chemistry

2023-06-25

Data

```
chem <- read.table("~/Downloads/momeg/a-level-chemistry.txt", dec = ".", sep = " ")
colnames(chem) <- c('Board','Ascore','Total','GNum','Gender','Age','InstType','LEAID','L
EAInst','StuID')
chem <- chem[1:500,]
```

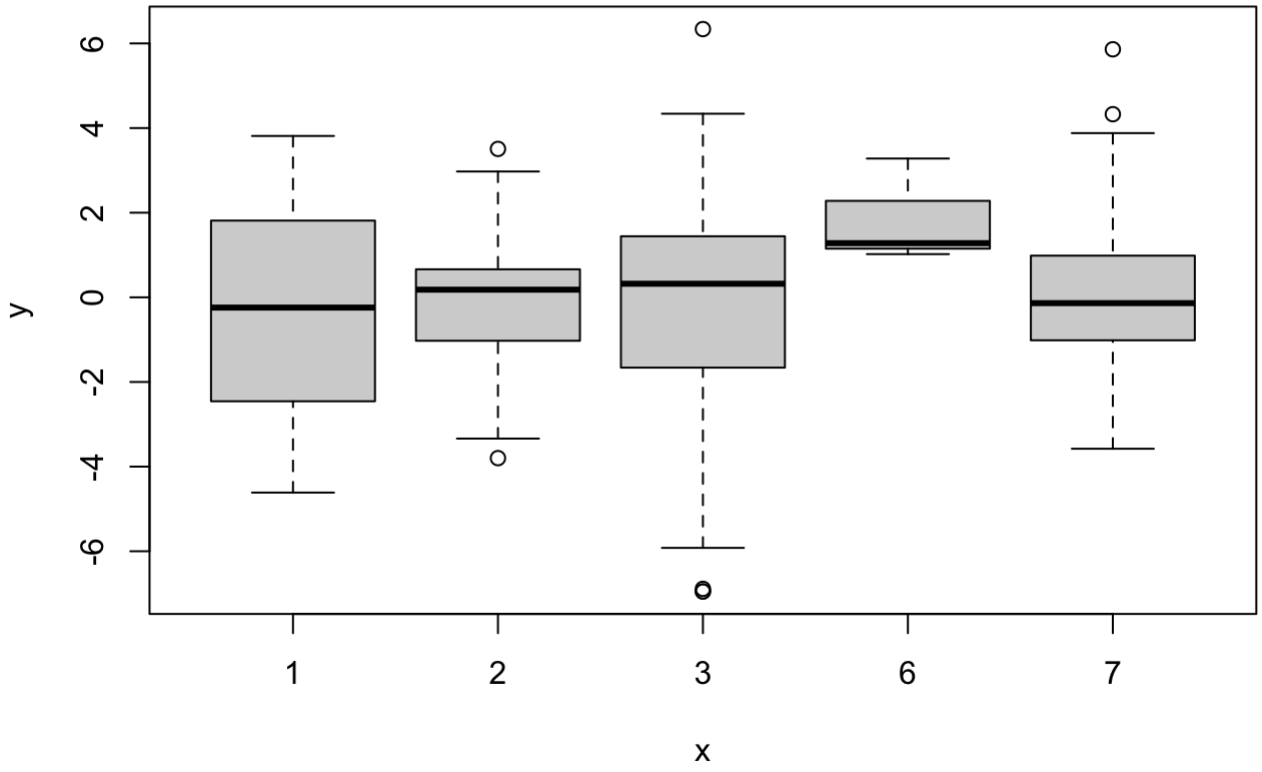
Tree

```
chemLM <- lmertree(Ascore ~ 1 | (1 | Board) |
                     Total + GNum + Gender + Age + LEAInst + InstType + LEAID, data = ch
em, cluster = Board)
```

