NRSG 741 Homework 7

Tommy Flynn

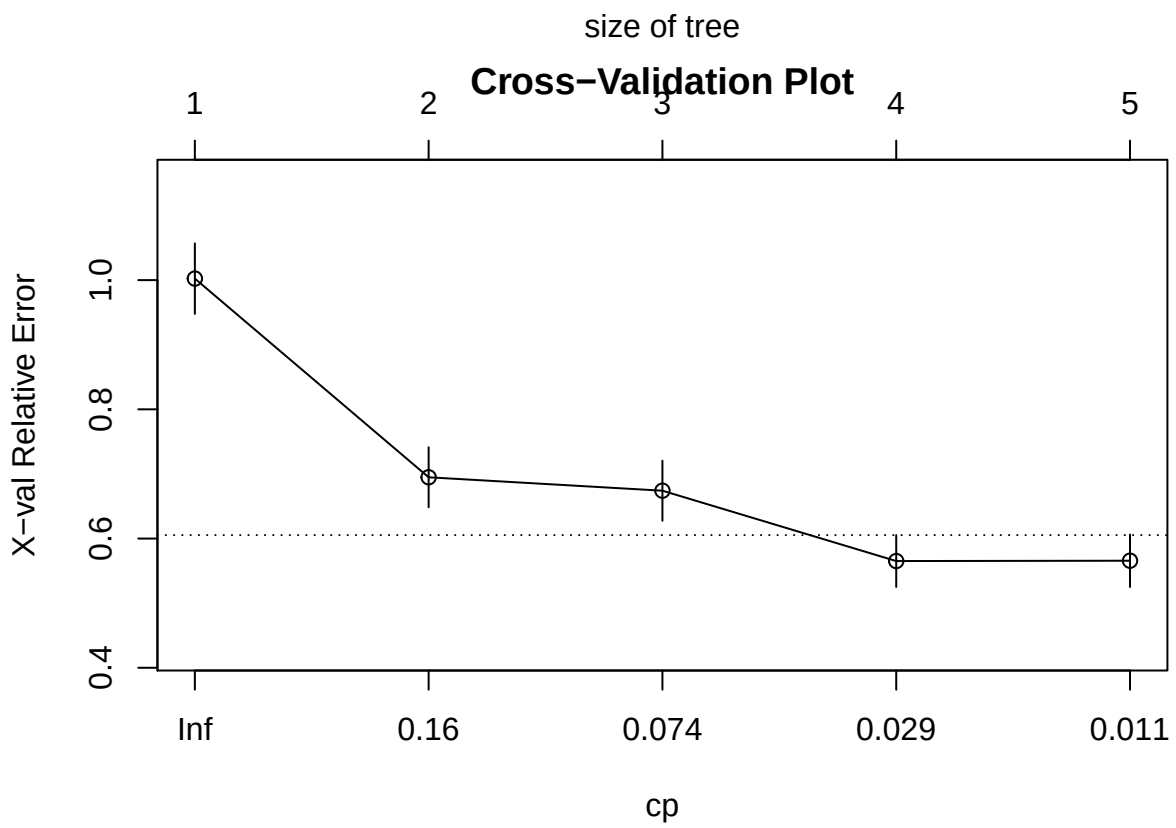
4/8/2018

Problem 1 Answer

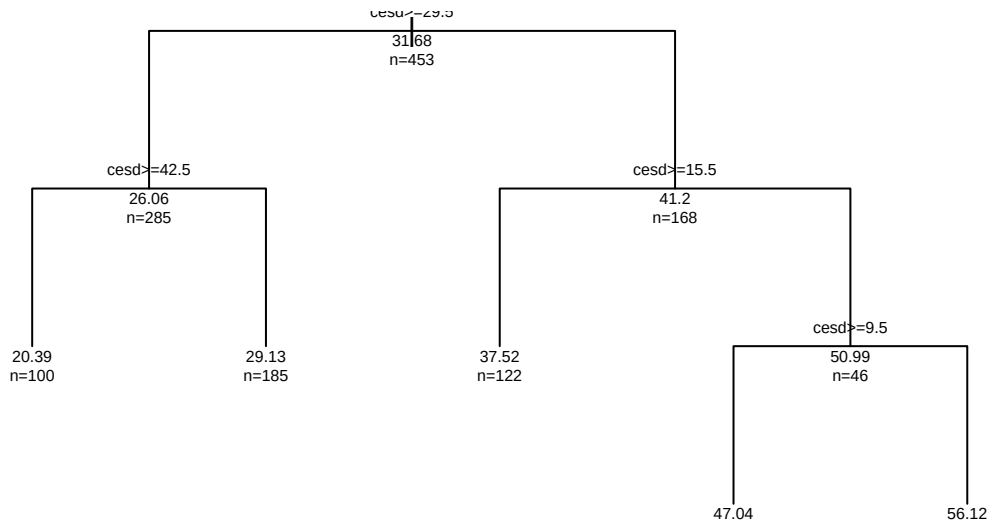
```
#read in dataset to helpdata df
helpdata <- read_spss("helpmkh.sav")

#Let's do it all in one go...
h1 <- helpdata %>%
  select(age, female, pss_fr, homeless,
         pcs, mcs, cesd) %>%
  mutate(cesd_16 = as.numeric(cesd >= 16), mcs_45 = as.numeric(mcs < 45))

# fit a regression tree model to the cesd as the outcome
# and using the mcs as the only predictor
fitmcs <- rpart(mcs ~ cesd, data = h1)
printcp(fitmcs) # Display the results
```

```
##
## Regression tree:
## rpart(formula = mcs ~ cesd, data = h1)
##
## Variables actually used in tree construction:
## [1] cesd
##
## Root node error: 74512/453 = 164.48
##
## n= 453
##
##         CP nsplit rel error  xerror     xstd
## 1 0.325298      0   1.00000 1.00238 0.054623
## 2 0.081349      1   0.67470 0.69486 0.046441
## 3 0.066496      2   0.59335 0.67409 0.046520
## 4 0.012496      3   0.52686 0.56518 0.040140
## 5 0.010000      4   0.51436 0.56571 0.040614
```

```
plotcp(fitmcs, main = "Cross-Validation Plot") # Visualize cross-validation results
```

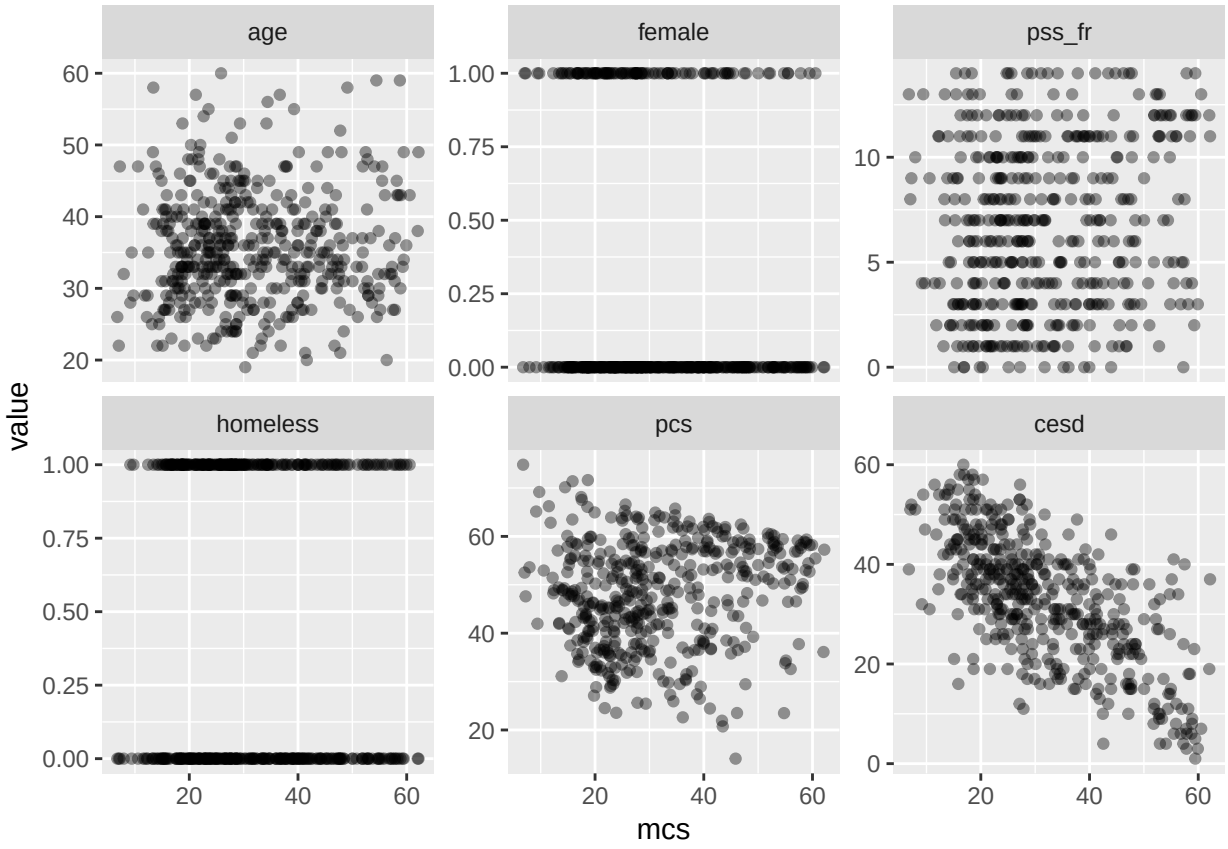



```
# plot tree
plot(fitmcs, uniform = TRUE, compress = FALSE, main = "MCS Regression Tree")
text(fitmcs, use.n = TRUE, all = TRUE, cex = 0.5)
```



