STAT_101C HW6

Junhyuk Jang

5/31/2017

SID : 004 728 134 LEC : 2 DIS : 2B

```
#(a)
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
require(tree)
```

```
## Loading required package: tree
```

```
library(ggplot2)
library(tree)
set.seed(9876)
setwd("/Users/junhyukjang/Desktop/UCLA_Academic/Spring 2017/STAT 101_C/HW")
birth <- read.csv("better2000births.csv")
attach(birth)
head(birth)
```

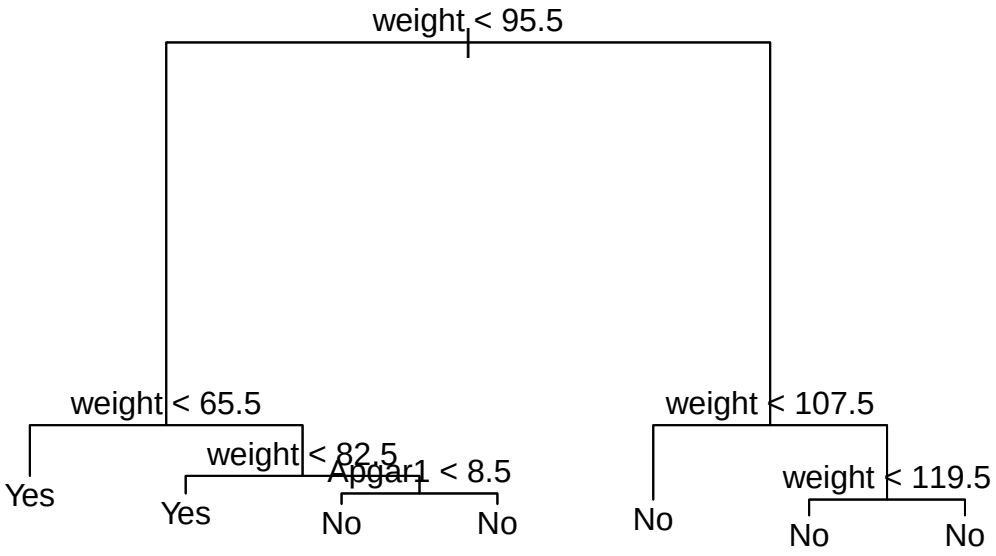
```
##   Gender Premie weight Apgar1 Fage Mage Feduc Meduc TotPreg Visits
## 1   Male     No    124      8   31   25    13    14       1     13
## 2 Female     No    177      8   36   26     9    12       2     11
## 3   Male     No    107      3   30   16    12     8       2     10
## 4 Female     No    144      6   33   37    12    14       2     12
## 5   Male     No    117      9   36   33    10    16       2     19
## 6 Female     No     98      4   31   29    14    16       3     20
##     Marital Racemom Racedad Hispmom Hispdad Gained     Habit MomPriorCond
## 1   Married   White   White NotHisp NotHisp     40 NonSmoker         None
## 2 Unmarried   White   White Mexican Mexican     20 NonSmoker         None
## 3 Unmarried   White Unknown Mexican Unknown     70 NonSmoker At Least One
## 4 Unmarried   White   White NotHisp NotHisp     50 NonSmoker         None
## 5   Married   White   Black NotHisp NotHisp     40 NonSmoker At Least One
## 6   Married   White   White NotHisp NotHisp     21 NonSmoker         None
##   BirthDef    DelivComp BirthComp
## 1     None At Least One      None
## 2     None At Least One      None
## 3     None At Least One      None
## 4     None At Least One      None
## 5     None         None      None
## 6     None         None      None
```

```
train=sample(1:nrow(birth),nrow(birth)/2)
test=birth[-train,]
```

```r
# Misclassification error using a tree method.
tree_m <- tree(Premie~., birth,subset=train)
summary(tree_m) #training misclassification rate is approximately 0.05606
```

```
##
## Classification tree:
## tree(formula = Premie ~ ., data = birth, subset = train)
## Variables actually used in tree construction:
## [1] "weight" "Apgar1"
## Number of terminal nodes:  7
## Residual mean deviance:  0.2862 = 283.9 / 992
## Misclassification error rate: 0.05606 = 56 / 999
```

```r
plot(tree_m)
text(tree_m,pretty=0)
```




```r
pred_t <- predict(tree_m, test, type="class")
tb <- table(pred_t,test$Premie)
tb
```

