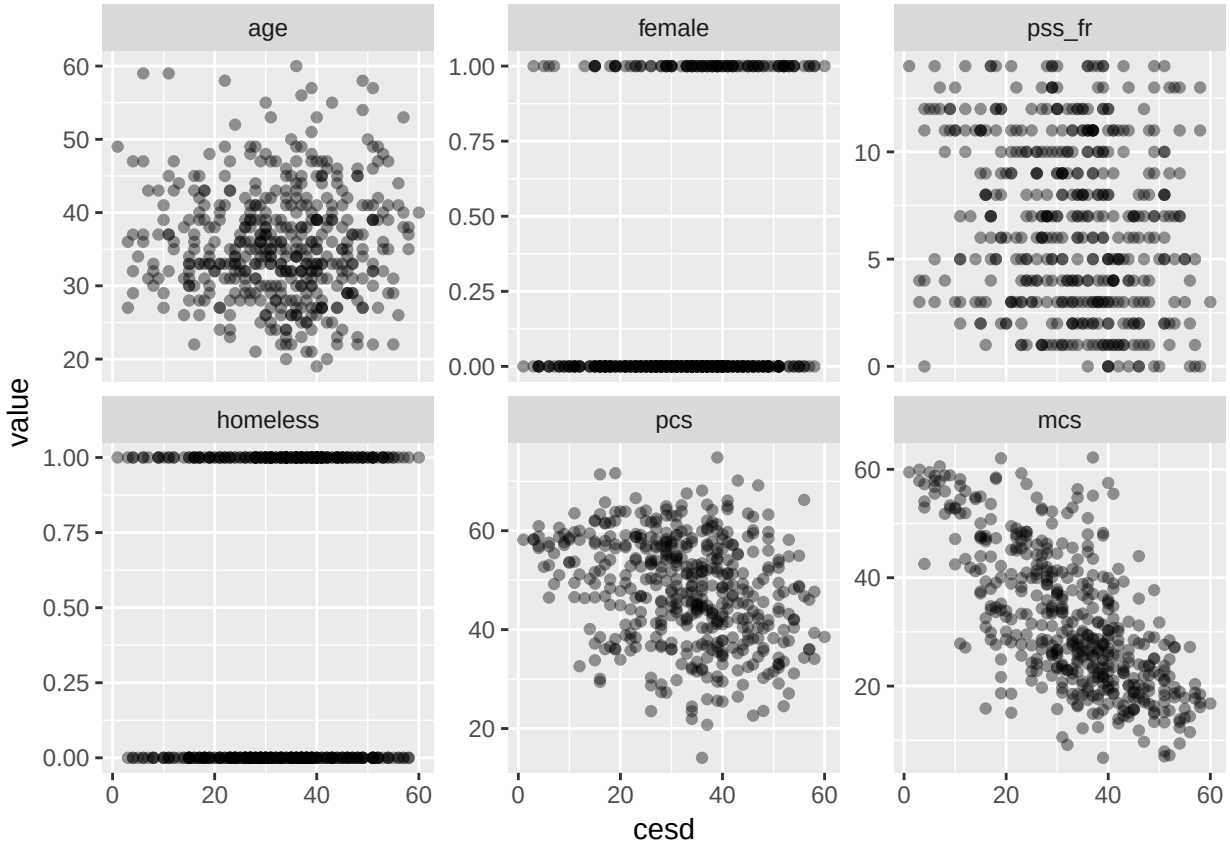
```
## Call:
## rpart::rpart(formula = cesd ~ mcs, data = h1)
##   n= 453
##
##           CP nsplit rel error    xerror       xstd
## 1 0.34035277      0 1.0000000 1.0052620 0.05884803
## 2 0.06309226      1 0.6596472 0.6702259 0.03866350
## 3 0.05362563      2 0.5965550 0.6276911 0.03773266
## 4 0.01387215      3 0.5429293 0.5853348 0.03653609
## 5 0.01000000      4 0.5290572 0.5702605 0.03588975
##
## Variable importance
## mcs
## 100
##
## Node number 1: 453 observations,    complexity param=0.3403528
##   mean=32.84768, MSE=156.266
##   left son=2 (187 obs) right son=3 (266 obs)
##   Primary splits:
##       mcs < 32.52559 to the right, improve=0.3403528, (0 missing)
##
## Node number 2: 187 observations,    complexity param=0.06309226
##   mean=24.14973, MSE=112.5979
##   left son=4 (48 obs) right son=5 (139 obs)
##   Primary splits:
##       mcs < 51.3962  to the right, improve=0.2121128, (0 missing)
##
## Node number 3: 266 observations,    complexity param=0.05362563
##   mean=38.96241, MSE=96.38956
##   left son=6 (140 obs) right son=7 (126 obs)
##   Primary splits:
##       mcs < 22.67163 to the right, improve=0.1480551, (0 missing)
##
## Node number 4: 48 observations
##   mean=15.83333, MSE=128.0556
##
## Node number 5: 139 observations,    complexity param=0.01387215
##   mean=27.02158, MSE=75.12903
##   left son=10 (56 obs) right son=11 (83 obs)
##   Primary splits:
##       mcs < 41.62456 to the right, improve=0.09403377, (0 missing)
##
## Node number 6: 140 observations
##   mean=35.37857, MSE=80.77811
##
## Node number 7: 126 observations
##   mean=42.94444, MSE=83.60802
##
## Node number 10: 56 observations
##   mean=23.78571, MSE=71.52551
##
```

```
## Node number 11: 83 observations
##   mean=29.20482, MSE=65.72913
```

```
# plot tree
plot(fitcesd, uniform = TRUE, compress = FALSE)
text(fitcesd, use.n = TRUE, all = TRUE, cex = 0.5)
```



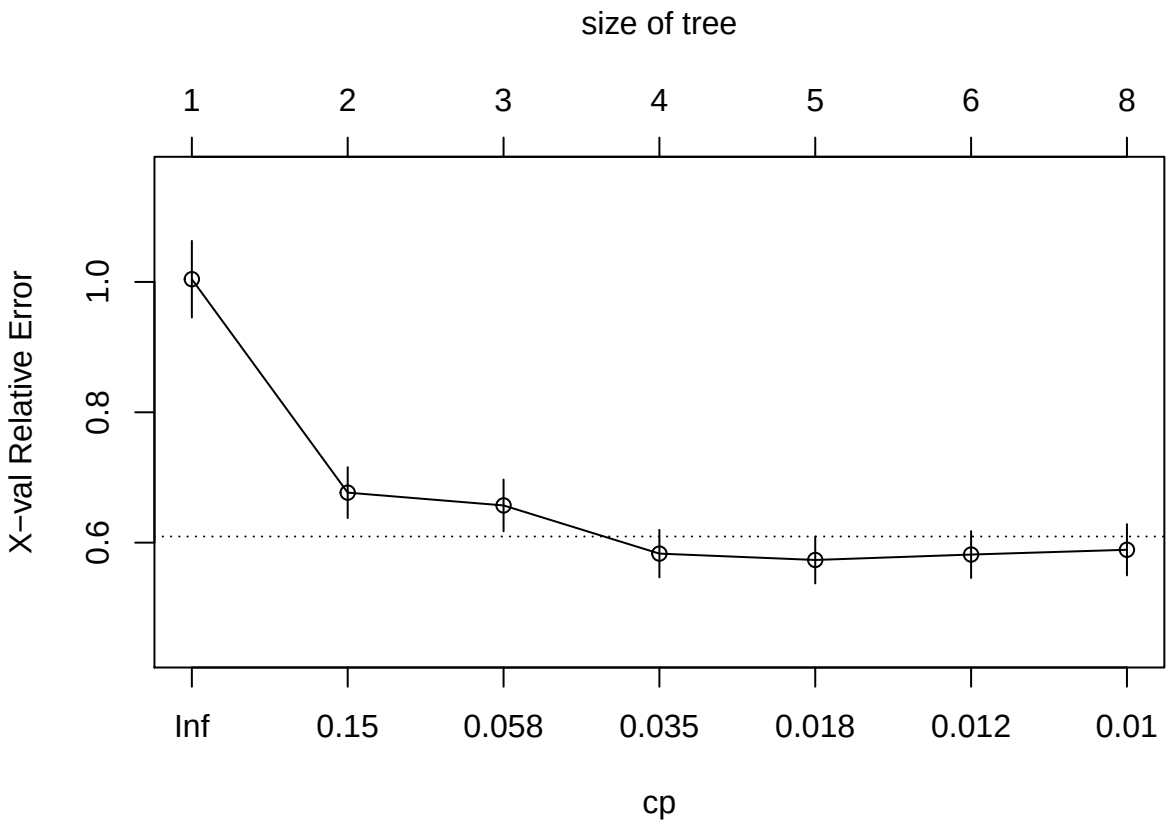

```
# all vars except the dictomous cesd_gte16
h1a <- h1[,1:7]

# Plot panels for each covariate
h1m <- reshape2::melt(h1a, id.vars = "cesd")
ggplot(h1m, aes(x=cesd, y=value)) +
  geom_point(alpha=0.4)+
  scale_color_brewer(palette="Set2")+
  facet_wrap(~variable, scales="free_y", ncol=3)
```




```
# fit a regression tree with all vars
fitall <- rpart::rpart(cesd ~ ., data = h1a)
# Now let's look at fitall
rpart::printcp(fitall) # Display the results
```

```
##
## Regression tree:
## rpart::rpart(formula = cesd ~ ., data = h1a)
##
## Variables actually used in tree construction:
## [1] mcs pcs
##
## Root node error: 70788/453 = 156.27
##
## n= 453
##
##         CP nsplit rel error  xerror     xstd
## 1 0.340353      0   1.00000 1.00422 0.058792
## 2 0.063092      1   0.65965 0.67674 0.038955
## 3 0.053626      2   0.59655 0.65710 0.039666
## 4 0.022423      3   0.54293 0.58316 0.036388
## 5 0.013872      4   0.52051 0.57342 0.036002
## 6 0.010032      5   0.50663 0.58172 0.035947
## 7 0.010000      7   0.48657 0.58905 0.039295
```

```r
rpart::plotcp(fitall) # Visualize cross-validation results
```




```r
summary(fitall) # Detailed summary of fit
```