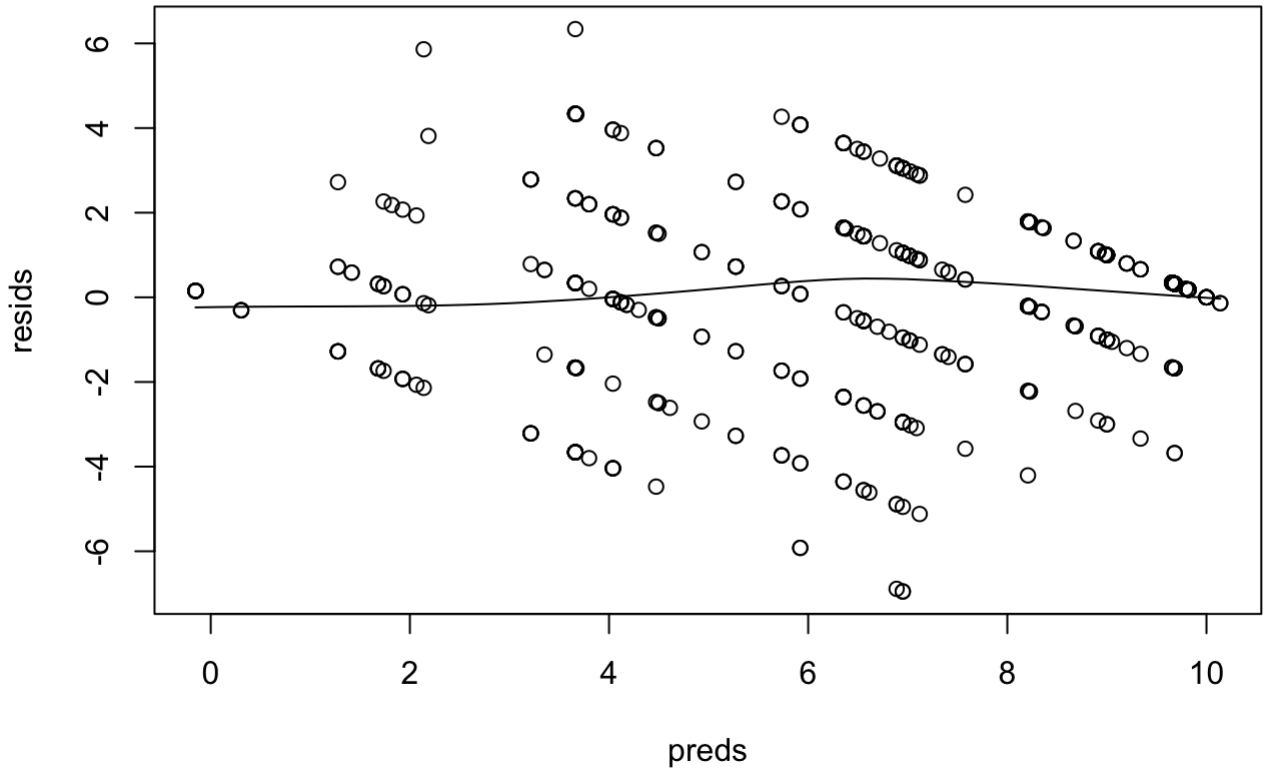
```
width(chemLM$tree)
```

```
## [1] 26
```

```
resids <- residuals(chemLM)
preds <- predict(chemLM)
plot(factor(chem$Board), resids)
```




```
scatter.smooth(preds, resids)
```




```
fligner.test(resids ~ chem$Board)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  resids by chem$Board
## Fligner-Killeen:med chi-squared = 9.6607, df = 4, p-value = 0.04655
```

```
bartlett.test(resids ~ chem$Board)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  resids by chem$Board
## Bartlett's K-squared = 14.233, df = 4, p-value = 0.006588
```

```
Chem_Train <- chem[1:300,]
cp_vals = 10^seq(-5, 5, length = 100)
colnames(Chem_Train) <- make.names(colnames(Chem_Train))
control = trainControl("repeatedcv", number = 10, repeats=10)

set.seed(2022)
Chem_Tree <- train(data=Chem_Train, Ascore ~ Total + GNum + Gender + Age + LEAInst + Ins
tType + LEAID ,  method="rpart", trControl=control,tuneGrid=expand.grid(cp=cp_vals))
```

```
## Warning in nominalTrainWorkflow(x = x, y = y, wts = weights, info = trainInfo,
## : There were missing values in resampled performance measures.
```

```
Chem_Best_Tree <- rpart(Ascore ~ Total + GNum + Gender + Age + LEAInst + InstType + LEAI
D, data=Chem_Train, cp=Chem_Tree$bestTune)
```

```
data <- head(chem$Ascore,10)
pre <- predict(chemLM, newdata = chem[1:10,])
sum((pre-data)^2)
```

```
## [1] 12.14109
```

```
rmse(data,pre)
```

```
## [1] 1.101866
```