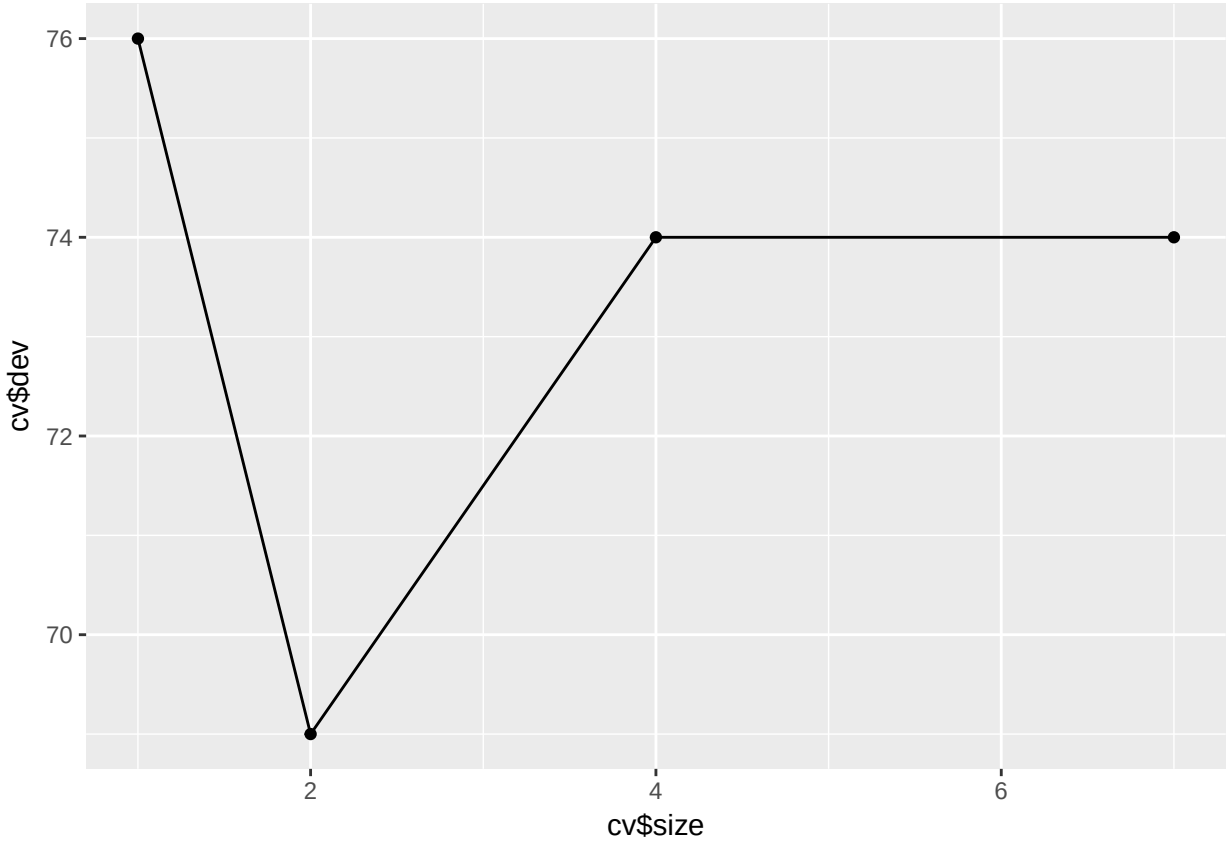
```
##
## pred_t  No Yes
##    No  904  51
##    Yes   4  40
```

```r
1-sum(diag(tb))/sum(tb)
```

```
## [1] 0.05505506
```

