Worksheet6

Sunny Lee

3/5/2021

1)

```
library(ISLR)
attach(Wage)
#?Wage
```

2)

```
library(tree)
```

```
## Warning: package 'tree' was built under R version 4.0.4
```

3)

```
#?tree
#?cv.tree
```

4)

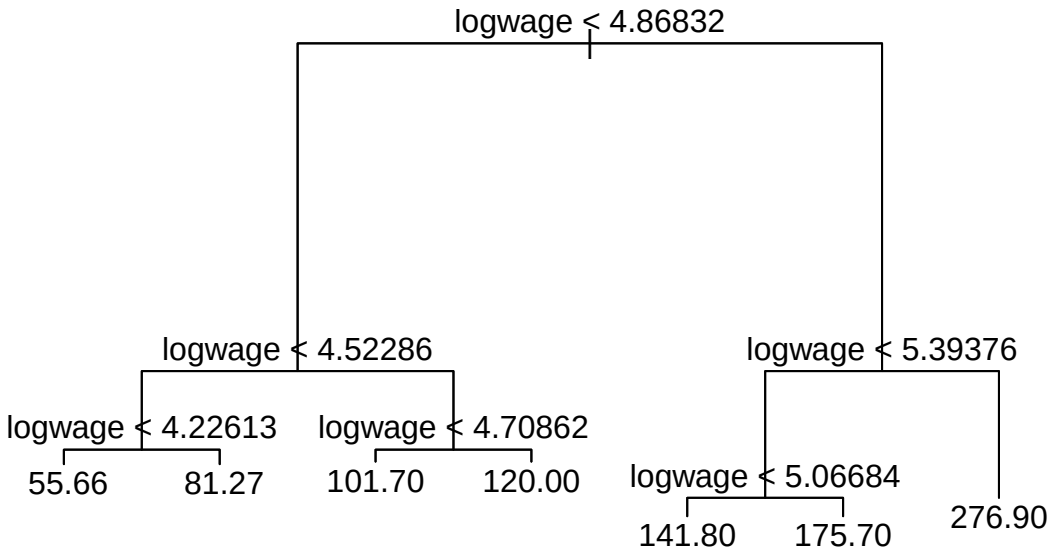
```
tree.wage <- tree(wage~., Wage)
```

5)

```
summary(tree.wage)
```

```
##
## Regression tree:
## tree(formula = wage ~ ., data = Wage)
## Variables actually used in tree construction:
## [1] "logwage"
## Number of terminal nodes:  7
## Residual mean deviance:  59.93 = 179400 / 2993
## Distribution of residuals:
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -49.45000  -5.54900  -0.01569   0.00000   6.59600  41.44000
```

```
plot(tree.wage)
text(tree.wage, pretty = 0)
```



6)

```
#?Wage
```

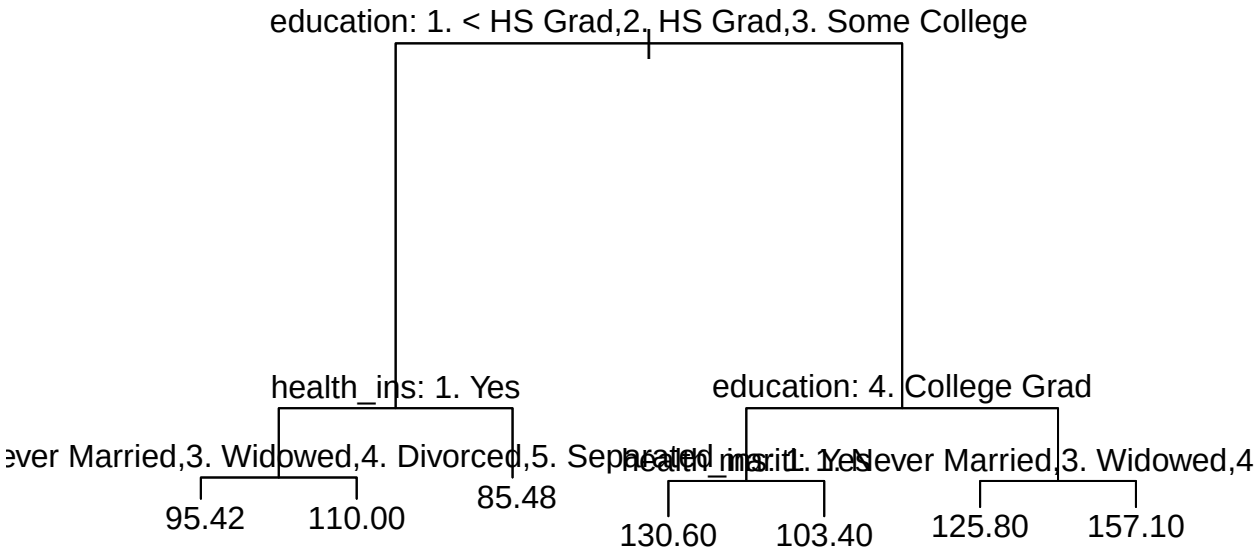
From the summary and the pretty tree construction, we find that logwage is the only predictor which was used in our tree. This is because logwage is the log transformation of the wage Which will force wage and logwage to have high collinearity.