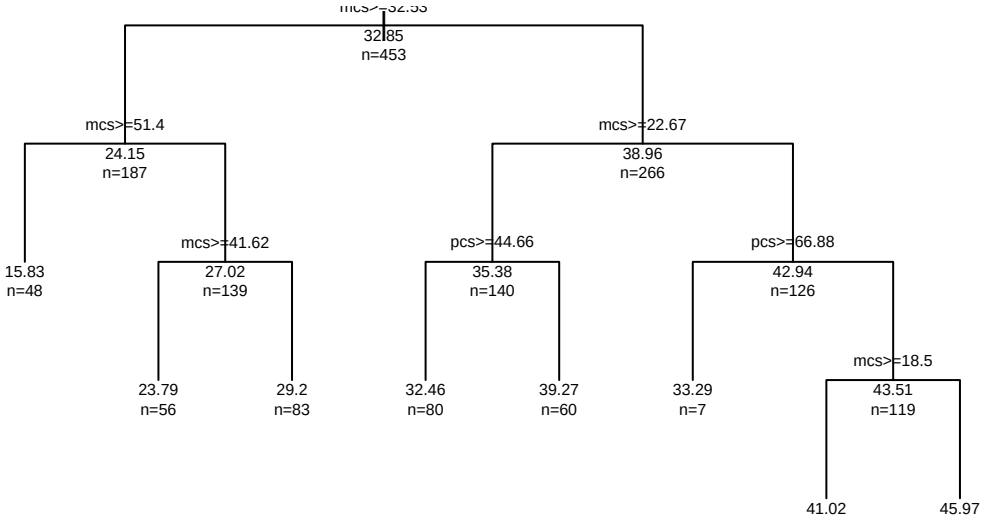
```
## Call:
## rpart::rpart(formula = cesd ~ ., data = h1a)
##   n= 453
##
##           CP nsplit rel error    xerror       xstd
## 1 0.34035277      0 1.0000000 1.0042201 0.05879232
## 2 0.06309226      1 0.6596472 0.6767410 0.03895500
## 3 0.05362563      2 0.5965550 0.6571024 0.03966574
## 4 0.02242335      3 0.5429293 0.5831605 0.03638849
## 5 0.01387215      4 0.5205060 0.5734209 0.03600228
## 6 0.01003176      5 0.5066338 0.5817197 0.03594676
## 7 0.01000000      7 0.4865703 0.5890488 0.03929521
##
## Variable importance
##    mcs    pcs pss_fr    age female
##     78     13      5      3      1
##
## Node number 1: 453 observations,    complexity param=0.3403528
##   mean=32.84768, MSE=156.266
##   left son=2 (187 obs) right son=3 (266 obs)
##   Primary splits:
```

```
##       mcs    < 32.52559 to the right, improve=0.340352800, (0 missing)
##       pcs    < 49.19916 to the right, improve=0.104572600, (0 missing)
##       female < 0.5      to the left,  improve=0.032302950, (0 missing)
##       pss_fr < 8.5      to the right, improve=0.029240370, (0 missing)
##       age    < 23.5     to the right, improve=0.007589837, (0 missing)
##   Surrogate splits:
##       pcs    < 56.1551  to the right, agree=0.634, adj=0.112, (0 split)
##       pss_fr < 10.5     to the right, agree=0.609, adj=0.053, (0 split)
##       age    < 21.5     to the left,  agree=0.592, adj=0.011, (0 split)
##
## Node number 2: 187 observations,    complexity param=0.06309226
##   mean=24.14973, MSE=112.5979
##   left son=4 (48 obs) right son=5 (139 obs)
##   Primary splits:
##       mcs    < 51.3962  to the right, improve=0.21211280, (0 missing)
##       pcs    < 46.0814  to the right, improve=0.07616853, (0 missing)
##       pss_fr < 11.5     to the right, improve=0.03161969, (0 missing)
##       age    < 22.5     to the right, improve=0.02449595, (0 missing)
##       female < 0.5      to the left,  improve=0.01088789, (0 missing)
##   Surrogate splits:
##       pss_fr < 11.5     to the right, agree=0.765, adj=0.083, (0 split)
##       age    < 58.5     to the right, agree=0.754, adj=0.042, (0 split)
##
## Node number 3: 266 observations,    complexity param=0.05362563
##   mean=38.96241, MSE=96.38956
##   left son=6 (140 obs) right son=7 (126 obs)
##   Primary splits:
##       mcs    < 22.67163 to the right, improve=0.14805510, (0 missing)
##       pcs    < 40.92127 to the right, improve=0.07769934, (0 missing)
##       pss_fr < 0.5      to the right, improve=0.03572097, (0 missing)
##       female < 0.5      to the left,  improve=0.03455917, (0 missing)
##       age    < 48.5     to the left,  improve=0.01737694, (0 missing)
##   Surrogate splits:
##       pss_fr   < 3.5      to the right, agree=0.583, adj=0.119, (0 split)
##       pcs      < 64.93552 to the left,  agree=0.560, adj=0.071, (0 split)
##       female   < 0.5      to the left,  agree=0.556, adj=0.063, (0 split)
##       age      < 46.5     to the left,  agree=0.553, adj=0.056, (0 split)
##       homeless < 0.5      to the right, agree=0.530, adj=0.008, (0 split)
##
## Node number 4: 48 observations
##   mean=15.83333, MSE=128.0556
##
## Node number 5: 139 observations,    complexity param=0.01387215
##   mean=27.02158, MSE=75.12903
##   left son=10 (56 obs) right son=11 (83 obs)
##   Primary splits:
##       mcs      < 41.62456 to the right, improve=0.09403377, (0 missing)
##       pcs      < 26.8635  to the right, improve=0.07496568, (0 missing)
##       pss_fr   < 3.5      to the right, improve=0.02872252, (0 missing)
##       age      < 33.5     to the left,  improve=0.01948280, (0 missing)
##       homeless < 0.5      to the left,  improve=0.01404178, (0 missing)
##   Surrogate splits:
##       pcs    < 22.26483 to the left,  agree=0.619, adj=0.054, (0 split)
##       pss_fr < 13.5     to the right, agree=0.612, adj=0.036, (0 split)
```