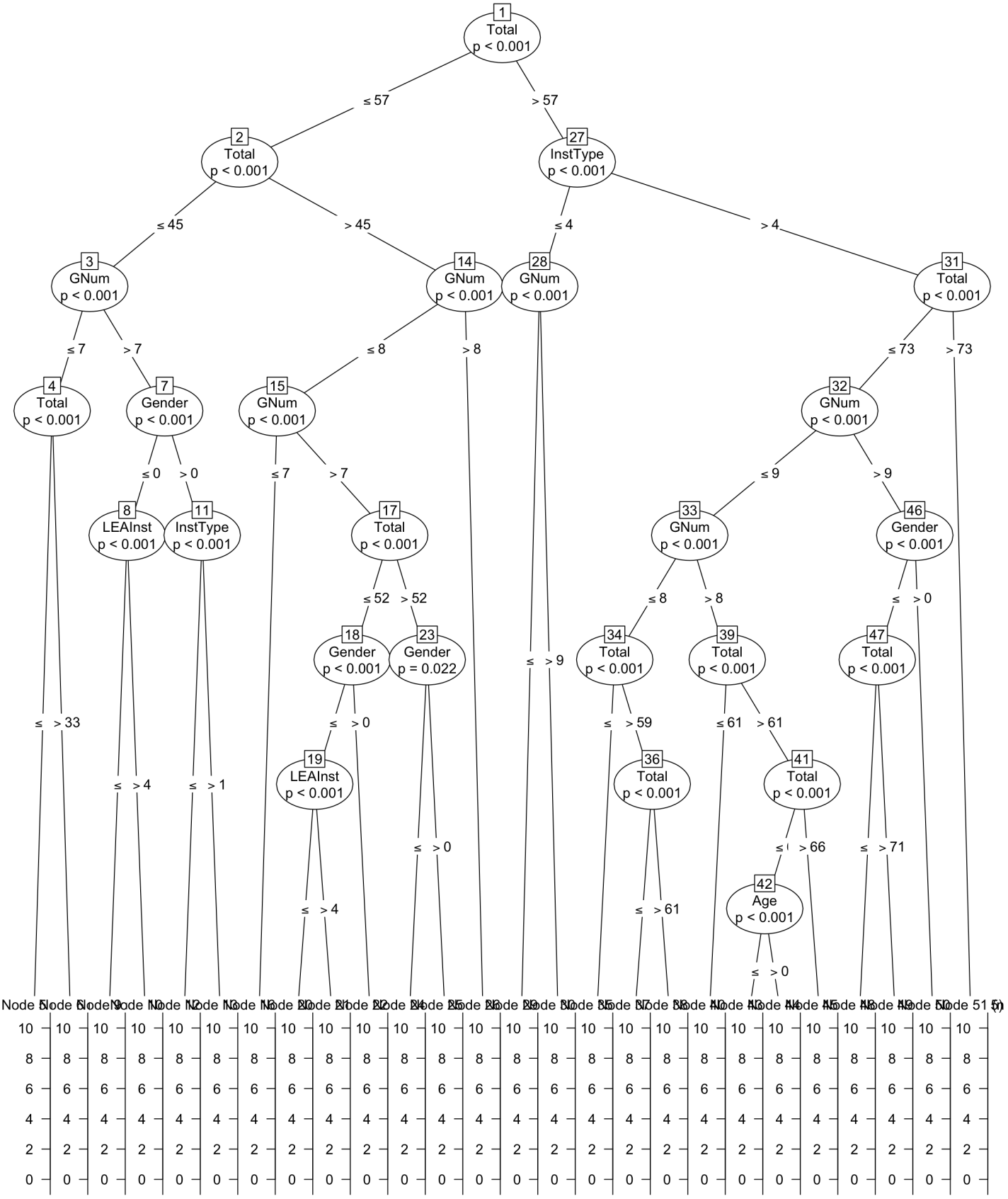
```
pre <- predict(Chem_Best_Tree, newdata = chem[1:10,])
sum((pre-data)^2)
```

```
## [1] 13.4973
```