Problem 2 Answer

```
# all vars except the dichotomous cesd_gte16 and mcs_lt45
h1a <- h1 %>%
  select(1:7)

# Melt the other variables down and link to cesd
h1m <- melt(h1a, id.vars = "mcs")

# Plot panels for each covariate
ggplot(h1m, aes(x=mcs, y=value)) +
  geom_point(alpha=0.4)+
  scale_color_brewer(palette="Set2")+
  facet_wrap(~variable, scales="free_y", ncol=3)
```



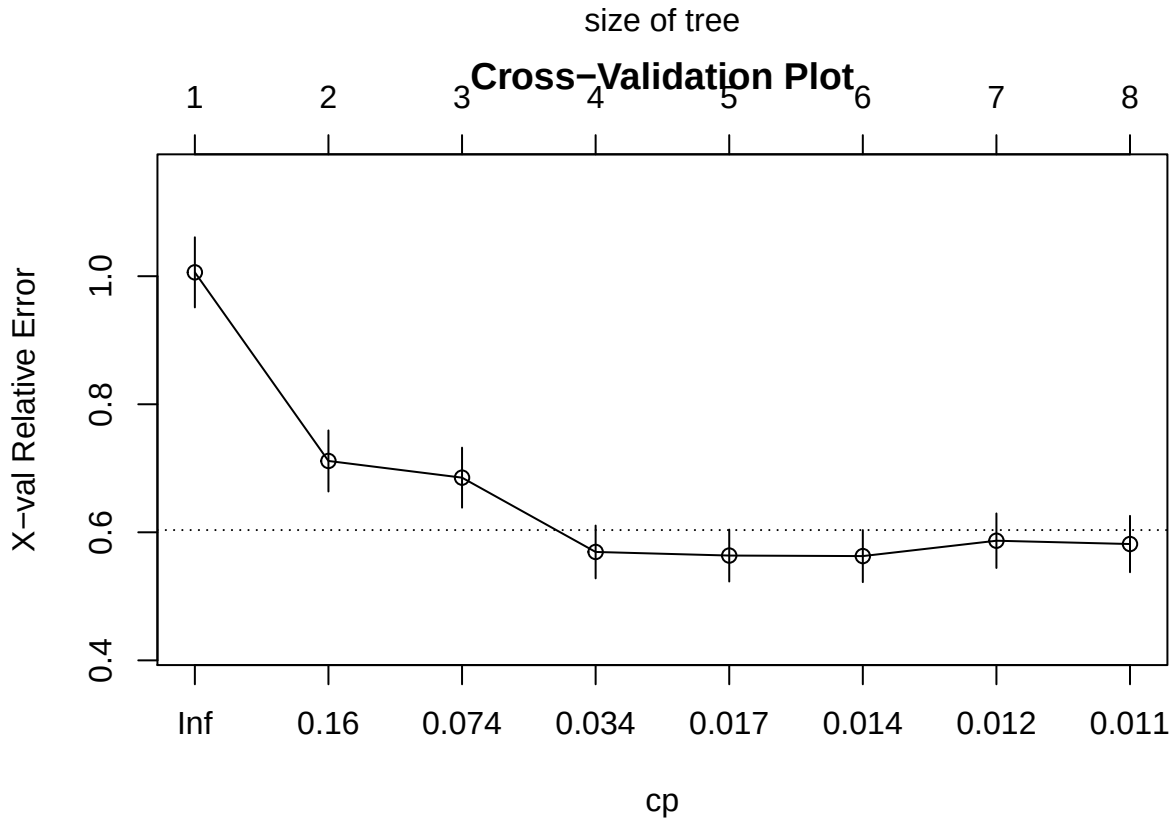
Problem 3 Answer

```
#regression tree of mcs with all other variables
fitall <- rpart(mcs ~ ., data = h1a)
#print the results
printcp(fitall)
```

```
##
## Regression tree:
```

```
## rpart(formula = mcs ~ ., data = h1a)
##
## Variables actually used in tree construction:
## [1] cesd pcs
##
## Root node error: 74512/453 = 164.48
##
## n= 453
##
##         CP nsplit rel error  xerror     xstd
## 1 0.325298      0   1.00000 1.00606 0.054768
## 2 0.081349      1   0.67470 0.71143 0.047695
## 3 0.066496      2   0.59335 0.68537 0.046866
## 4 0.017717      3   0.52686 0.56930 0.041247
## 5 0.015767      4   0.50914 0.56368 0.040577
## 6 0.012496      5   0.49337 0.56284 0.040680
## 7 0.012258      6   0.48088 0.58685 0.042596
## 8 0.010000      7   0.46862 0.58167 0.044025
```

```r
# Visualize cross-validation results
plotcp(fitall, main="Cross-Validation Plot")
```




