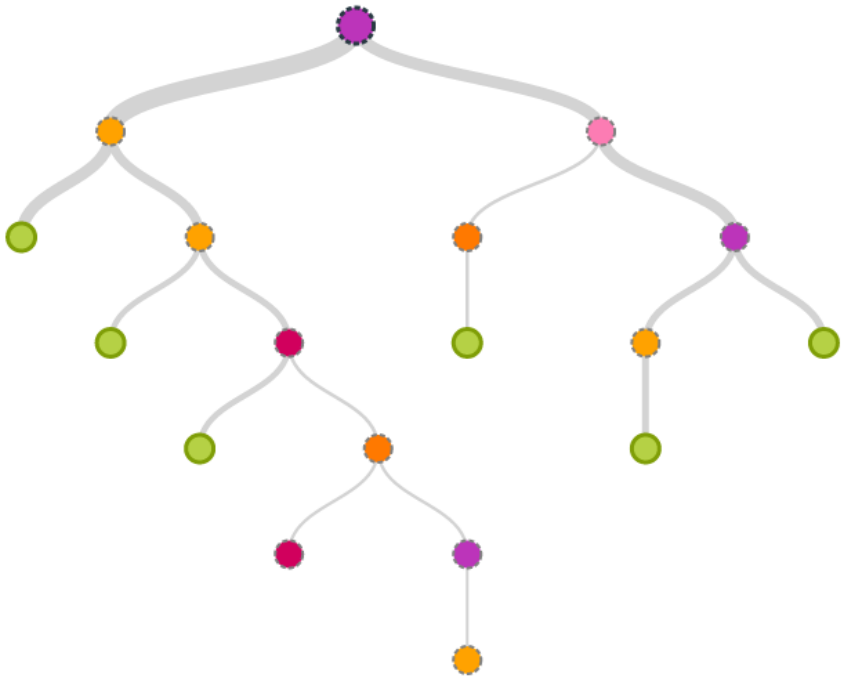


Decision Tree R Exercise: Regression Tree

Pilsung Kang
School of Industrial Management Engineering
Korea University

R Exercise: Regression Tree

- Data Set: Toyota Corolla Selling Price

Variable	Description	Variable	Description
		Guarantee_Period	Guarantee period in months
		ABS	Anti-Lock Brake System (Yes=1, No=0)
Price	Offer Price in EUROs	Airbag_1	Driver_Airbag (Yes=1, No=0)
Age_08_04	Age in months as in August 2004	Airbag_2	Passenger Airbag (Yes=1, No=0)
Mfg_Month	Manufacturing month (1-12)	Airco	Airconditioning (Yes=1, No=0)
Mfg_Year	Manufacturing Year	Automatic_airco	Automatic Airconditioning (Yes=1, No=0)
KM	Accumulated Kilometers on odometer	Boardcomputer	Boardcomputer (Yes=1, No=0)
Fuel_Type	Fuel Type (Petrol, Diesel, CNG)	CD_Player	CD Player (Yes=1, No=0)
HP	Horse Power	Central_Lock	Central Lock (Yes=1, No=0)
Met_Color	Metallic Color? (Yes=1, No=0)	Powered_Windows	Powered Windows (Yes=1, No=0)
Automatic	Automatic ((Yes=1, No=0)	Power_Steering	Power Steering (Yes=1, No=0)
CC	Cylinder Volume in cubic centimeters	Radio	Radio (Yes=1, No=0)
Doors	Number of doors	Mistlamps	Mistlamps (Yes=1, No=0)
Cylinders	Number of cylinders	Sport_Model	Sport Model (Yes=1, No=0)
Gears	Number of gear positions	Backseat_Divider	Backseat Divider (Yes=1, No=0)
Quarterly_Tax	Quarterly road tax in EUROs	Metallic_Rim	Metallic Rim (Yes=1, No=0)
Weight	Weight in Kilograms	Radio_cassette	Radio Cassette (Yes=1, No=0)
Mfr_Guarantee	Within Manufacturer's Guarantee period (Yes=1, No=0)	Parking_Assistant	Parking assistance system (Yes=1, No=0)
BOVAG_Guarantee	BOVAG (Dutch dealer network) Guarantee (Yes=1, No=0)	Tow_Bar	Tow Bar (Yes=1, No=0)