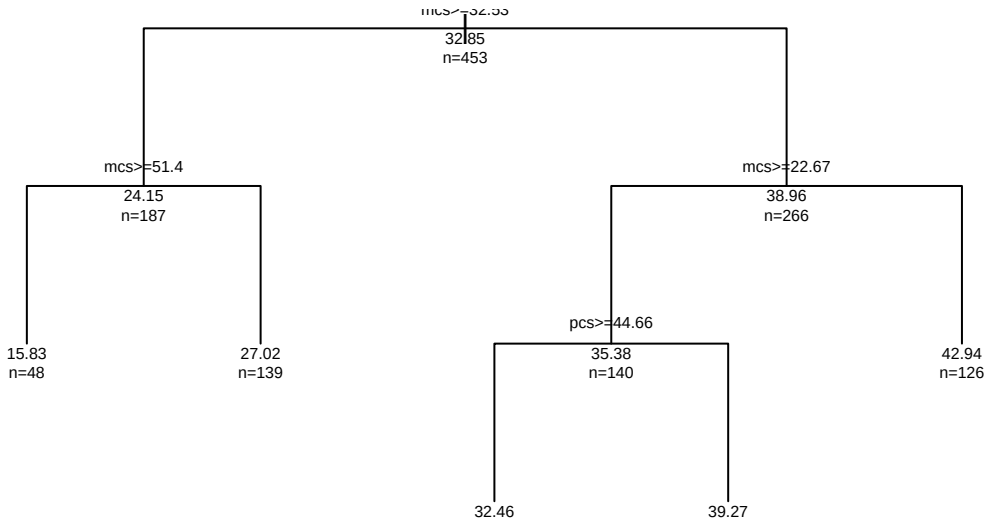
```
##
## Node number 6: 140 observations,    complexity param=0.02242335
##   mean=35.37857, MSE=80.77811
##   left son=12 (80 obs) right son=13 (60 obs)
##   Primary splits:
##       pcs    < 44.6562  to the right, improve=0.140359400, (0 missing)
##       pss_fr < 8.5      to the right, improve=0.069217610, (0 missing)
##       age    < 38.5     to the left,  improve=0.044384950, (0 missing)
##       mcs    < 27.62416 to the right, improve=0.021316600, (0 missing)
##       female < 0.5      to the left,  improve=0.007874331, (0 missing)
##   Surrogate splits:
##       age      < 36.5     to the left,  agree=0.686, adj=0.267, (0 split)
##       mcs      < 23.7272  to the right, agree=0.621, adj=0.117, (0 split)
##       homeless < 0.5      to the left,  agree=0.593, adj=0.050, (0 split)
##       pss_fr   < 0.5      to the right, agree=0.579, adj=0.017, (0 split)
##
## Node number 7: 126 observations,    complexity param=0.01003176
##   mean=42.94444, MSE=83.60802
##   left son=14 (7 obs) right son=15 (119 obs)
##   Primary splits:
##       pcs    < 66.88379 to the right, improve=0.06563616, (0 missing)
##       mcs    < 18.49567 to the right, improve=0.05724195, (0 missing)
##       female < 0.5      to the left,  improve=0.05365277, (0 missing)
##       age    < 48.5     to the left,  improve=0.03224087, (0 missing)
##       pss_fr < 12.5     to the left,  improve=0.02139818, (0 missing)
##
## Node number 10: 56 observations
##   mean=23.78571, MSE=71.52551
##
## Node number 11: 83 observations
##   mean=29.20482, MSE=65.72913
##
## Node number 12: 80 observations
##   mean=32.4625, MSE=73.84859
##
## Node number 13: 60 observations
##   mean=39.26667, MSE=63.56222
##
## Node number 14: 7 observations
##   mean=33.28571, MSE=117.9184
##
## Node number 15: 119 observations,    complexity param=0.01003176
##   mean=43.51261, MSE=75.77925
##   left son=30 (59 obs) right son=31 (60 obs)
##   Primary splits:
##       mcs    < 18.49567 to the right, improve=0.08082019, (0 missing)
##       female < 0.5      to the left,  improve=0.04062765, (0 missing)
##       pcs    < 35.99184 to the right, improve=0.03933964, (0 missing)
##       age    < 48.5     to the left,  improve=0.03130734, (0 missing)
##       pss_fr < 12.5     to the left,  improve=0.02626137, (0 missing)
##   Surrogate splits:
##       pcs      < 46.51692 to the left,  agree=0.672, adj=0.339, (0 split)
##       pss_fr   < 8.5      to the left,  agree=0.622, adj=0.237, (0 split)
##       age      < 31.5     to the right, agree=0.613, adj=0.220, (0 split)
```

```
##       homeless < 0.5      to the right, agree=0.571, adj=0.136, (0 split)
##       female   < 0.5      to the left,  agree=0.521, adj=0.034, (0 split)
##
## Node number 30: 59 observations
##   mean=41.01695, MSE=69.67768
##
## Node number 31: 60 observations
##   mean=45.96667, MSE=69.63222
```

```r
plot(fitall, uniform = TRUE, compress = FALSE, main = "Regression Tree for CESD Scores from HELP(h1) Da
text(fitall, use.n = TRUE, all = TRUE, cex = 0.5)
```

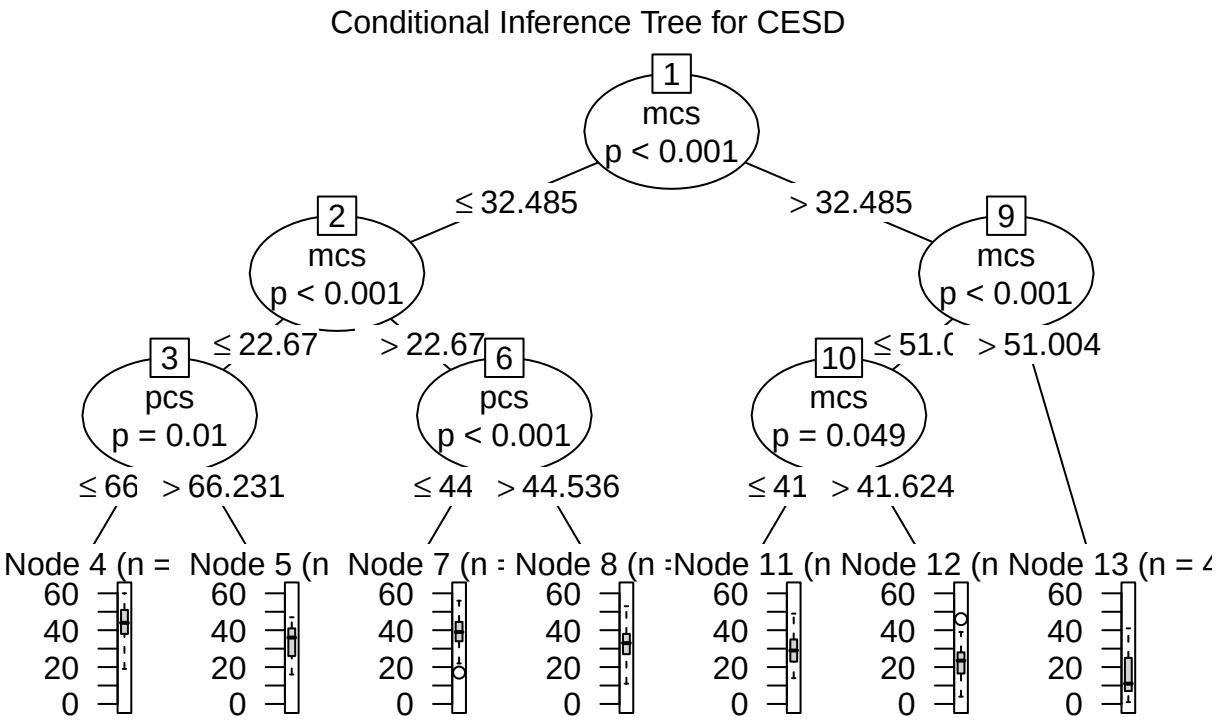



```r
# Prune the tree
pfit <- prune(fitall, cp = fitall$cptable[which.min(fitall$cptable[, "xerror"]), "CP"])

# Plot the pruned tree
plot(pfit, uniform = TRUE, compress = FALSE, main = "Pruned Regression Tree for Ozone")
text(pfit, use.n = TRUE, all = TRUE, cex = 0.5)
```



better graphics with `party`

```
library(party)
fitallp <- party::ctree(cesd ~ ., data = h1a)
plot(fitallp, main = "Conditional Inference Tree for CESD")
```



decision trees - outcome is a group or class

```
# begin with a logistic regression - depressed or not
glm1 <- glm(cesd_gte16 ~ age + female + pss_fr + homeless +
              pcs + mcs, data = h1)
summary(glm1)
```

```
##
## Call:
## glm(formula = cesd_gte16 ~ age + female + pss_fr + homeless +
##     pcs + mcs, data = h1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.90801  -0.06647   0.02642   0.14484   0.51900
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.5529621  0.0948791  16.368  < 2e-16 ***
## age         -0.0018844  0.0016377  -1.151 0.250517
## female      -0.0309994  0.0294192  -1.054 0.292584
## pss_fr      -0.0034492  0.0031460  -1.096 0.273509
## homeless    -0.0045879  0.0250875  -0.183 0.854978
## pcs         -0.0039722  0.0011870  -3.346 0.000888 ***
## mcs         -0.0114878  0.0009709 -11.832  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.0668409)
##
##     Null deviance: 41.329  on 452  degrees of freedom
## Residual deviance: 29.811  on 446  degrees of freedom
## AIC: 68.939
##
## Number of Fisher Scoring iterations: 2
```