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

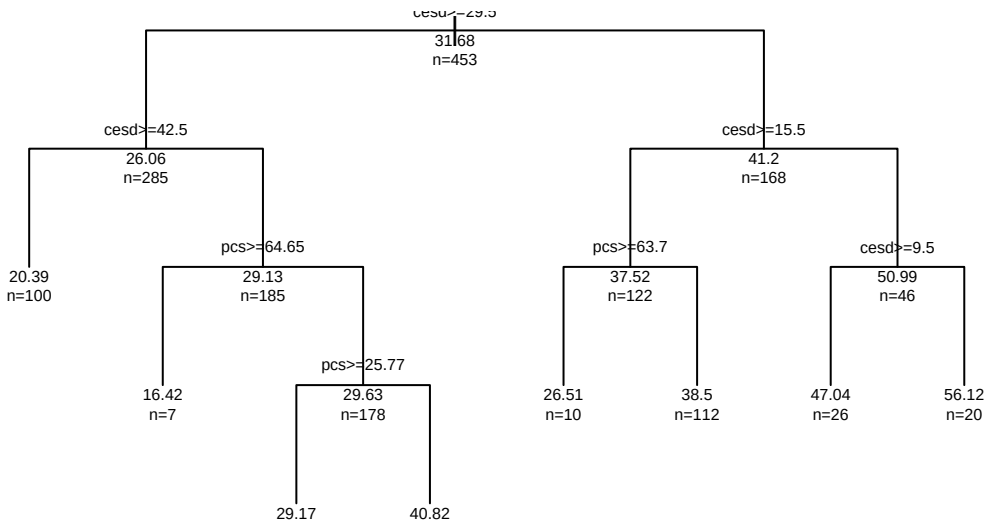
```
## Call:
## rpart(formula = mcs ~ ., data = h1a)
##   n= 453
##
```

```
##           CP nsplit rel error    xerror       xstd
## 1 0.32529813      0 1.0000000 1.0060646 0.05476772
## 2 0.08134904      1 0.6747019 0.7114336 0.04769476
## 3 0.06649553      2 0.5933528 0.6853686 0.04686590
## 4 0.01771736      3 0.5268573 0.5693023 0.04124748
## 5 0.01576737      4 0.5091399 0.5636840 0.04057714
## 6 0.01249609      5 0.4933726 0.5628367 0.04067991
## 7 0.01225792      6 0.4808765 0.5868503 0.04259597
## 8 0.01000000      7 0.4686186 0.5816730 0.04402462
##
## Variable importance
##   cesd    pcs    age pss_fr
##     83     14      1      1
##
## Node number 1: 453 observations,    complexity param=0.3252981
##   mean=31.67668, MSE=164.4847
##   left son=2 (285 obs) right son=3 (168 obs)
##   Primary splits:
##       cesd   < 29.5     to the right, improve=0.325298100, (0 missing)
##       pcs    < 49.46132 to the left,  improve=0.064711670, (0 missing)
##       pss_fr < 10.5     to the left,  improve=0.039318510, (0 missing)
##       female < 0.5      to the right, improve=0.014091560, (0 missing)
##       age    < 42.5     to the left,  improve=0.005473724, (0 missing)
##   Surrogate splits:
##       pcs < 56.34591 to the left,  agree=0.669, adj=0.107, (0 split)
##       age < 57.5     to the left,  agree=0.631, adj=0.006, (0 split)
##
## Node number 2: 285 observations,    complexity param=0.06649553
##   mean=26.06057, MSE=100.1894
##   left son=4 (100 obs) right son=5 (185 obs)
##   Primary splits:
##       cesd   < 42.5     to the right, improve=0.173520000, (0 missing)
##       pcs    < 24.47511 to the right, improve=0.057879990, (0 missing)
##       pss_fr < 10.5     to the left,  improve=0.015219690, (0 missing)
##       age    < 22.5     to the right, improve=0.005742931, (0 missing)
##       female < 0.5      to the right, improve=0.001903900, (0 missing)
##   Surrogate splits:
##       pss_fr < 0.5      to the left,  agree=0.660, adj=0.03, (0 split)
##       pcs    < 68.64778 to the right, agree=0.653, adj=0.01, (0 split)
##
## Node number 3: 168 observations,    complexity param=0.08134904
##   mean=41.20401, MSE=129.2805
##   left son=6 (122 obs) right son=7 (46 obs)
##   Primary splits:
##       cesd   < 15.5     to the right, improve=0.279083400, (0 missing)
##       pcs    < 62.7532  to the right, improve=0.113215200, (0 missing)
##       pss_fr < 10.5     to the left,  improve=0.053187210, (0 missing)
##       age    < 48.5     to the left,  improve=0.036737610, (0 missing)
##       female < 0.5      to the right, improve=0.007177787, (0 missing)
##   Surrogate splits:
##       age < 58.5     to the left,  agree=0.738, adj=0.043, (0 split)
##
## Node number 4: 100 observations
##   mean=20.38941, MSE=43.95751
```

```
##
## Node number 5: 185 observations,    complexity param=0.01576737
##   mean=29.12606, MSE=103.8029
##   left son=10 (7 obs) right son=11 (178 obs)
##   Primary splits:
##       pcs    < 64.65134 to the right, improve=0.061178900, (0 missing)
##       age    < 22.5     to the right, improve=0.031248410, (0 missing)
##       cesd   < 37.5     to the right, improve=0.020833690, (0 missing)
##       pss_fr < 10.5     to the left,  improve=0.015175680, (0 missing)
##       female < 0.5      to the left,  improve=0.004355548, (0 missing)
##
## Node number 6: 122 observations,    complexity param=0.01771736
##   mean=37.51566, MSE=103.6988
##   left son=12 (10 obs) right son=13 (112 obs)
##   Primary splits:
##       pcs    < 63.69606 to the right, improve=0.10434930, (0 missing)
##       age    < 47.5     to the left,  improve=0.02626159, (0 missing)
##       cesd   < 24.5     to the right, improve=0.02348926, (0 missing)
##       female < 0.5      to the right, improve=0.02256241, (0 missing)
##       pss_fr < 2.5      to the right, improve=0.01295167, (0 missing)
##
## Node number 7: 46 observations,    complexity param=0.01249609
##   mean=50.98616, MSE=65.35702
##   left son=14 (26 obs) right son=15 (20 obs)
##   Primary splits:
##       cesd     < 9.5      to the right, improve=0.30970460, (0 missing)
##       pcs      < 59.57495 to the right, improve=0.16249370, (0 missing)
##       pss_fr   < 11.5     to the left,  improve=0.13099300, (0 missing)
##       age      < 40       to the left,  improve=0.06604375, (0 missing)
##       homeless < 0.5      to the left,  improve=0.00873942, (0 missing)
##   Surrogate splits:
##       pss_fr   < 11.5     to the left,  agree=0.674, adj=0.25, (0 split)
##       pcs      < 54.5861  to the left,  agree=0.652, adj=0.20, (0 split)
##       age      < 46       to the left,  agree=0.609, adj=0.10, (0 split)
##       homeless < 0.5      to the left,  agree=0.609, adj=0.10, (0 split)
##
## Node number 10: 7 observations
##   mean=16.41837, MSE=35.31025
##
## Node number 11: 178 observations,    complexity param=0.01225792
##   mean=29.6258, MSE=99.89614
##   left son=22 (171 obs) right son=23 (7 obs)
##   Primary splits:
##       pcs      < 25.77119 to the right, improve=0.051365510, (0 missing)
##       age      < 22.5     to the right, improve=0.029936490, (0 missing)
##       pss_fr   < 10.5     to the left,  improve=0.022699840, (0 missing)
##       cesd     < 37.5     to the right, improve=0.020642200, (0 missing)
##       homeless < 0.5      to the right, improve=0.002448012, (0 missing)
##
## Node number 12: 10 observations
##   mean=26.50685, MSE=30.97799
##
## Node number 13: 112 observations
##   mean=38.49859, MSE=98.40465
```