R Exercise: Regression Tree

- Define the performance evaluation function

```r
# Performance evaluation function for regression -------------------------
perf_eval_reg <- function(tgt_y, pre_y){
    # RMSE
    rmse <- sqrt(mean((tgt_y - pre_y)^2))
    # MAE
    mae <- mean(abs(tgt_y - pre_y))
    # MAPE
    mape <- 100*mean(abs((tgt_y - pre_y)/tgt_y))
    return(c(rmse, mae, mape))
}

# Performance table initialization
Perf_table <- matrix(0, nrow = 2, ncol = 3)
colnames(Perf_table)<- c("RMSE", "MAE", 'MAPE’)
rownames(Perf_table)<- c("MLR", "Regression Tree")
Perf_table
```

✓ perf_eval_reg( ) function

- Arguments: target values & predicted values
- Outputs: RMSE, MAE, MAPE

R Exercise: Regression Tree

- Load the dataset and Convert "Fule_Type" variable to three dummy variables

```
# Load the dataset
corolla <- read.csv("ToyotaCorolla.csv")

# Regression model 1: multivariate linear regression (MLR)
id_idx <- c(1,2)
category_idx <- 8

# Transform a categorical variable into a set of binary variables
dummy_p <- rep(0,nrow(corolla))
dummy_d <- rep(0,nrow(corolla))
dummy_c <- rep(0,nrow(corolla))

p_idx <- which(corolla$Fuel_Type == "Petrol")
d_idx <- which(corolla$Fuel_Type == "Diesel")
c_idx <- which(corolla$Fuel_Type == "CNG")

dummy_p[p_idx] <- 1
dummy_d[d_idx] <- 1
dummy_c[c_idx] <- 1

Fuel <- data.frame(dummy_p, dummy_d, dummy_c)
names(Fuel) <- c("Petrol","Diesel","CNG")

# Prepare the data for MLR
corolla_mlr_data <- cbind(corolla[,-c(id_idx, category_idx)], Fuel)
```

R Exercise: Regression Tree

- Train an MLR model and evaluate it with the test dataset

```
# Split the data into the training/validation sets
set.seed(12345)
trn_idx <- sample(1:nrow(corolla), round(0.7*nrow(corolla)))

MLR_trn <- corolla_mlr_data[trn_idx,]
MLR_tst <- corolla_mlr_data[-trn_idx,]

# Train the MLR
MLR_corolla <- lm(Price ~ ., data = MLR_trn)
MLR_corolla

# Performance Measure
MLR_corolla_haty <- predict(MLR_corolla, newdata = MLR_tst)
Perf_table[1,] <- perf_eval_reg(MLR_tst$Price, MLR_corolla_haty)
Perf_table
```

	RMSE	MAE	MAPE
MLR	1244.73	882.77	8.92
Regression Tree			

R Exercise: Regression Tree

- Install the necessary package, divide the dataset, and train the full tree

```
# Regression model 2: Regression Tree
# Install the necessary package

install.packages("tree")
library(tree)

corolla_rt_data <- corolla[,-id_idx]
RT_trn <- corolla_rt_data[trn_idx,]
RT_tst <- corolla_rt_data[-trn_idx,]

# Training the tree
RT_corolla <- tree(Price ~ ., RT_trn)
summary(RT_corolla)
```

✓ (Note) Variable <Fuel_type> is not converted to dummy variables

- Decision tree can handle both numeric and categorical variables

✓ Use the same function tree to train the regression tree

- Target variable <Price> is numeric → Regression tree is constructed

R Exercise: Regression Tree

- Check the result

```
> summary(RT_corolla)

Regression tree:
tree(formula = Price ~ ., data = RT_trn)
Variables actually used in tree construction:
[1] "Age_08_04" "Weight"    "HP"
Number of terminal nodes:  8
Residual mean deviance:  1769000 = 1.764e+09 / 997
Distribution of residuals:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-6565.00  -747.00   -12.97     0.00   787.60  5212.00
```