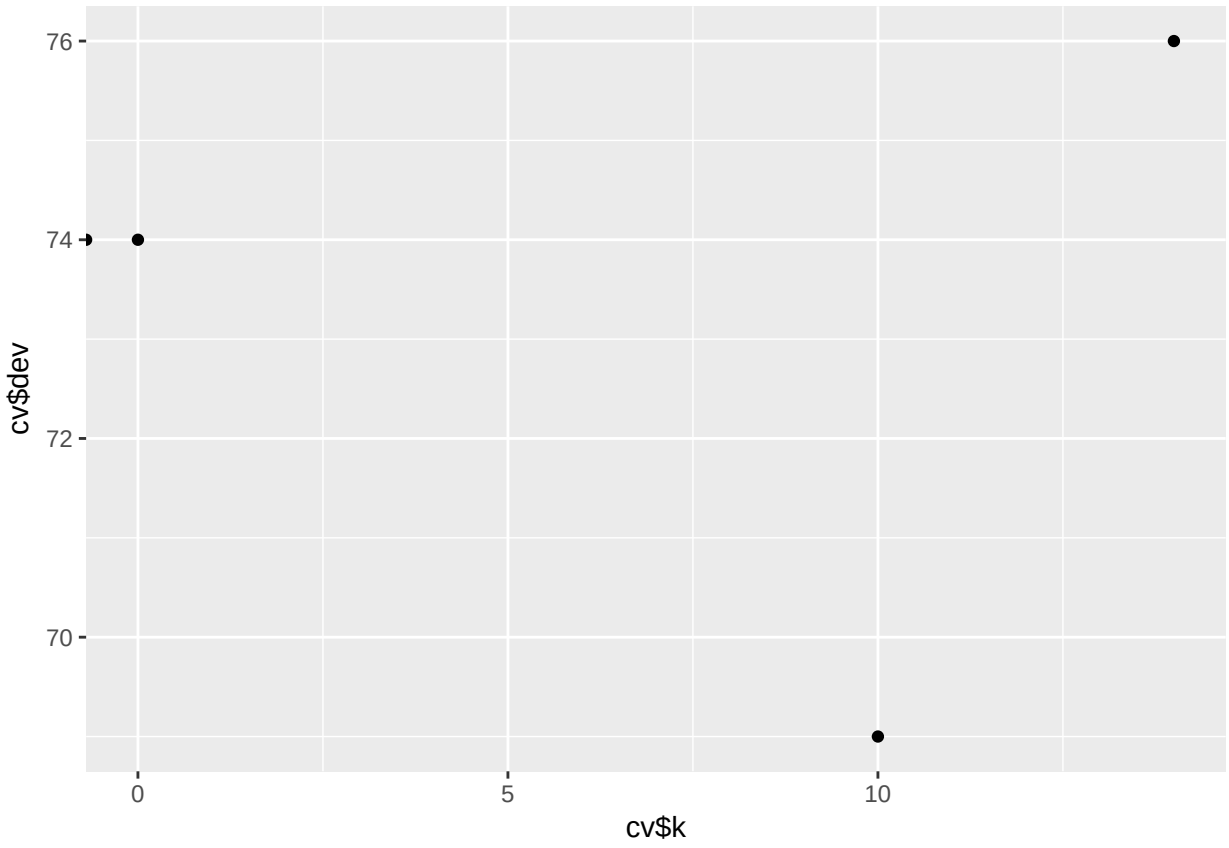
```r
# The testing missclassification error is approximately 0.0551 which is really
# similar to training missclassification error. This result tells that
# it is not overfitted.
```

```
#(b)
cv <- cv.tree(tree_m,FUN=prune.misclass)
qplot(x=cv$size, y = cv$dev, geom = c("point", "line"))
```