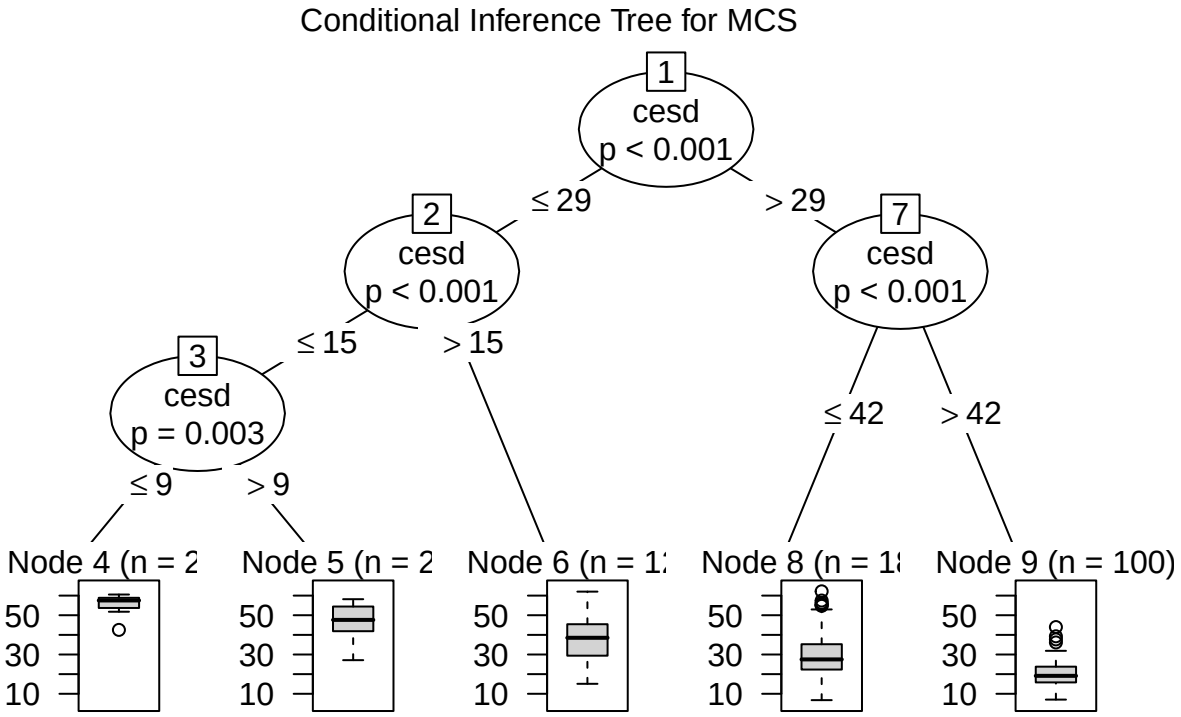
```
##
## Node number 14: 26 observations
##   mean=47.04024, MSE=67.29195
##
## Node number 15: 20 observations
##   mean=56.11586, MSE=16.28645
##
## Node number 22: 171 observations
##   mean=29.16748, MSE=95.51594
##
## Node number 23: 7 observations
##   mean=40.8217, MSE=76.41866
```

```
#regression tree for mcs from all other variables
plot(fitall, uniform = TRUE, compress = FALSE, main = "Regression Tree for MCS Scores from HELP")
text(fitall, use.n = TRUE, all = TRUE, cex = 0.5)
```



Problem 4 Answer

```
fitallp <- ctree(mcs ~ ., data = h1a)
plot(fitallp, main = "Conditional Inference Tree for MCS")
```



Problem 5 Answer

```
glm1 <- glm(mcs_45 ~ age + female + pss_fr + homeless +
              pcs + cesd, data = h1)
summary(glm1)
```

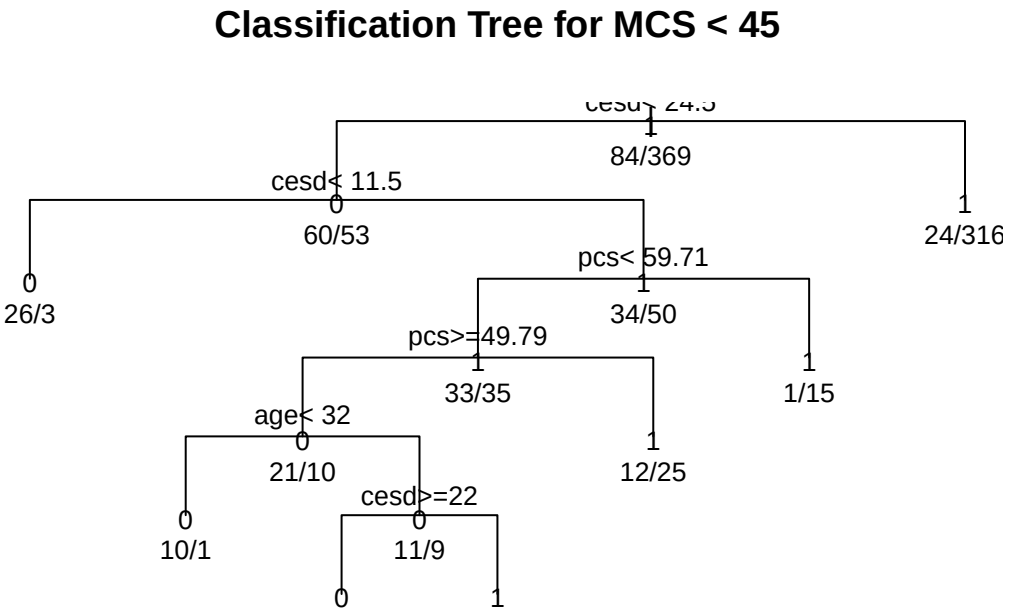
```
##
## Call:
## glm(formula = mcs_45 ~ age + female + pss_fr + homeless + pcs +
##     cesd, data = h1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.96035  -0.10332   0.08078   0.21806   0.62498
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.3611168  0.1386939   2.604  0.00953 **
## age         -0.0023080  0.0021130  -1.092  0.27529
## female       0.0202380  0.0382212   0.529  0.59672
## pss_fr      -0.0036606  0.0040882  -0.895  0.37104
## homeless     0.0172706  0.0323939   0.533  0.59420
## pcs          0.0005446  0.0015809   0.344  0.73064
## cesd         0.0158725  0.0013519  11.741  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for gaussian family taken to be 0.1114291)
##
```

```
##     Null deviance: 68.424  on 452  degrees of freedom
## Residual deviance: 49.697  on 446  degrees of freedom
## AIC: 300.46
##
## Number of Fisher Scoring iterations: 2
```

This model is similar to the model for CESD, although PCS is not significant in this model.

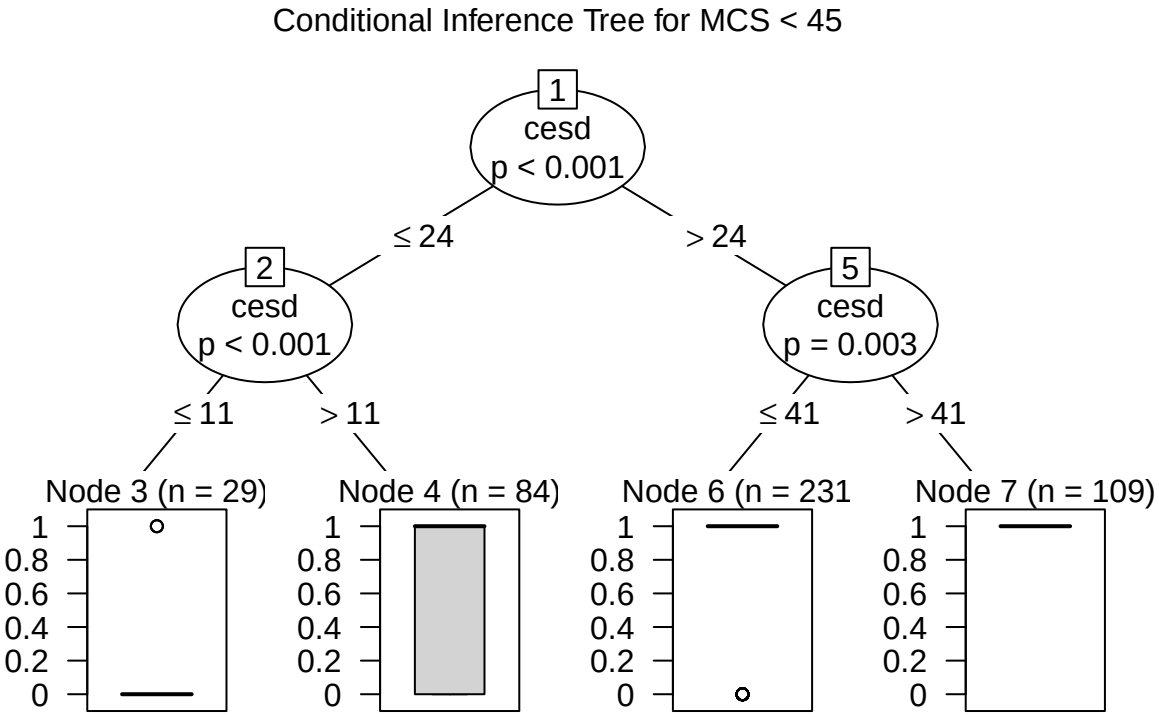
Problem 6 Answer

```
fitk <- rpart(mcs_45 ~ age + female + pss_fr +
                       homeless + pcs + cesd,
                     method = "class", data = h1)
#printcp(fitk)
#plotcp(fitk)
#summary(fitk)
plot(fitk, uniform = TRUE, main = "Classification Tree for MCS < 45")
text(fitk, use.n = TRUE, all = TRUE, cex = 0.8)
```