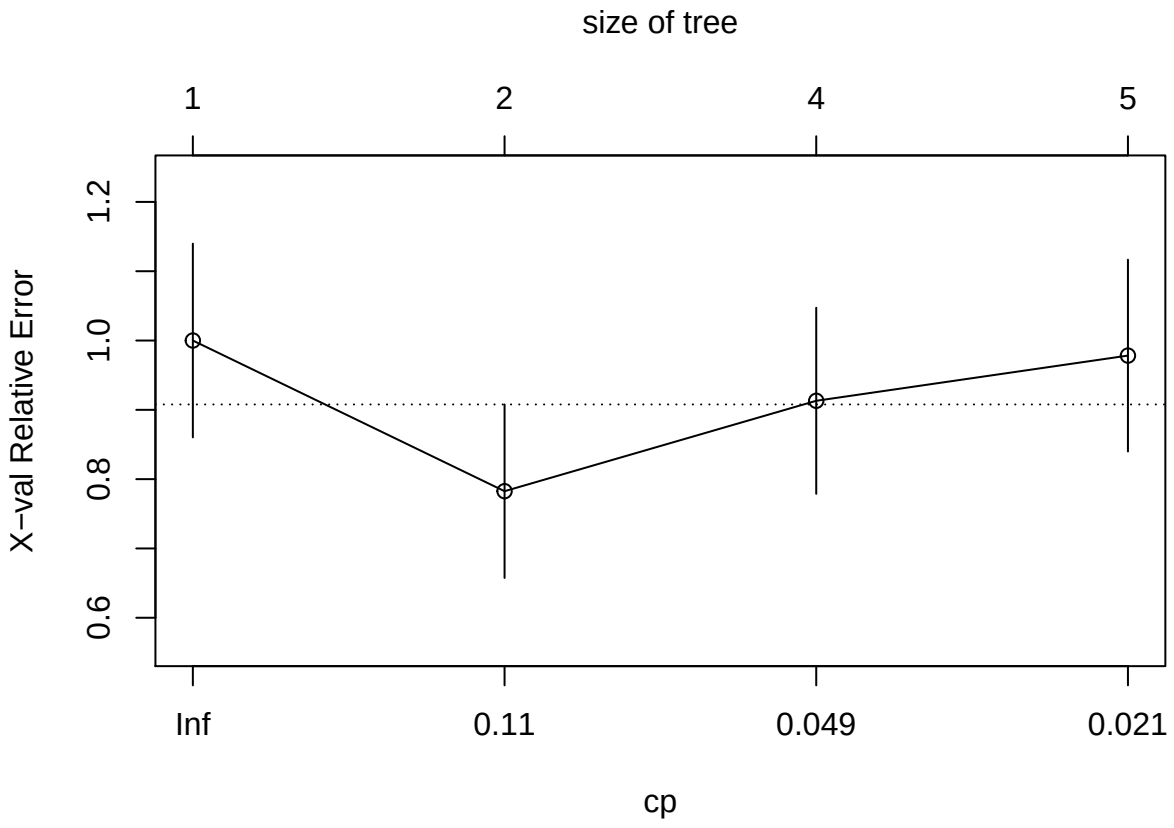
```
fitk <- rpart::rpart(cesd_gte16 ~ age + female + pss_fr +
                       homeless + pcs + mcs,
                     method = "class", data = h1)
class(fitk)
```

```
## [1] "rpart"
```

```
# Display the results
rpart::printcp(fitk)
```

```
##
## Classification tree:
## rpart::rpart(formula = cesd_gte16 ~ age + female + pss_fr + homeless +
##     pcs + mcs, data = h1, method = "class")
##
## Variables actually used in tree construction:
## [1] age mcs pcs
##
## Root node error: 46/453 = 0.10155
##
## n= 453
##
##         CP nsplit rel error  xerror    xstd
## 1 0.239130      0   1.00000 1.00000 0.13976
## 2 0.054348      1   0.76087 0.78261 0.12514
## 3 0.043478      3   0.65217 0.91304 0.13420
## 4 0.010000      4   0.60870 0.97826 0.13840
```

```
#Visualize the cross-validation results
rpart::plotcp(fitk)
```



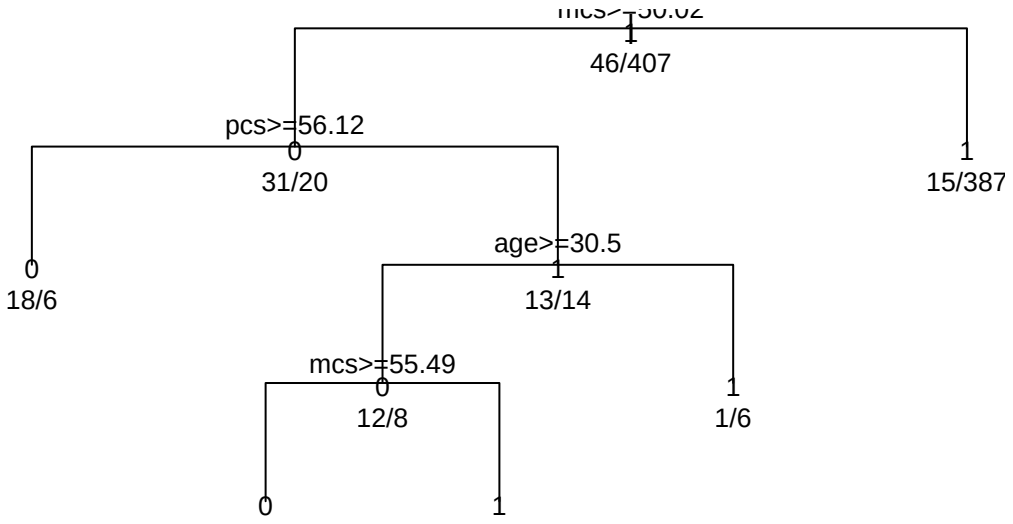

```
# Get a detailed summary of the splits
summary(fitk)
```

```
## Call:
## rpart::rpart(formula = cesd_gte16 ~ age + female + pss_fr + homeless +
##     pcs + mcs, data = h1, method = "class")
##   n= 453
##
##           CP nsplit rel error    xerror      xstd
## 1 0.23913043      0 1.0000000 1.0000000 0.1397556
## 2 0.05434783      1 0.7608696 0.7826087 0.1251447
## 3 0.04347826      3 0.6521739 0.9130435 0.1341957
## 4 0.01000000      4 0.6086957 0.9782609 0.1383979
##
## Variable importance
##      mcs      age      pcs   pss_fr homeless
##       84        8        5        1        1
##
## Node number 1: 453 observations,    complexity param=0.2391304
##   predicted class=1  expected loss=0.1015453  P(node) =1
##     class counts:    46   407
##    probabilities: 0.102 0.898
##   left son=2 (51 obs) right son=3 (402 obs)
##   Primary splits:
##       mcs      < 50.02446 to the right, improve=29.4635100, (0 missing)
##       pcs      < 49.19916 to the right, improve= 4.2774340, (0 missing)
```

```
##       pss_fr   < 10.5     to the right, improve= 3.6879600, (0 missing)
##       age      < 25.5     to the right, improve= 0.7580753, (0 missing)
##       homeless < 0.5      to the left,  improve= 0.1845446, (0 missing)
##   Surrogate splits:
##       age < 58.5     to the right, agree=0.89, adj=0.02, (0 split)
##
## Node number 2: 51 observations,    complexity param=0.05434783
##   predicted class=0  expected loss=0.3921569  P(node) =0.1125828
##     class counts:    31    20
##    probabilities: 0.608 0.392
##   left son=4 (24 obs) right son=5 (27 obs)
##   Primary splits:
##       pcs      < 56.1216  to the right, improve=1.83224400, (0 missing)
##       age      < 28.5     to the right, improve=1.68385500, (0 missing)
##       mcs      < 52.79105 to the right, improve=0.65918000, (0 missing)
##       homeless < 0.5      to the right, improve=0.11695130, (0 missing)
##       pss_fr   < 8.5      to the right, improve=0.09467787, (0 missing)
##   Surrogate splits:
##       mcs      < 54.23909 to the left,  agree=0.647, adj=0.250, (0 split)
##       homeless < 0.5      to the left,  agree=0.588, adj=0.125, (0 split)
##       age      < 37.5     to the left,  agree=0.569, adj=0.083, (0 split)
##       female   < 0.5      to the left,  agree=0.569, adj=0.083, (0 split)
##       pss_fr   < 12.5     to the right, agree=0.569, adj=0.083, (0 split)
##
## Node number 3: 402 observations
##   predicted class=1  expected loss=0.03731343  P(node) =0.8874172
##     class counts:    15   387
##    probabilities: 0.037 0.963
##
## Node number 4: 24 observations
##   predicted class=0  expected loss=0.25  P(node) =0.05298013
##     class counts:    18     6
##    probabilities: 0.750 0.250
##
## Node number 5: 27 observations,    complexity param=0.05434783
##   predicted class=1  expected loss=0.4814815  P(node) =0.05960265
##     class counts:    13    14
##    probabilities: 0.481 0.519
##   left son=10 (20 obs) right son=11 (7 obs)
##   Primary splits:
##       age    < 30.5     to the right, improve=2.16719600, (0 missing)
##       mcs    < 54.81272 to the right, improve=1.81481500, (0 missing)
##       pcs    < 53.12609 to the left,  improve=1.21832400, (0 missing)
##       pss_fr < 4.5      to the right, improve=0.72433860, (0 missing)
##       female < 0.5      to the left,  improve=0.05291005, (0 missing)
##
## Node number 10: 20 observations,    complexity param=0.04347826
##   predicted class=0  expected loss=0.4  P(node) =0.04415011
##     class counts:    12     8
##    probabilities: 0.600 0.400
##   left son=20 (12 obs) right son=21 (8 obs)
##   Primary splits:
##       mcs      < 55.49419 to the right, improve=1.35000000, (0 missing)
##       pcs      < 50.26239 to the right, improve=0.40000000, (0 missing)
```

```
##       pss_fr   < 6        to the right, improve=0.26666670, (0 missing)
##       homeless < 0.5      to the right, improve=0.14545450, (0 missing)
##       age      < 34.5     to the right, improve=0.01758242, (0 missing)
##   Surrogate splits:
##       pss_fr   < 4.5      to the right, agree=0.70, adj=0.250, (0 split)
##       age      < 44       to the left,  agree=0.65, adj=0.125, (0 split)
##       homeless < 0.5      to the left,  agree=0.65, adj=0.125, (0 split)
##       pcs      < 50.20288 to the left,  agree=0.65, adj=0.125, (0 split)
##
## Node number 11: 7 observations
##   predicted class=1  expected loss=0.1428571  P(node) =0.01545254
##     class counts:     1     6
##    probabilities: 0.143 0.857
##
## Node number 20: 12 observations
##   predicted class=0  expected loss=0.25  P(node) =0.02649007
##     class counts:     9     3
##    probabilities: 0.750 0.250
##
## Node number 21: 8 observations
##   predicted class=1  expected loss=0.375  P(node) =0.01766004
##     class counts:     3     5
##    probabilities: 0.375 0.625
```

```r
# Plot the tree
plot(fitk, uniform = TRUE,
     main = "Classification Tree for CESD => 16")
text(fitk, use.n = TRUE, all = TRUE, cex = 0.8)
```