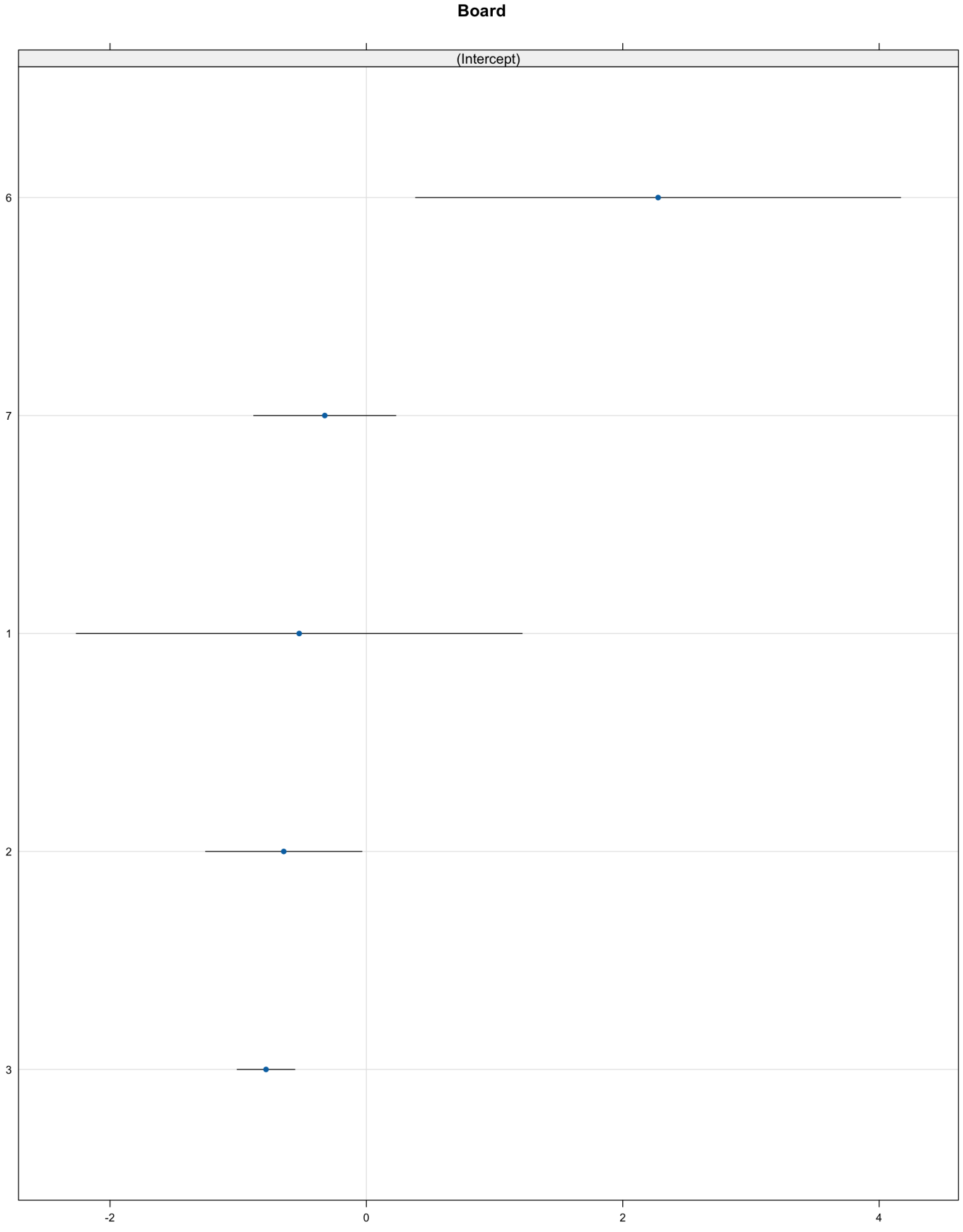
```
rmse(data,pre)
```

```
## [1] 1.161779
```