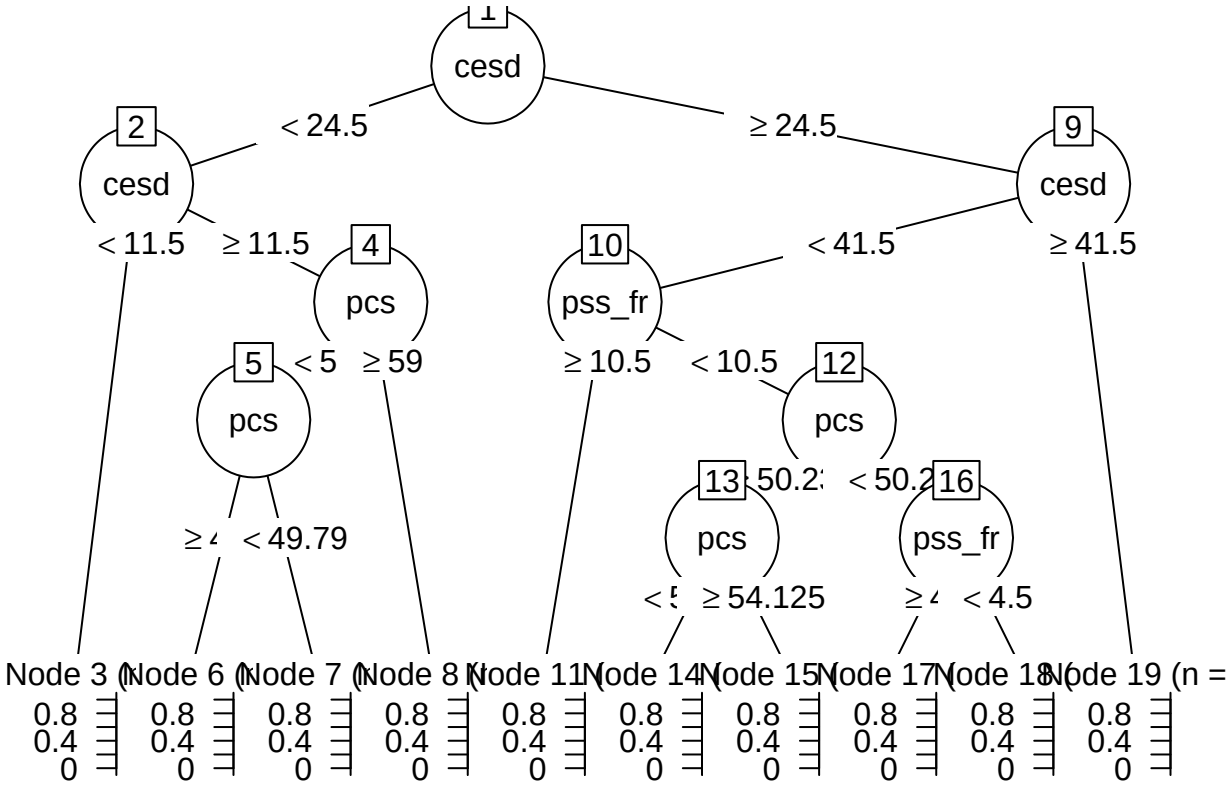
Problem 7 Answer

```
fitallpk <- party::ctree(mcs_45 ~ age + female + pss_fr +
                           homeless + pcs + cesd, data = h1)
plot(fitallpk, main = "Conditional Inference Tree for MCS < 45")
```



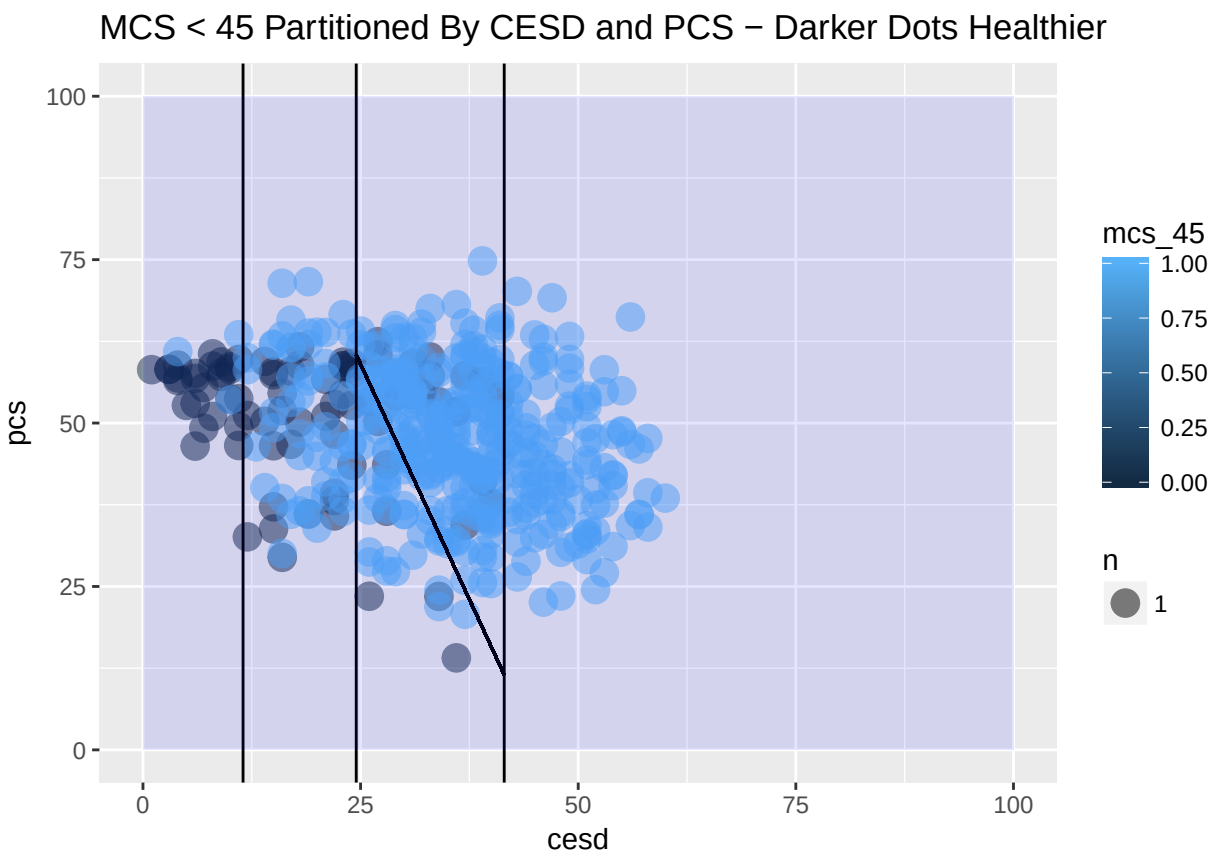
Problem 8 Answer

```
whoNeedsTherapy <- rpart::rpart(mcs_45 ~ age + female + pss_fr + homeless + pcs + cesd, data = h1,
                                control = rpart.control(cp = 0.001, minbucket = 20))
plot(as.party(whoNeedsTherapy))
```



Extra Credit Answer

```
ggplot(data = h1, aes(x = cesd, y = pcs)) +
  geom_count(aes(color = mcs_45), alpha = 0.5) +
  geom_vline(xintercept = 24.5) +
  geom_vline(xintercept = 41.5) +
  geom_vline(xintercept = 11.5) +
  geom_segment(x = 24.5, xend = 41.5, y = 60.442, yend = 11.5) +
  annotate("rect", xmin = 0, xmax = 100, ymin = 0, ymax = 100, fill = "blue", alpha = 0.1) +
  ggtitle("MCS < 45 Partitioned By CESD and PCS - Darker Dots Healthier")
```