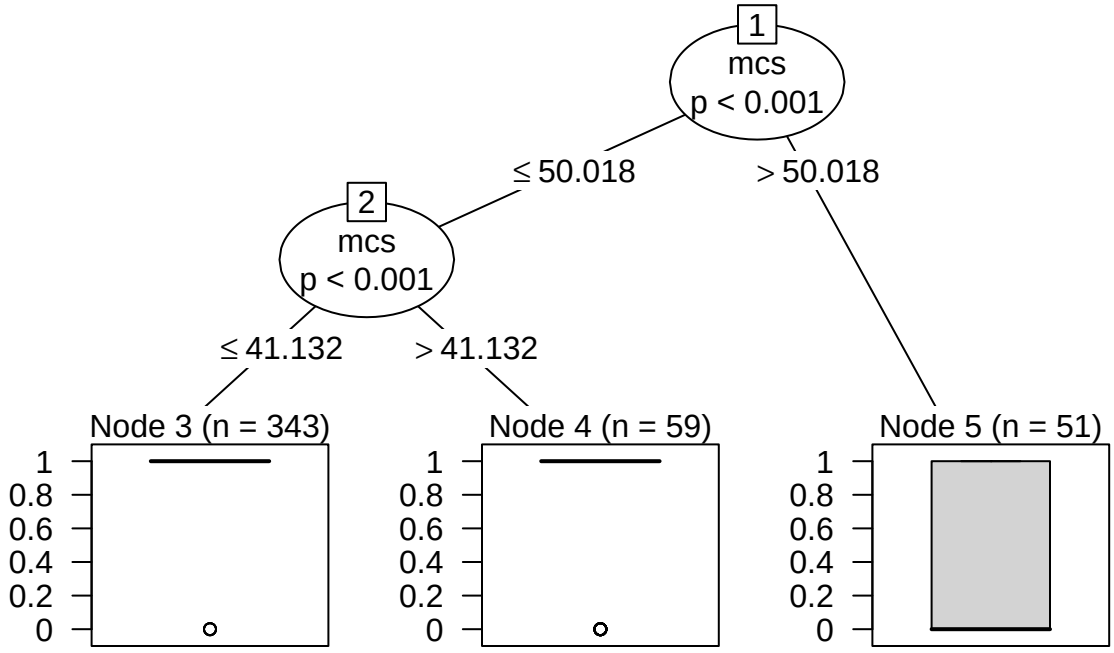

```
# Prune the tree
prune_fitk <- prune(fitk, cp = fitk$cptable[which.min(fitk$cptable[, "xerror"]), "CP"])
class(prune_fitk)
```

```
## [1] "rpart"
```

```
# look at cesd_gte16 with ctree from party

h1b <- h1 %>%
  select(age, female, pss_fr, homeless, pcs, mcs, cesd_gte16)
fitallpk <- party::ctree(cesd_gte16 ~ ., data = h1b)
class(fitallpk)
```

```
## [1] "BinaryTree"
## attr(,"package")
## [1] "party"
```

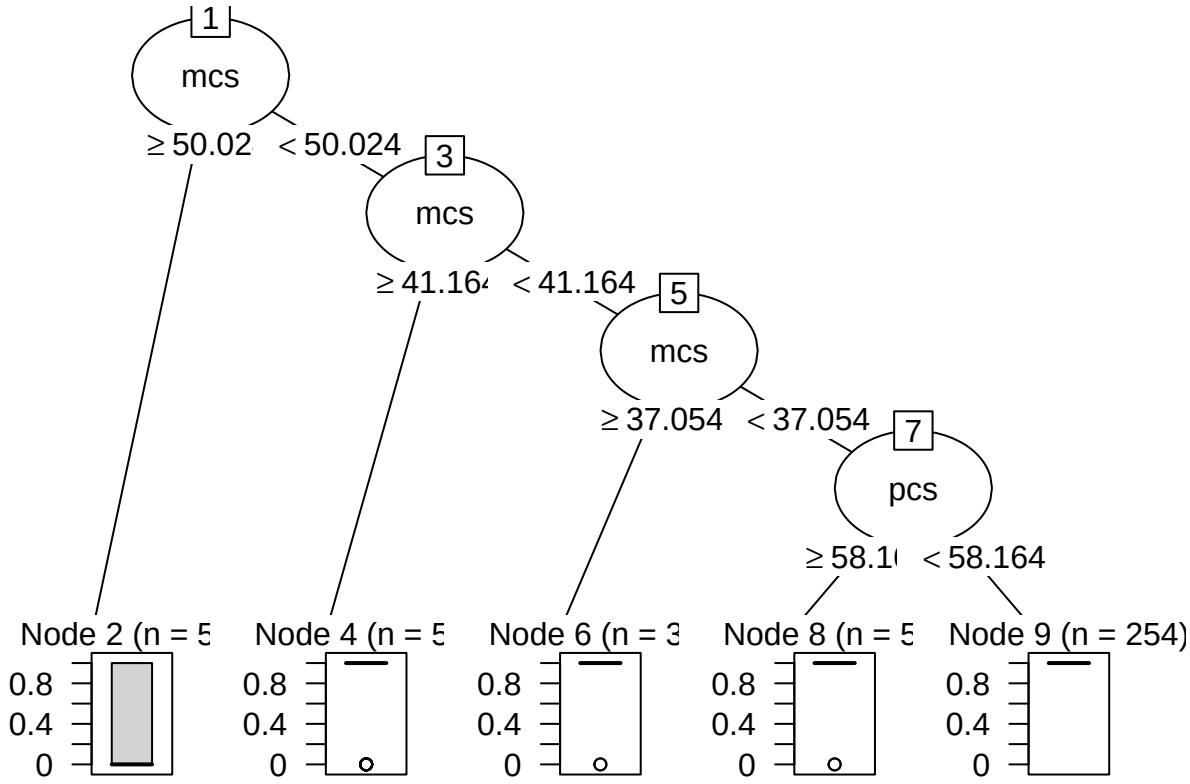
```
plot(fitallpk, main = "Conditional Inference Tree for CESD => 16")
```