7)

```
tree.wage1 <- tree(wage~.-logwage, Wage)
summary(tree.wage1)
```

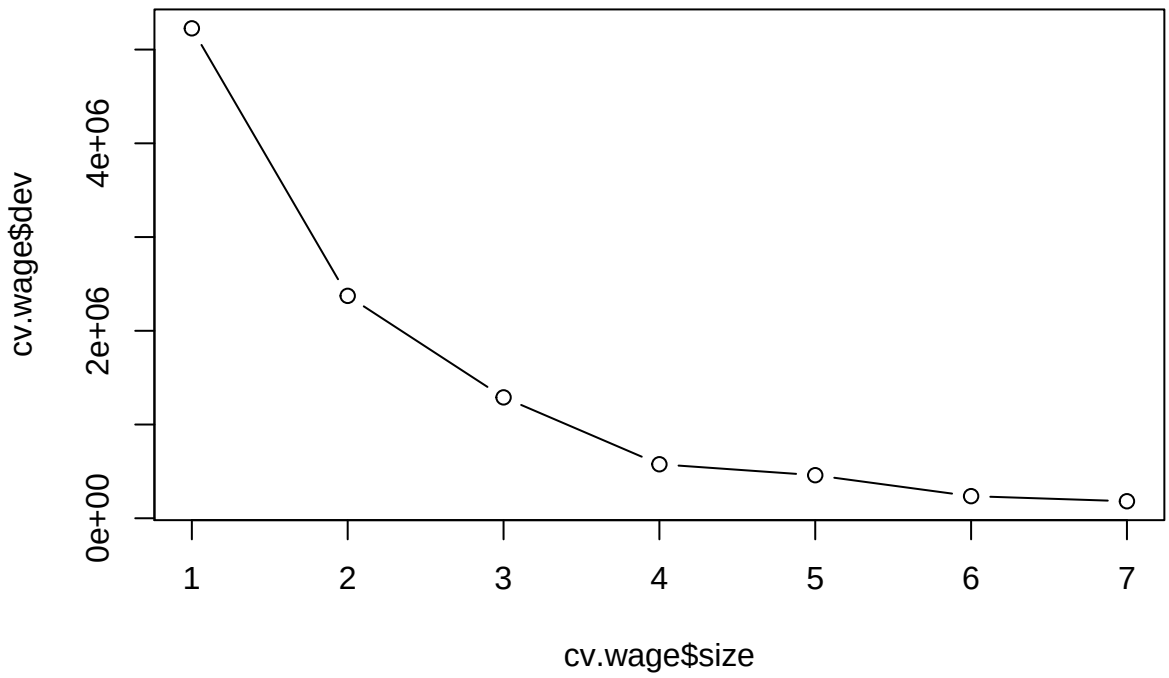
```
##
## Regression tree:
## tree(formula = wage ~ . - logwage, data = Wage)
## Variables actually used in tree construction:
## [1] "education"  "health_ins" "maritl"
## Number of terminal nodes:  7
## Residual mean deviance:  1246 = 3729000 / 2993
## Distribution of residuals:
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -116.000  -20.720   -4.199    0.000   14.590  224.100
```

```
plot(tree.wage1)
text(tree.wage1, pretty = 0)
```



8)

```
cv.wage <- cv.tree(tree.wage)
plot(cv.wage$size, cv.wage$dev, type = 'b')
```



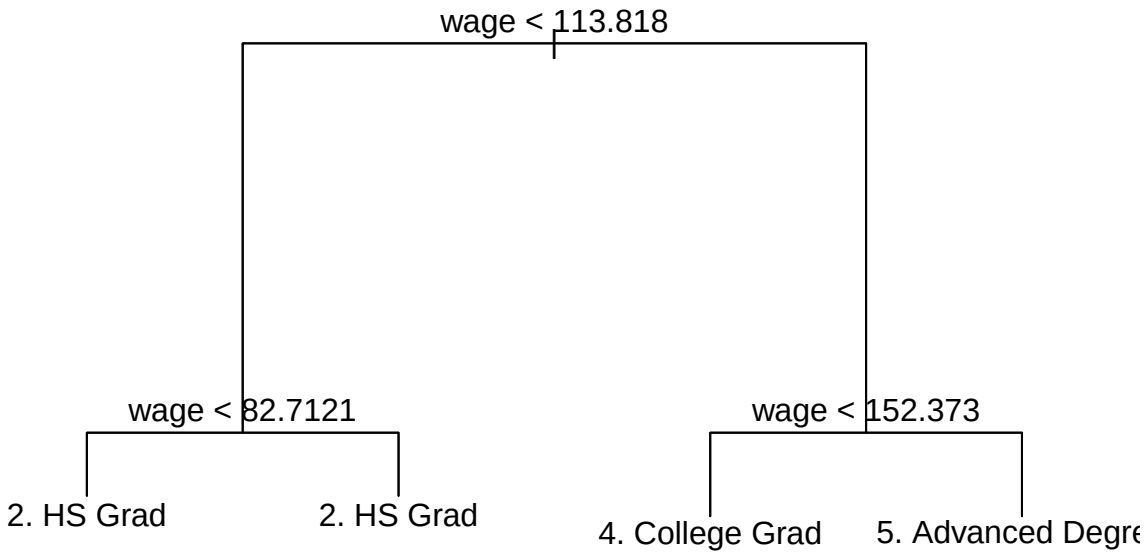
From the plot above, we seem to have a pretty good downsizing of dev each time we add a new node up to 6. Between 6 and 7 however, there seems to be minimal difference between dev = 6 and dev = 7.

9)

```
tree.education <- tree(education~.-logwage, Wage)
summary(tree.education)
```

```
##
## Classification tree:
## tree(formula = education ~ . - logwage, data = Wage)
## Variables actually used in tree construction:
## [1] "wage"
## Number of terminal nodes:  4
## Residual mean deviance:  2.752 = 8246 / 2996
## Misclassification error rate: 0.5957 = 1787 / 3000
```

```
plot(tree.education)
text(tree.education, pretty = 0 )
```



From the tree above, we find that wage is the most significant predictor, as our tree is split only based on the wage variable.