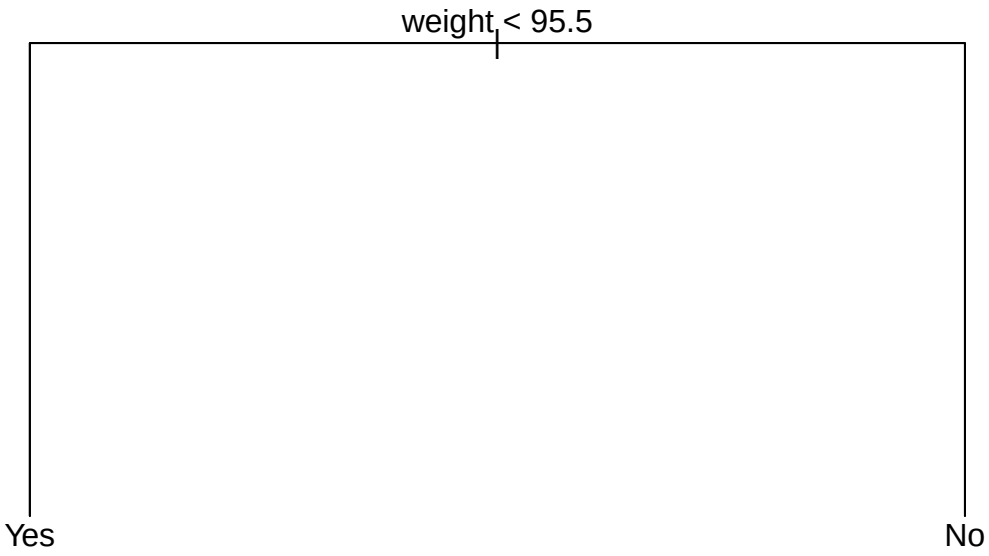



```
qplot(x=cv$k, y=cv$dev, geom="point","line")
```

```
## Warning: Ignoring unknown parameters: NA
```




```
prun_t <- prune.misclass(tree_m, best=2)
plot(prun_t)
text(prun_t,pretty=0)
```




```
pre <- predict(prun_t,test,type="class")
tb2 <- table(pre,test$Premie)
1-sum(diag(tb2))/sum(tb2)
```

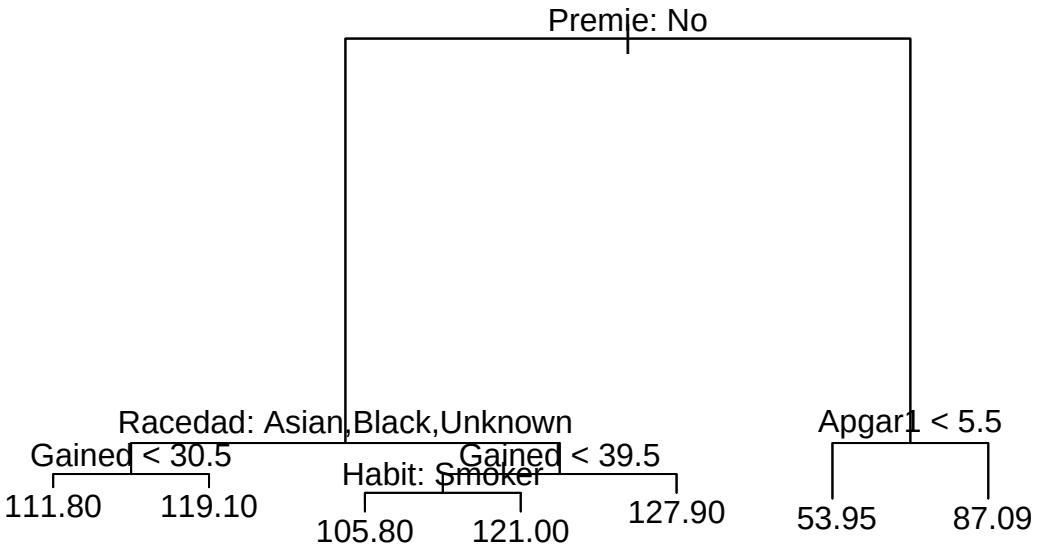
```
## [1] 0.08908909
```

```
# It is improved from 0.05606 to 0.08908909.
# Pruning do have an effect.

#(c)
# The pruned tree cannot tell us whether smoking is a potential cause of
# premature births. Instead, the tree can tell whether baby's births prematurely or not
# depending on the baby's birth weight. For example, if a baby's weight is less than 95.5,
# the baby is expected to be born prematurely and if a baby's weight is more than
# 95.5, the baby is expected to be not born prematurely.

#(d)
# The testing misclassification error I've got is approximately 8.9% which is
# 0.1% less than the simple prediction of the prematurely born baby(9%).
```

```
# Q2
# (a)
t <- tree(weight~., birth,subset=train)
summary(t)
```

```
##
## Regression tree:
## tree(formula = weight ~ ., data = birth, subset = train)
## Variables actually used in tree construction:
## [1] "Premie"  "Racedad" "Gained"  "Habit"   "Apgar1"
## Number of terminal nodes:  7
## Residual mean deviance:  247.2 = 245300 / 992
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -51.0200  -9.8080  -0.1143   0.0000   9.8990  55.9800
```

```
plot(t)
text(t,pretty=0)
```