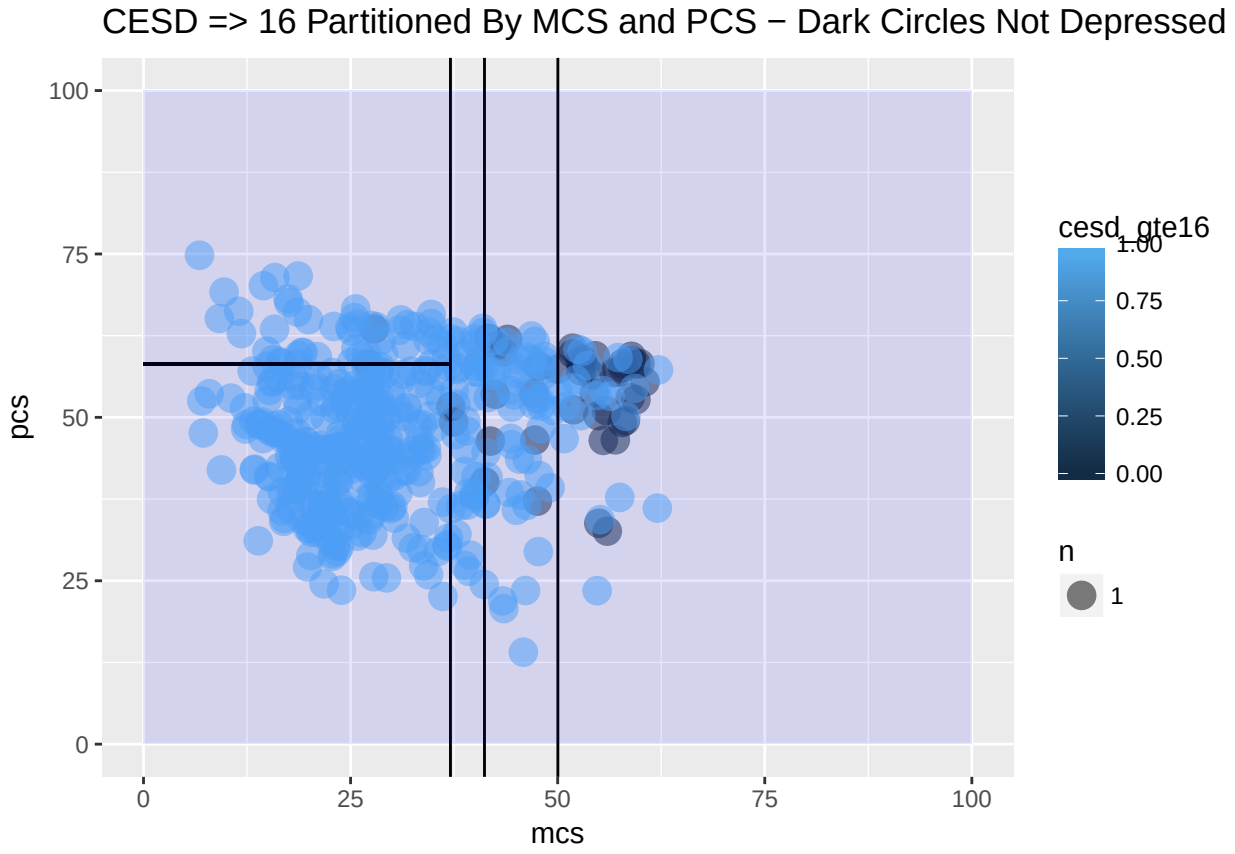

```
# Recursive partitioning of CESD on age,
# female, pss_fr, homeless, pcs, mcs
whoIsDepressed <- rpart::rpart(cesd_gte16 ~ age + female +
                                 pss_fr + homeless + pcs + mcs,
                               data = h1b,
                               control = rpart.control(cp = 0.001,
                                                       minbucket = 20))

whoIsDepressed
```

```
## n= 453
##
## node), split, n, deviance, yval
##       * denotes terminal node
##
##  1) root 453 41.328920 0.8984547
##    2) mcs>=50.02446 51 12.156860 0.3921569 *
##    3) mcs< 50.02446 402 14.440300 0.9626866
##      6) mcs>=41.16363 59  8.949153 0.8135593 *
##      7) mcs< 41.16363 343  3.953353 0.9883382
##       14) mcs>=37.05422 38  1.894737 0.9473684 *
##       15) mcs< 37.05422 305  1.986885 0.9934426
##         30) pcs>=58.16405 51  1.921569 0.9607843 *
##         31) pcs< 58.16405 254  0.000000 1.0000000 *
```

```
library(partykit)
# Plot the tree
plot(partykit::as.party(whoIsDepressed))
```




```
# EXTRA CREDIT
# Graph as partition
# using the break points shown from the
# conditional tree
ggplot(data = h1b, aes(x = mcs, y = pcs)) +
  geom_count(aes(color = cesd_gte16), alpha = 0.5) +
  geom_vline(xintercept = 50.024) +
  geom_vline(xintercept = 41.164) +
  geom_vline(xintercept = 37.054) +
  geom_segment(x = 37.054, xend = 0, y = 58.164, yend = 58.164) +
  annotate("rect", xmin = 0, xmax = 100, ymin = 0, ymax = 100, fill = "blue", alpha = 0.1) +
  ggtitle("CESD => 16 Partitioned By MCS and PCS - Dark Circles Not Depressed")
```



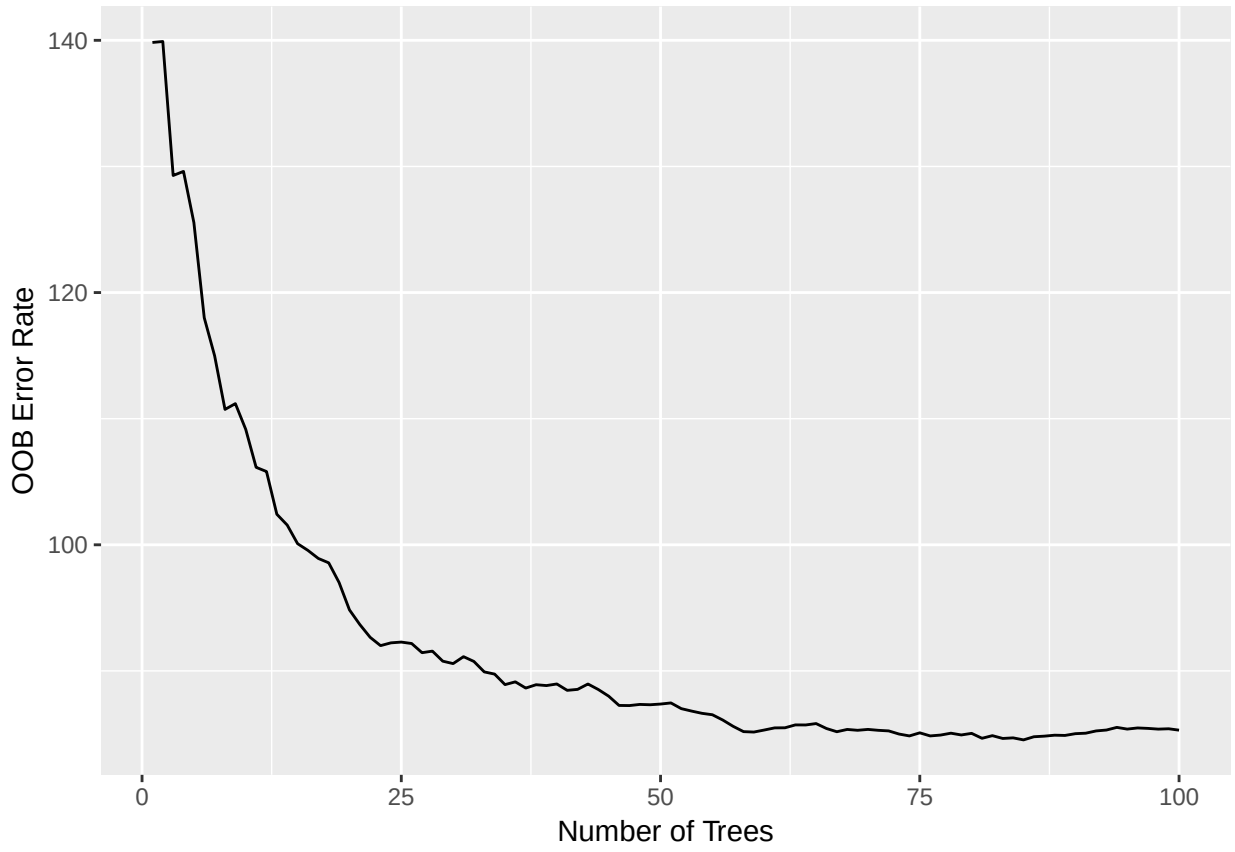
random forests approach

```
# library(RColorBrewer)
# library(plot3D)
# library(parallel)
library(randomForestSRC)
library(ggRandomForests)

h1a <- as.data.frame(h1a)
set.seed(131)
# Random Forest for the ozone dataset
fitallrf <- randomForestSRC::rfsrc(cesd ~ .,
                                   data = h1a, ntree = 100,
                                   tree.err=TRUE)
# view the results
fitallrf
```

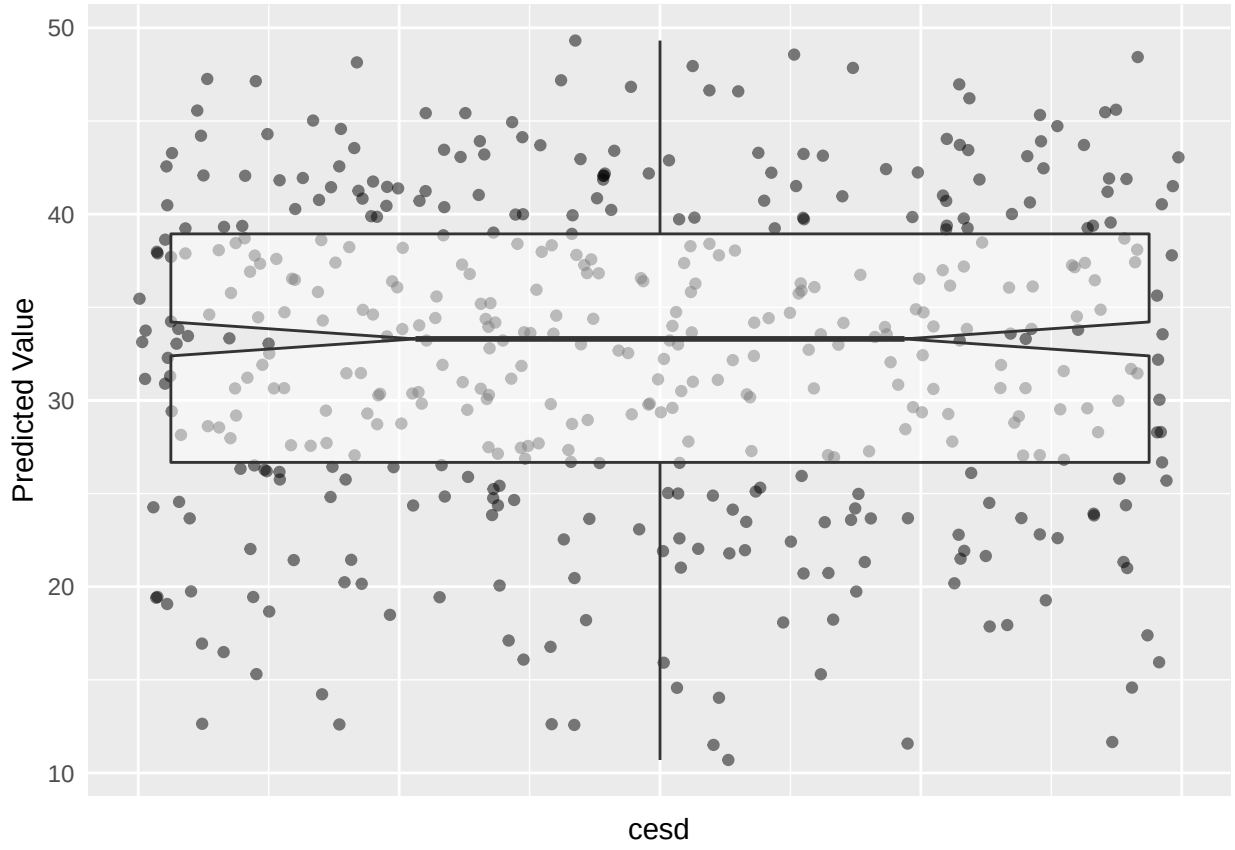
```
##                          Sample size: 453
##                      Number of trees: 100
##            Forest terminal node size: 5
##        Average no. of terminal nodes: 91.16
## No. of variables tried at each split: 2
##               Total no. of variables: 6
##                             Analysis: RF-R
##                               Family: regr
##                      Splitting rule: mse
```

```
##                 % variance explained: 45.53
##                           Error rate: 85.3
```