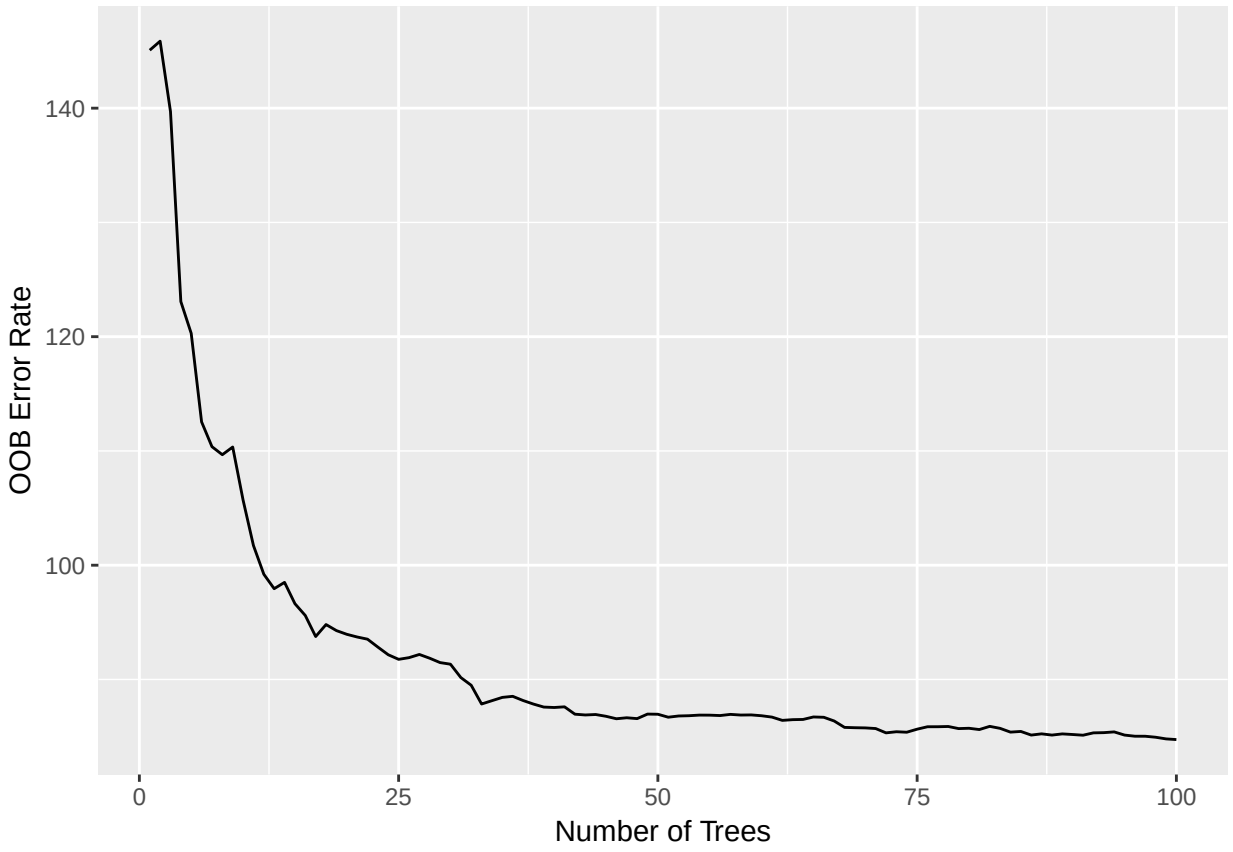


Problem 9 Answer

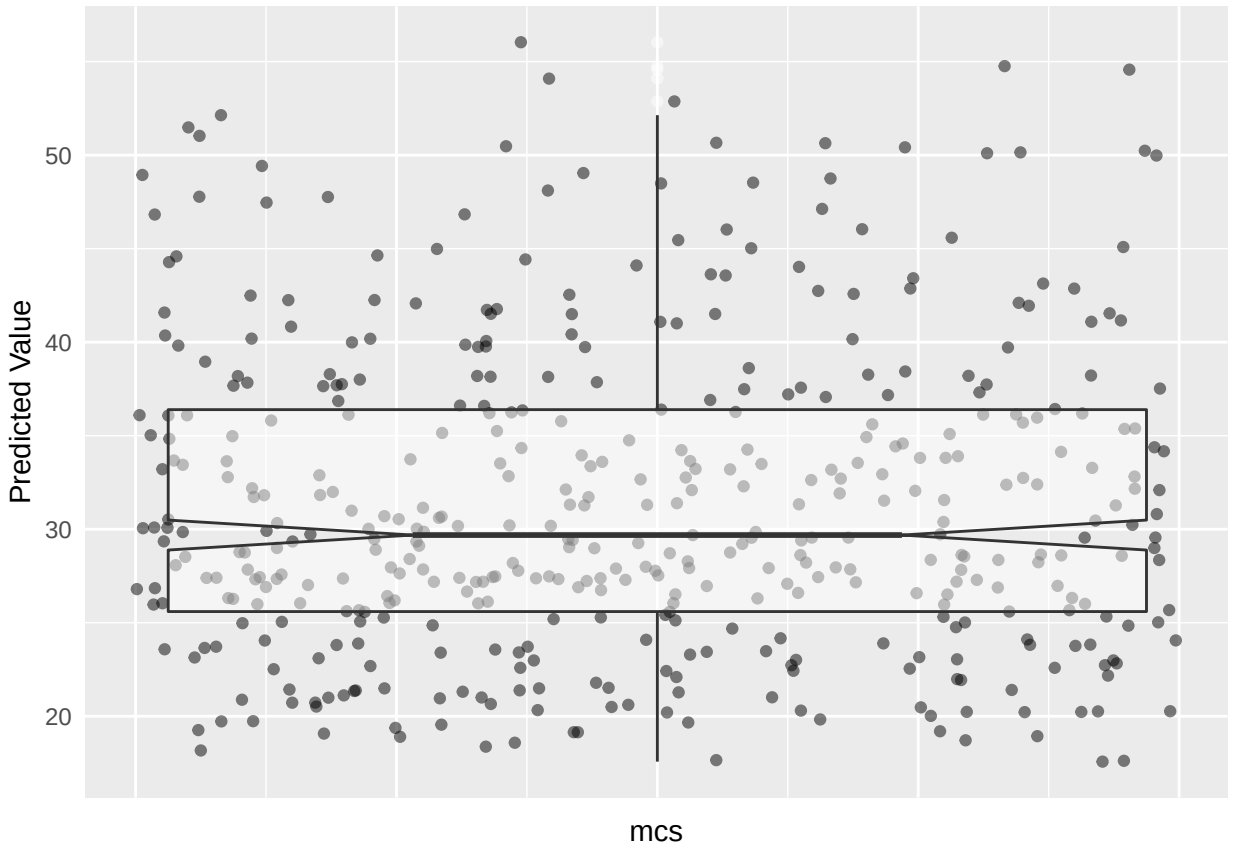
```
h1 <- as.data.frame(h1)
set.seed(131)
# Random Forest for the h1 dataset
fitallrf <- rfsrc(mcs ~ age + female + pss_fr + homeless + pcs + cesd,
                                  data = h1, ntree = 100, tree.err=TRUE)
# view the results
fitallrf
```

```
##                          Sample size: 453
##                      Number of trees: 100
##            Forest terminal node size: 5
##        Average no. of terminal nodes: 90.85
## No. of variables tried at each split: 2
##               Total no. of variables: 6
##                             Analysis: RF-R
##                               Family: regr
##                       Splitting rule: mse
##                % variance explained: 48.6
##                            Error rate: 84.74
```