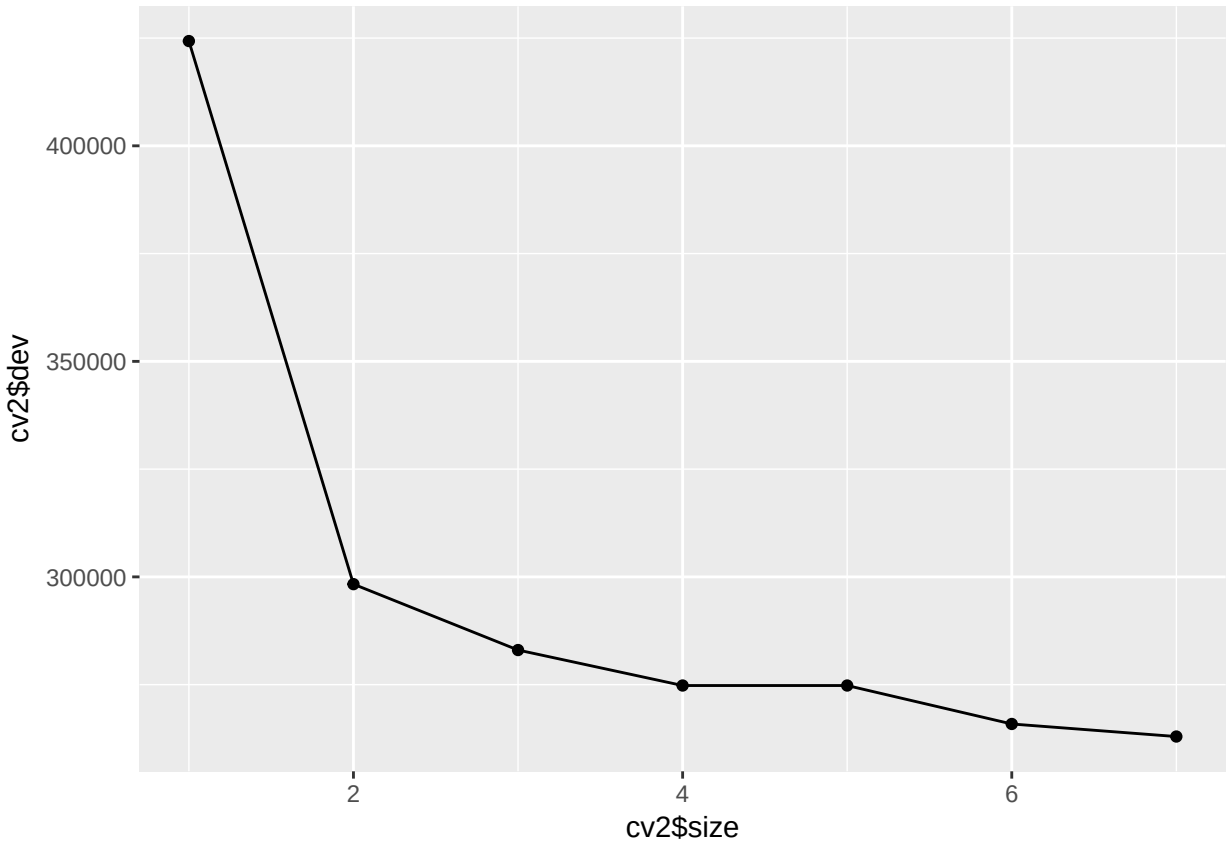
```r
p <- predict(t, test)
mean((test$weight-p)^2)
```

```
## [1] 271.4958
```

```r
# The MSE I have got is 271.4958.

#(b)
cv2 <- cv.tree(t,FUN=prune.tree)
cv2
```