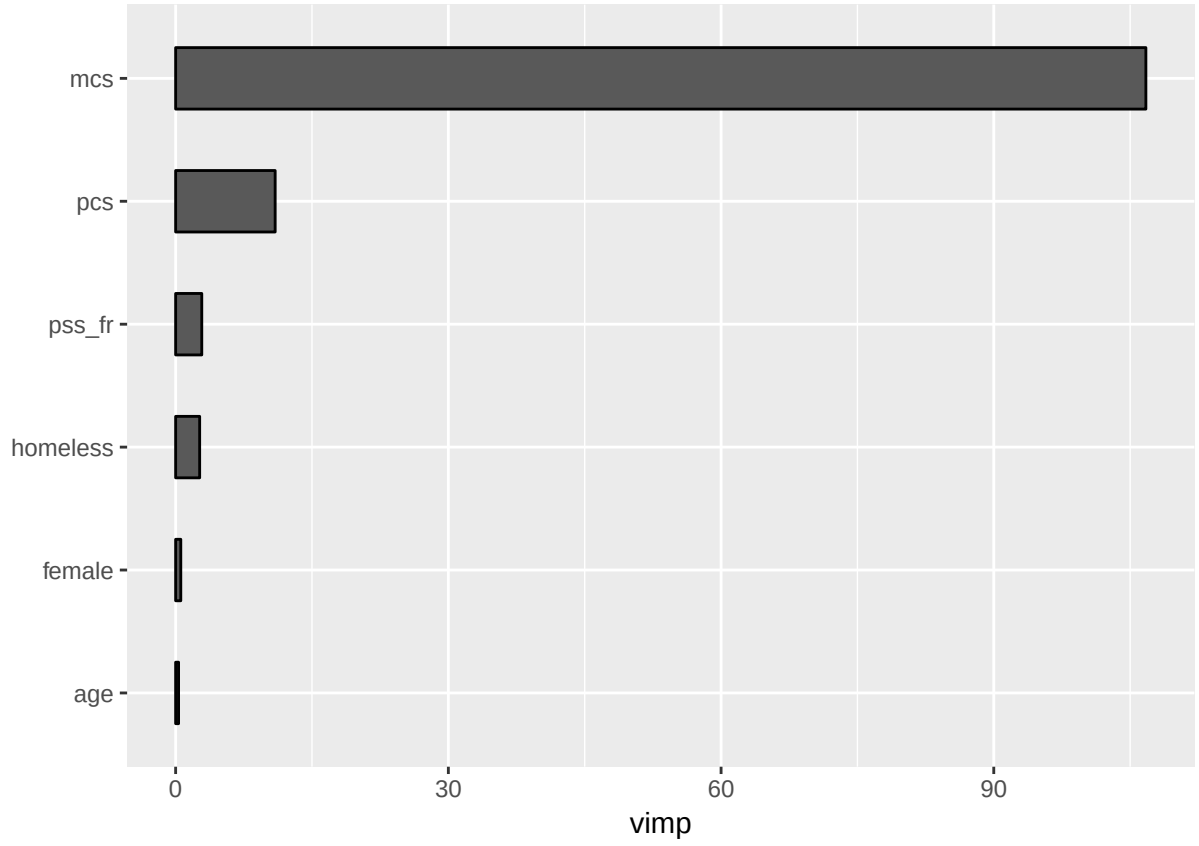
```
gg_e <- ggRandomForests::gg_error(fitallrf)
plot(gg_e)
```




```
# Plot the predicted cesd values
plot(ggRandomForests::gg_rfsrc(fitallrf), alpha = 0.5)
```




```
# Plot the VIMP rankins of independent variables
plot(ggRandomForests::gg_vimp(fitallrf))
```




```
# Select the variables
varsel_cesd <- randomForestSRC::var.select(fitallrf)
```

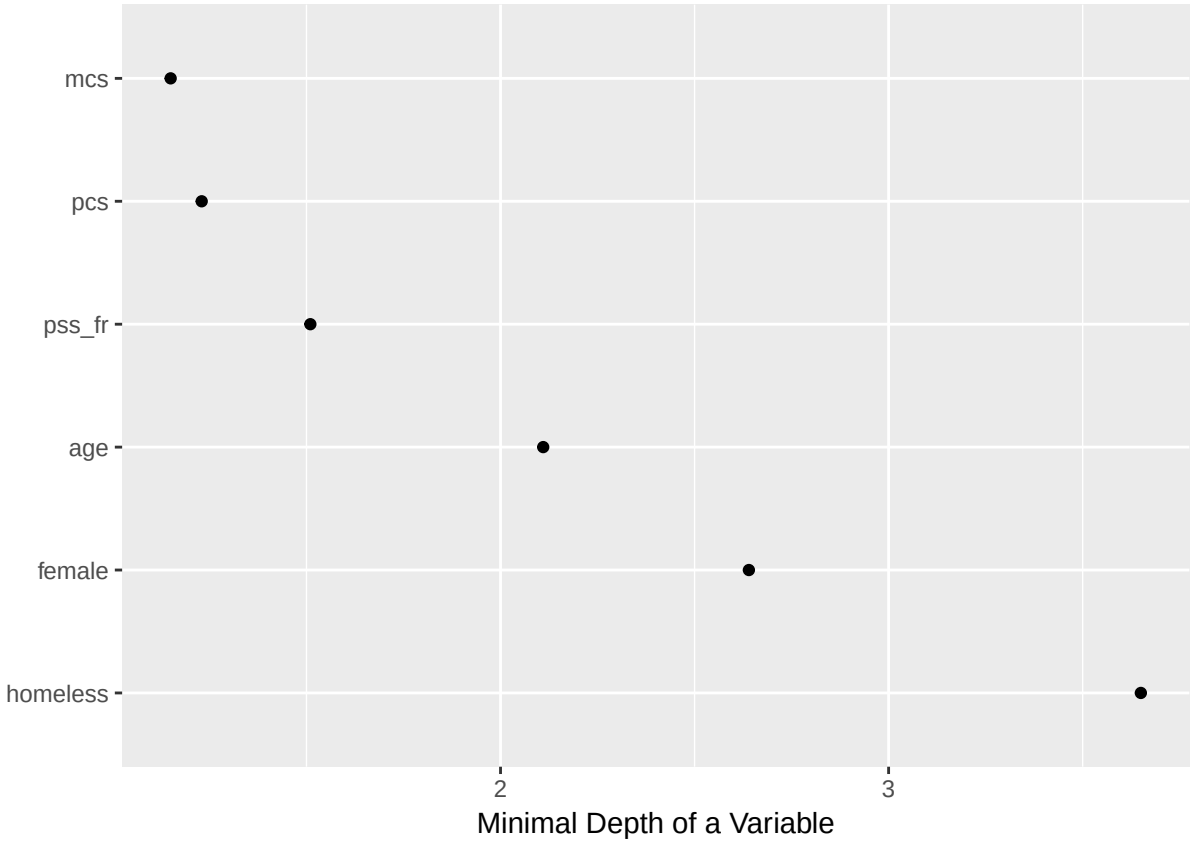
```
## minimal depth variable selection ...
##
##
## ------------------------------------------------------------
## family             : regr
## var. selection     : Minimal Depth
## conservativeness   : medium
## x-weighting used?  : TRUE
## dimension          : 6
## sample size        : 453
## ntree              : 100
## nsplit             : 0
## mtry               : 2
## nodesize           : 5
## refitted forest    : FALSE
## model size         : 6
## depth threshold    : 5.6833
## PE (true OOB)      : 85.3018
##
##
## Top variables:
##          depth vimp
## mcs       1.15   NA
```

```
## pcs       1.23   NA
## pss_fr    1.51   NA
## age       2.11   NA
## female    2.64   NA
## homeless  3.65   NA
## ------------------------------------------------------------
```