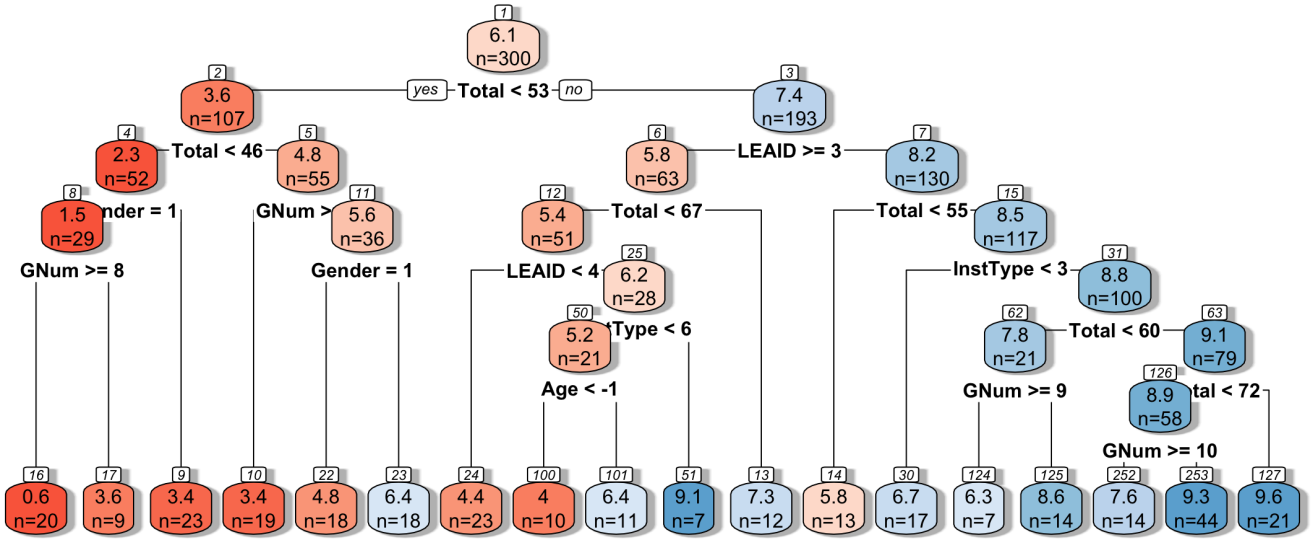
```
plot(chemLM)
```


1
Total
p < 0.001
≤ 57
> 57
2
Total
p < 0.001
27
InstType
p < 0.001
≤ 45
> 45
≤ 4
> 4
3
GNum
p < 0.001
14
GNum
p < 0.001
28
GNum
p < 0.001
31
Total
p < 0.001
≤ 7
> 7
≤ 8
> 8
≤ 73
> 73
4
Total
p < 0.001
7
Gender
p < 0.001
15
GNum
p < 0.001
32
GNum
p < 0.001
≤ 0
> 0
≤ 7
> 7
≤ 9
> 9
8
LEAInst
p < 0.001
11
InstType
p < 0.001
17
Total
p < 0.001
33
GNum
p < 0.001
46
Gender
p < 0.001
≤ 52
> 52
≤ 8
> 8
≤
> 0
18
Gender
p < 0.001
23
Gender
p = 0.022
≤ > 9
34
Total
p < 0.001
39
Total
p < 0.001
47
Total
p < 0.001
≤ > 33
≤ > 4
≤ > 1
≤ > 0
19
LEAInst
p < 0.001
≤ > 0
≤ > 59
36
Total
p < 0.001
≤ 61
> 61
41
Total
p < 0.001
≤ > 71
≤ > 4
≤ > 61
≤ > 66
42
Age
p < 0.001
≤ > 0
Node 5
Node 6
Node 9
Node 10
Node 12
Node 13
Node 16
Node 20
Node 21
Node 22
Node 24
Node 25
Node 26
Node 29
Node 30
Node 35
Node 37
Node 38
Node 40
Node 43
Node 44
Node 45
Node 48
Node 49
Node 50
Node 51
5)
10
8
6
4
2
0

```
## $Board
```




```
rpart.plot(Chem_Best_Tree, box.palette="RdBu", shadow.col="gray", nn=TRUE, cex=1, extra=
1)
```


1
6.1
n=300
yes
Total < 53
no
2
3.6
n=107
3
7.4
n=193
4
2.3
n=52
Total < 46
5
4.8
n=55
6
5.8
n=63
LEAID >= 3
7
8.2
n=130
8
1.5
n=29
nder = 1
GNum >
11
5.6
n=36
12
5.4
n=51
Total < 67
Total < 55
15
8.5
n=117
GNum >= 8
Gender = 1
LEAID < 4
25
6.2
n=28
InstType < 3
31
8.8
n=100
50
5.2
n=21
Type < 6
62
7.8
n=21
Total < 60
63
9.1
n=79
Age < -1
GNum >= 9
126
8.9
n=58
tal < 72
GNum >= 10
16
0.6
n=20
17
3.6
n=9
9
3.4
n=23
10
3.4
n=19
22
4.8
n=18
23
6.4
n=18
24
4.4
n=23
100
4
n=10
101
6.4
n=11
51
9.1
n=7
13
7.3
n=12
14
5.8
n=13
30
6.7
n=17
124
6.3
n=7
125
8.6
n=14
252
7.6
n=14
253
9.3
n=44
127
9.6
n=21