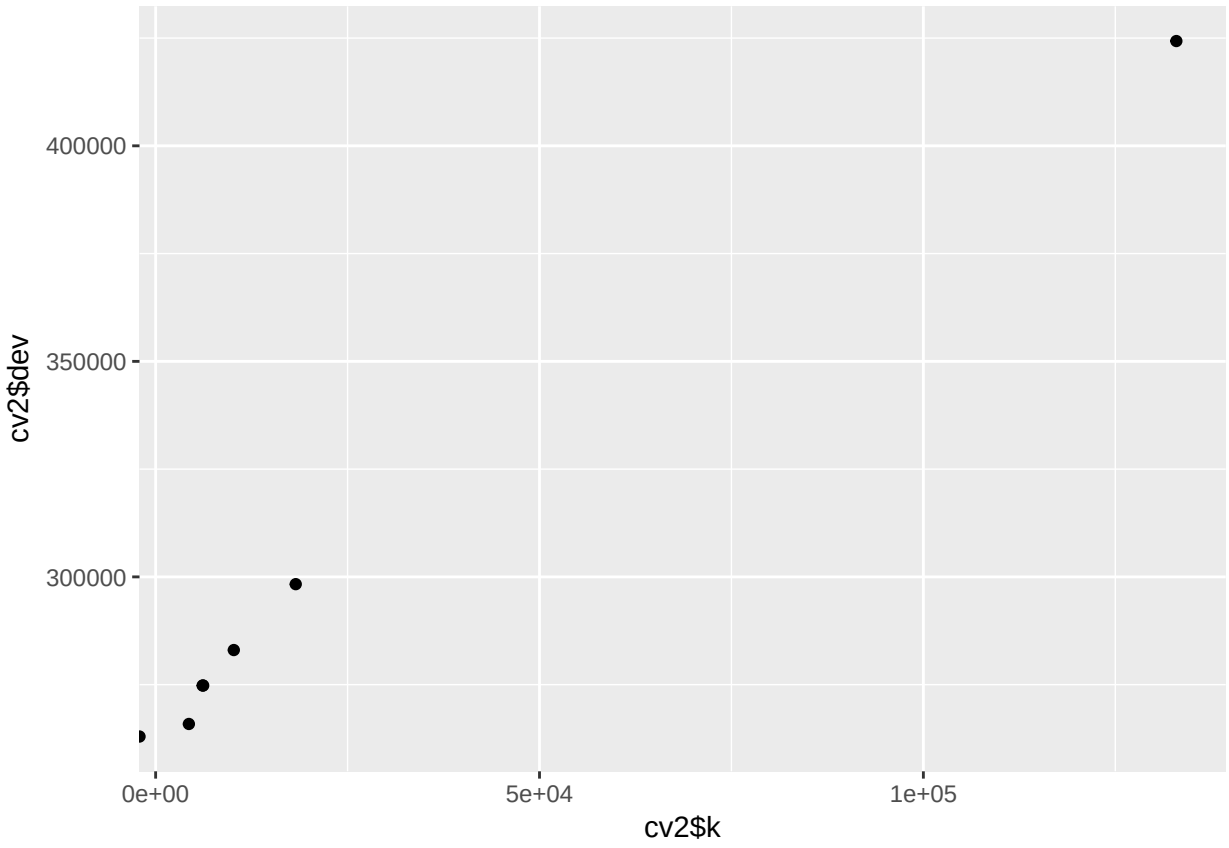
```
## $size
## [1] 7 6 5 4 3 2 1
##
## $dev
## [1] 262965.1 265880.6 274800.6 274800.6 283029.7 298315.3 424305.3
##
## $k
## [1]       -Inf   4331.117   6121.062   6193.148  10182.069  18248.564
## [7] 132951.883
##
## $method
## [1] "deviance"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```