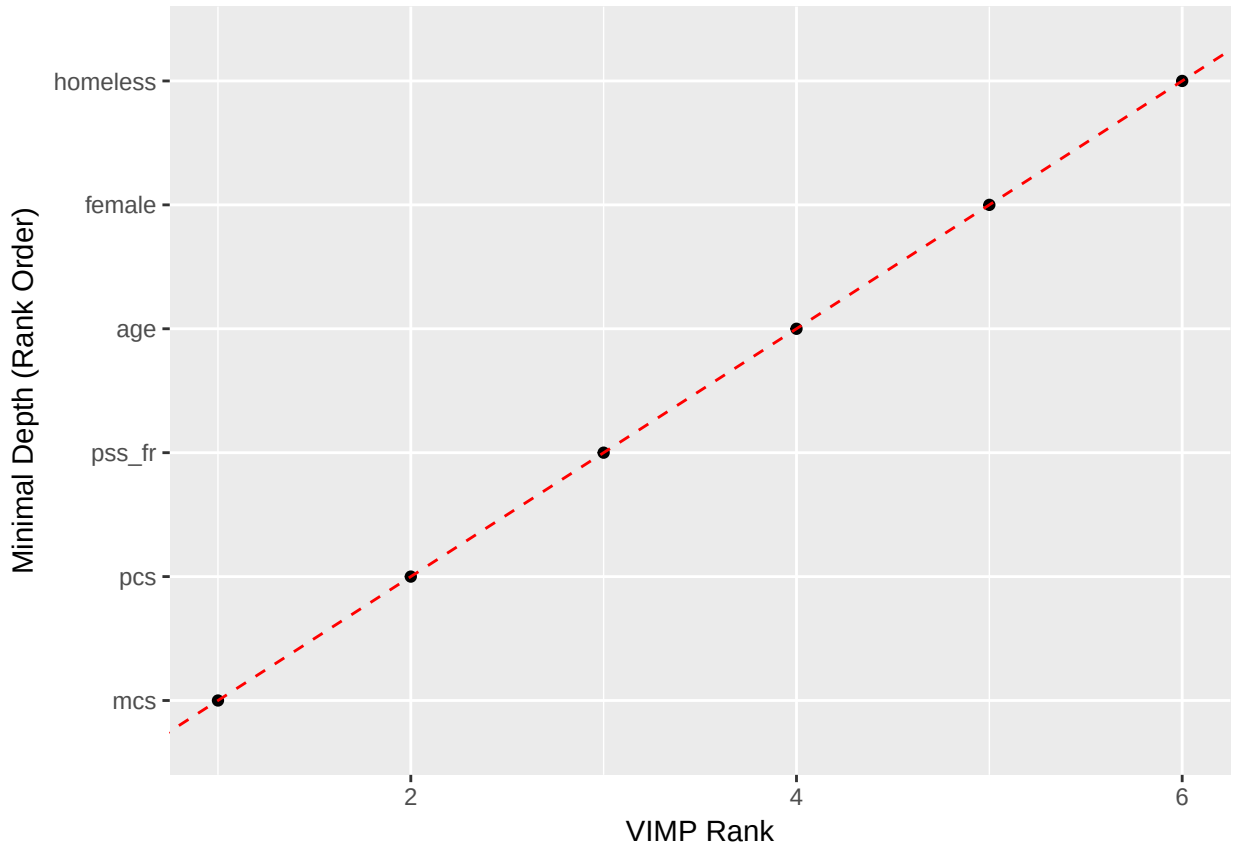
```
glimpse(varsel_cesd)
```

```
## List of 6
##  $ err.rate       : num 85.3
##  $ modelsize      : int 6
##  $ topvars        : chr [1:6] "mcs" "pcs" "pss_fr" "age" ...
##  $ varselect      :'data.frame': 6 obs. of  2 variables:
##   ..$ depth: num [1:6] 1.15 1.23 1.51 2.11 2.64 3.65
##   ..$ vimp : num [1:6] NA NA NA NA NA NA
##  $ rfsrc.refit.obj: NULL
##  $ md.obj         :List of 11
##   ..$ order                : num [1:6, 1:2] 2.11 2.64 1.51 3.65 1.23 1.15 3.57 6.23 5.37 4.77 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   ..$ count                : Named num [1:6] 0.1539 0.0816 0.1107 0.1075 0.091 ...
##   .. ..- attr(*, "names")= chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ nodes.at.depth       : num [1:10000, 1:100] 2 4 7 7 10 14 12 11 9 7 ...
##   ..$ sub.order            : NULL
##   ..$ threshold            : num 5.68
##   ..$ threshold.1se        : num 5.88
##   ..$ topvars              : chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ topvars.1se          : chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ percentile           : Named num [1:6] 0.194 0.26 0.141 0.357 0.12 ...
##   .. ..- attr(*, "names")= chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ density              : Named num [1:21] 0.0641 0.0968 0.1314 0.1307 0.1008 ...
##   .. ..- attr(*, "names")= chr [1:21] "0" "1" "2" "3" ...
##   ..$ second.order.threshold: num 10.1
```

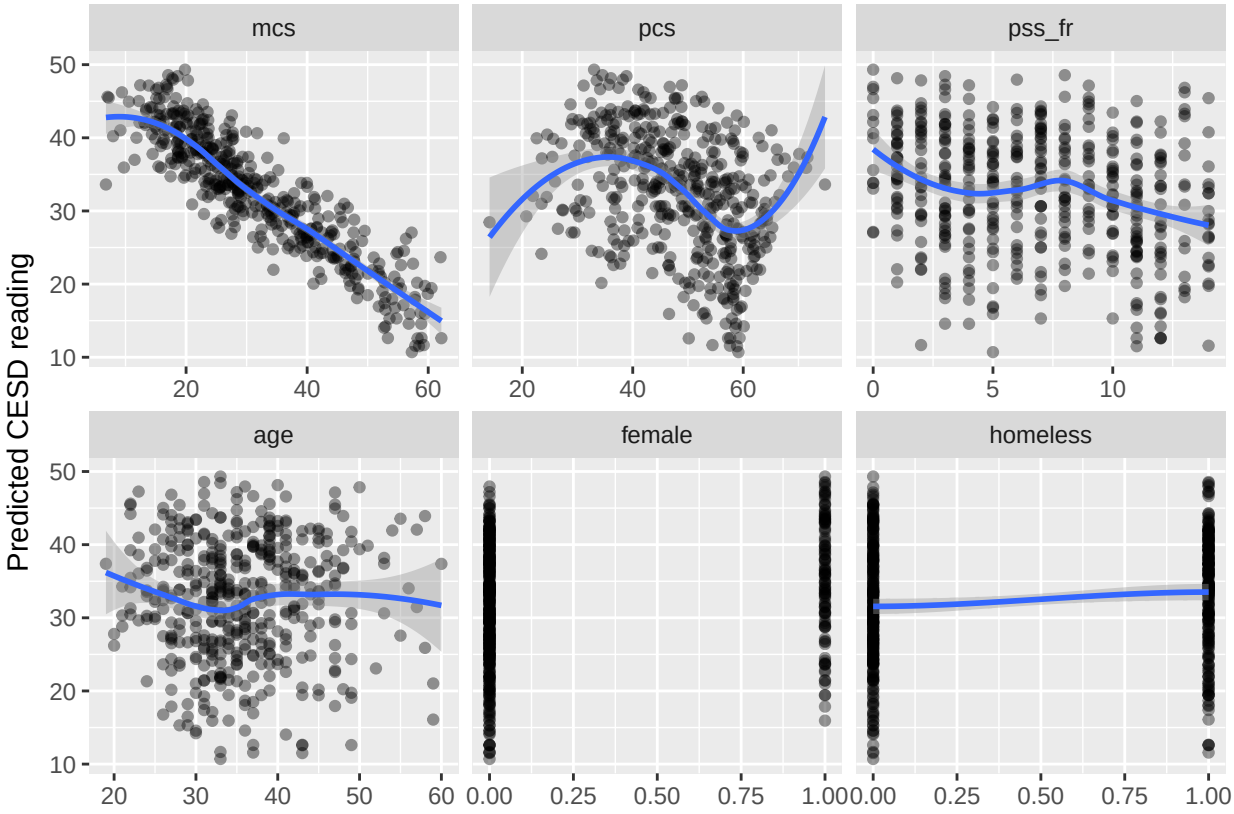
```
# Save the gg_minimal_depth object for later use
gg_md <- ggRandomForests::gg_minimal_depth(varsel_cesd)
# Plot the object
plot(gg_md)
```




```
# Plot minimal depth v VIMP
gg_mdVIMP <- ggRandomForests::gg_minimal_vimp(gg_md)
plot(gg_mdVIMP)
```




```
#Create the variable dependence object from the random forest
gg_v <- ggRandomForests::gg_variable(fitallrf)

# Use the top ranked minimal depth variables only, plotted in minimal depth rank order
xvar <- gg_md$topvars

# Plot the variable list in a single panel plot
plot(gg_v, xvar = xvar, panel = TRUE, alpha = 0.4) +
  labs(y="Predicted CESD reading", x="")
```



Use R markdown to complete your homework and show all of your code and output in your final report - Turn in a PDF of your report to Canvas. Include a link to your Github repo for Homework 7