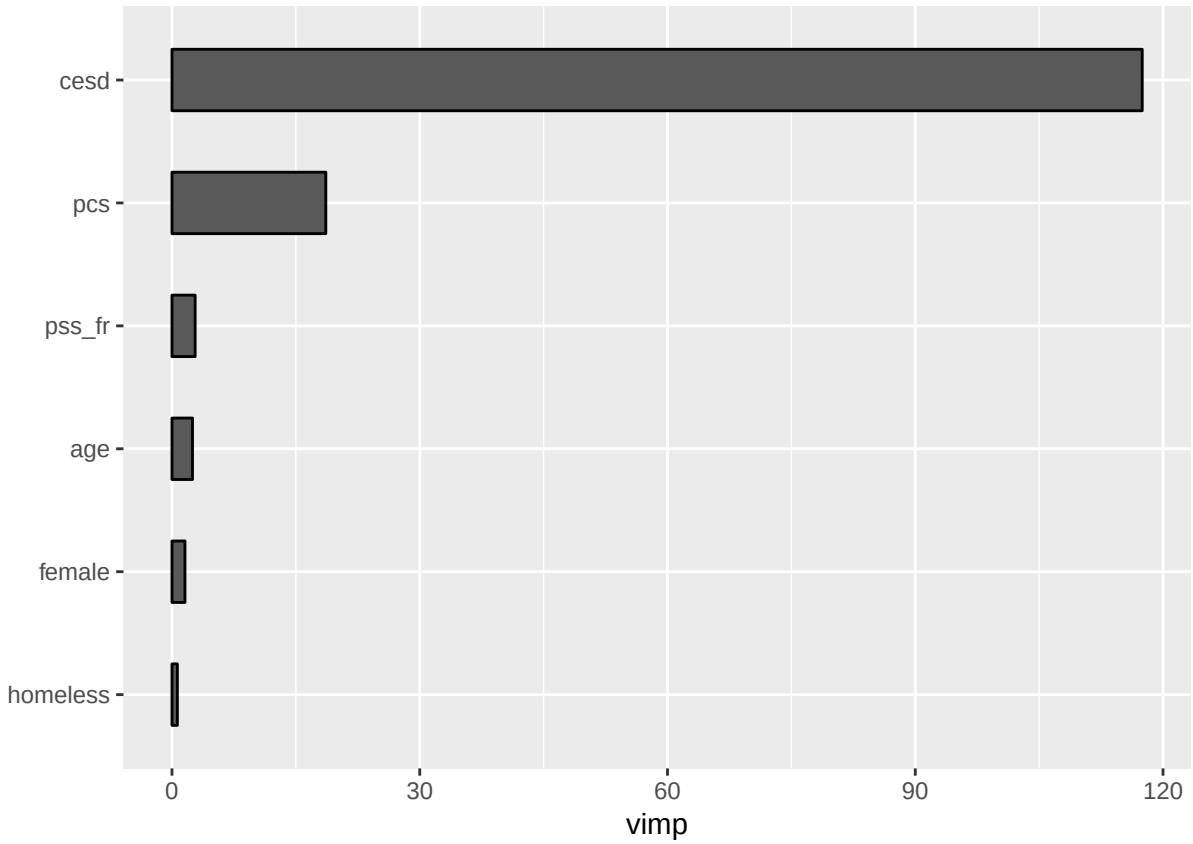
```
gg_e <- gg_error(fitallrf)
plot(gg_e)
```




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




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




```
# Select the variables
varsel_mcs <- var.select(fitallrf)
```

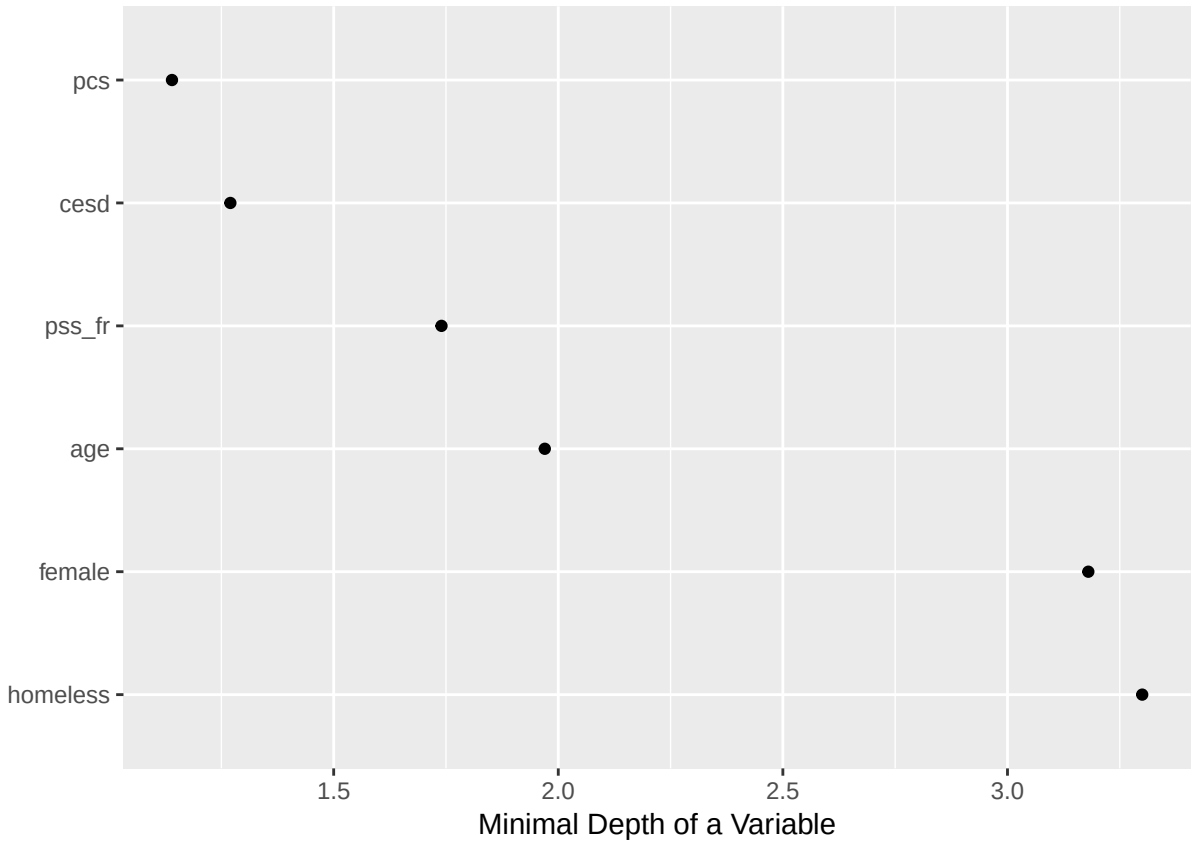
```
## minimal depth variable selection ...
##
##
## ----------------------------------------------------------
## family             : regr
## var. selection     : Minimal Depth
## conservativeness   : medium
## x-weighting used?  : TRUE
## dimension          : 6
## sample size        : 453
## ntree              : 100
## nsplit             : 0
## mtry               : 2
## nodesize           : 5
## refitted forest    : FALSE
## model size         : 6
## depth threshold    : 5.9024
## PE (true OOB)      : 84.7368
##
##
## Top variables:
##          depth vimp
## pcs       1.14   NA
## cesd      1.27   NA
```

```
## pss_fr    1.74   NA
## age       1.97   NA
## female    3.18   NA
## homeless  3.30   NA
## ----------------------------------------------------------
```