```r
qplot(x=cv2$size, y = cv2$dev, geom = c("point", "line"))
```

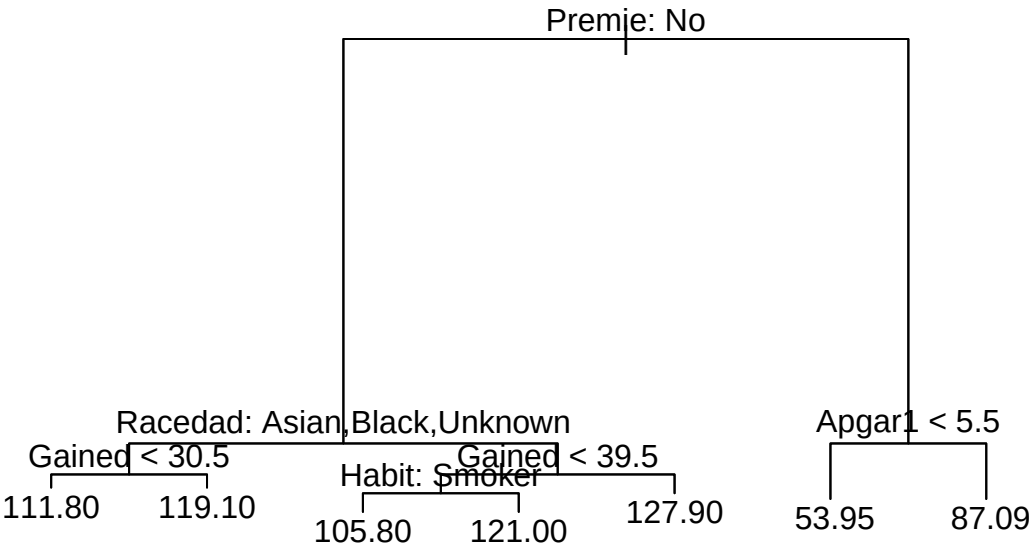



```
qplot(x=cv2$k, y=cv2$dev, geom="point","line")
```

```
## Warning: Ignoring unknown parameters: NA
```




```
# I can say the best size is the size of 7.

prune_7 <- prune.tree(t, best=7)
plot(prune_7)
text(prune_7,pretty=0)
```




```
pre7 <- predict(prune_7,test)
mean((test$weight-pre7)^2)
```

```
## [1] 271.4958
```

```
# The MSE I have got here is same as the above which is 271.4958.

#(c)
# When it comes to the baby who is not expected to be born prematurely,
# the factors 1. Racedad:Asian,Black,Unknown, 2. Gained, 3. Habit, 4. Apgar1
# are the important factors. The important factors have no influence with each others.
# In particular, the number of visits is not important predictor based on my pruned
# tree model.
```