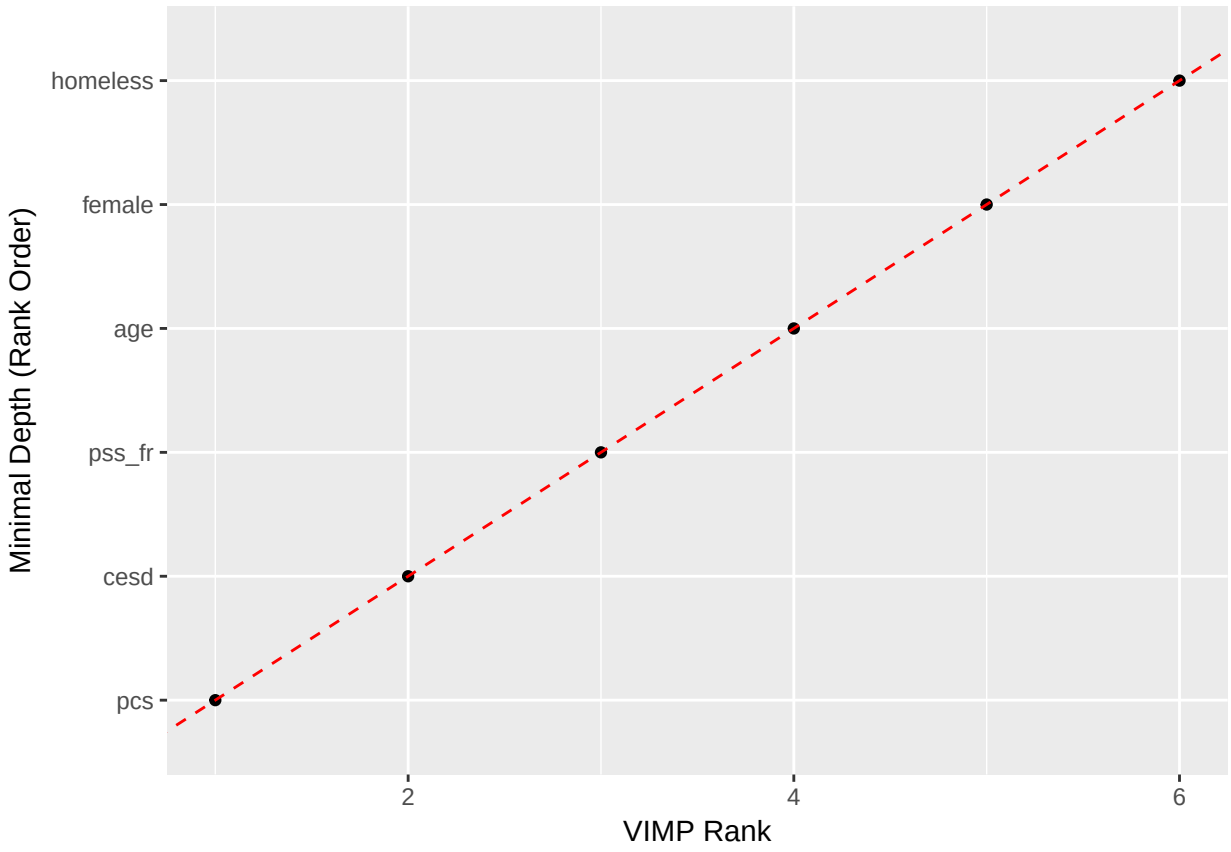
```
glimpse(varsel_mcs)
```

```
## List of 6
##  $ err.rate       : num 84.7
##  $ modelsize      : int 6
##  $ topvars        : chr [1:6] "pcs" "cesd" "pss_fr" "age" ...
##  $ varselect      :'data.frame': 6 obs. of  2 variables:
##   ..$ depth: num [1:6] 1.14 1.27 1.74 1.97 3.18 3.3
##   ..$ 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] 1.97 3.18 1.74 3.3 1.14 1.27 4.04 5.38 5.76 4.64 ...
##   .. ..- attr(*, "dimnames")=List of 2
##   ..$ count                 : Named num [1:6] 0.1396 0.0918 0.1192 0.0993 0.092 ...
##   .. ..- attr(*, "names")= chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ nodes.at.depth        : num [1:10000, 1:100] 2 4 5 9 10 13 13 13 7 3 ...
##   ..$ sub.order             : NULL
##   ..$ threshold             : num 5.9
##   ..$ threshold.1se         : num 6.1
##   ..$ topvars               : chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ topvars.1se           : chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ percentile            : Named num [1:6] 0.172 0.299 0.161 0.303 0.104 ...
##   .. ..- attr(*, "names")= chr [1:6] "age" "female" "pss_fr" "homeless" ...
##   ..$ density               : Named num [1:23] 0.0612 0.0906 0.1222 0.1232 0.0981 ...
##   .. ..- attr(*, "names")= chr [1:23] "0" "1" "2" "3" ...
##   ..$ second.order.threshold: num 10.4
```

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

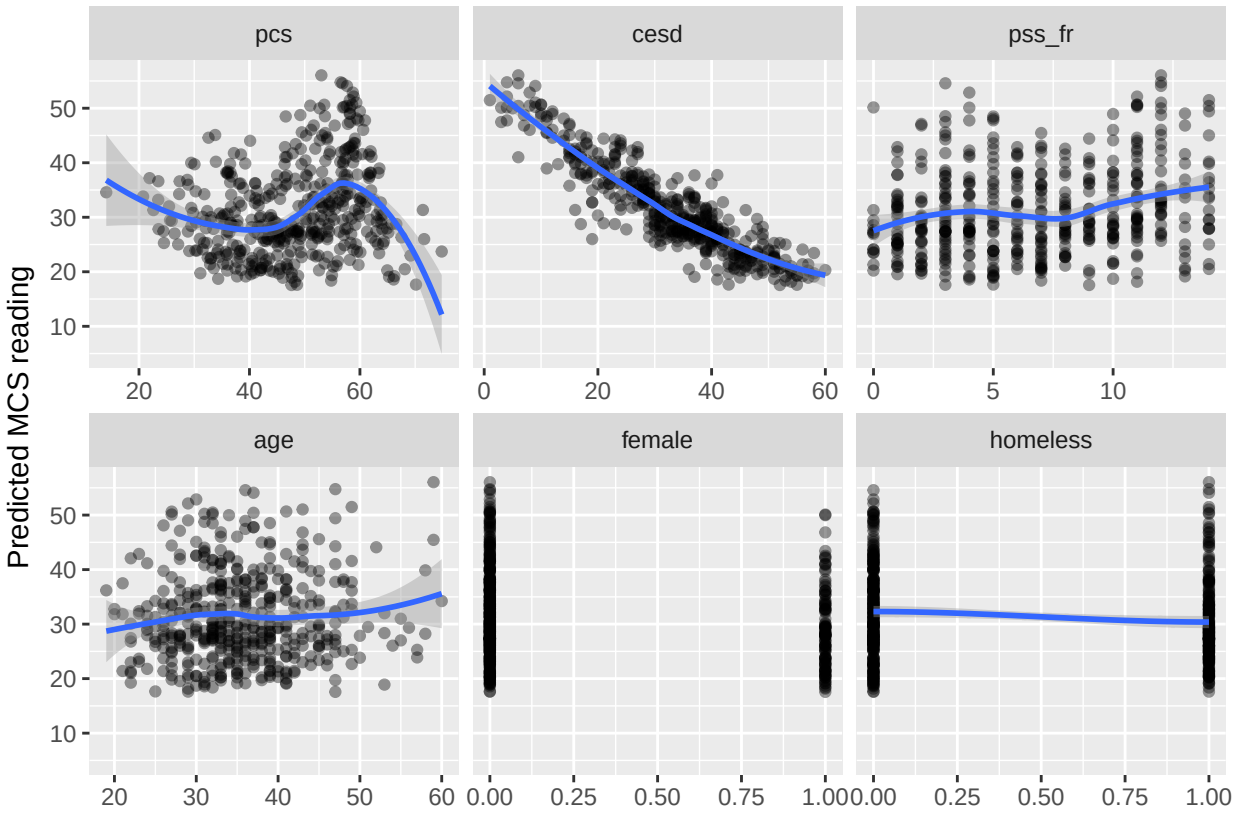



```
# Plot minimal depth v VIMP
gg_mdVIMP <- gg_minimal_vimp(gg_md)
plot(gg_mdVIMP)
```



Problem 10 Answer

```
#Create the variable dependence object from the random forest
gg_v <- 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 MCS reading", x="")
```


pcs
cesd
pss_fr
age
female
homeless
Predicted MCS reading
50
40
30
20
10
20
40
60
0
20
40
60
0
5
10
20
30
40
50
60
0.00
0.25
0.50
0.75
1.00