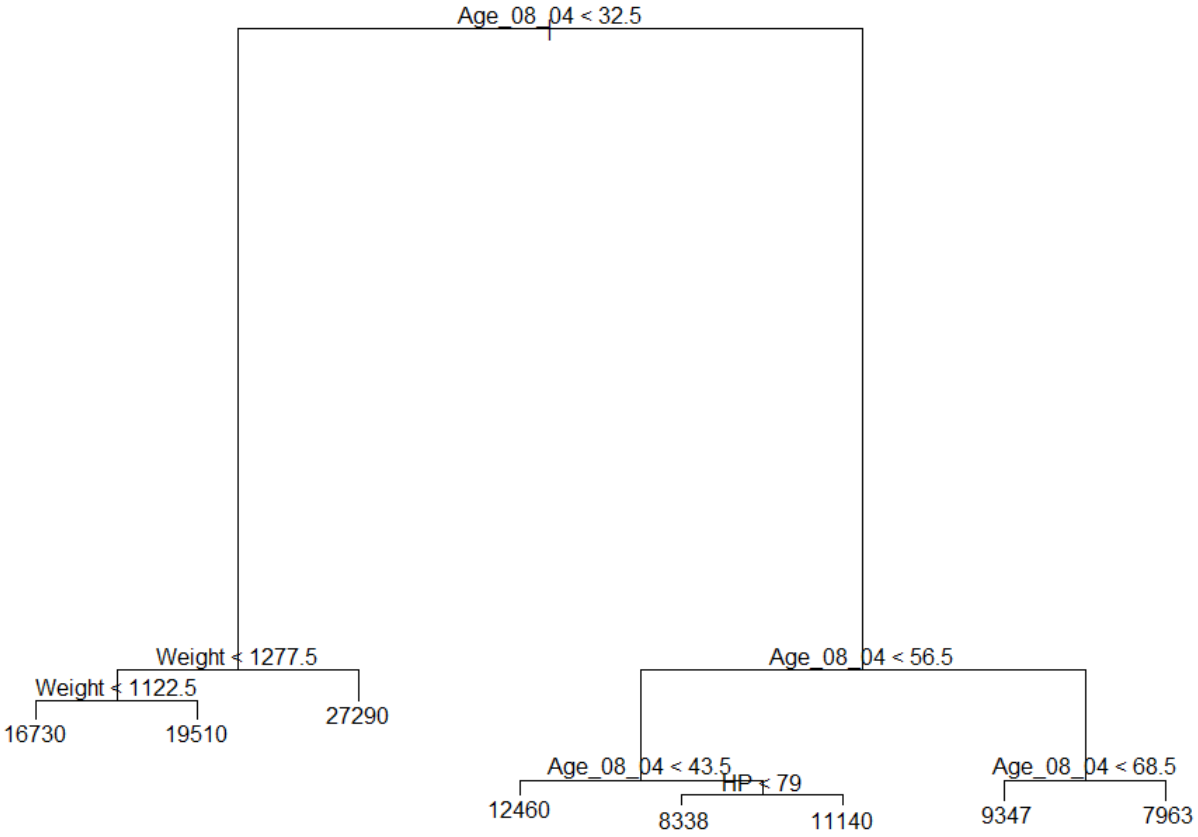
✓ Only tree variables (Age_08_04, Weight, HP) are used to construct the regression tree

✓ The number of leaf (terminal) node is 8

R Exercise: Regression Tree

- Check the result

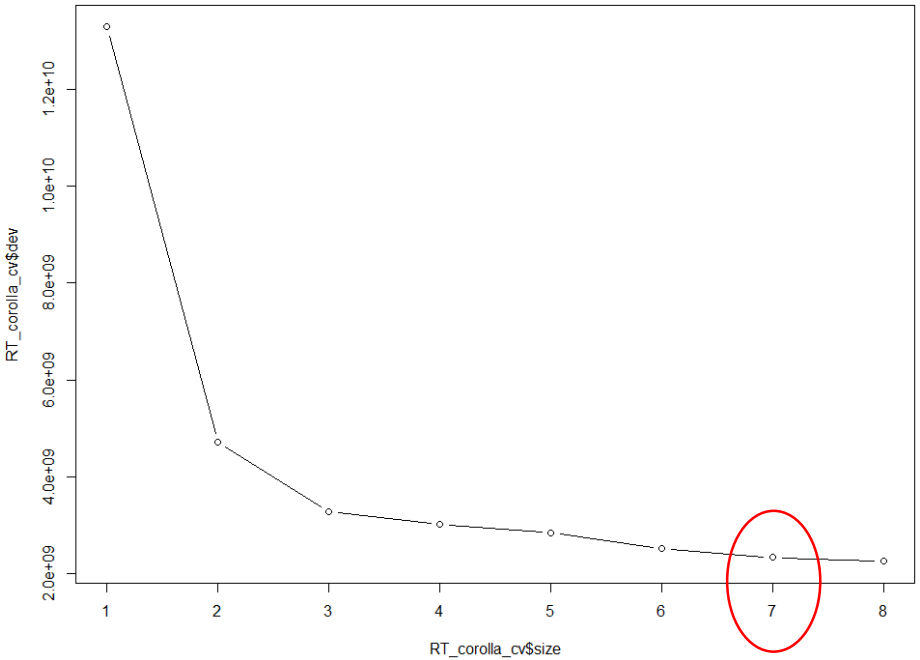
```
# Plot the tree
plot(RT_corolla)
text(RT_corolla, pretty = 1)
```



R Exercise: Regression Tree

- Prune the tree based on 10-fold cross-validation

```r
# Find the best tree
set.seed(12345)
RT_corolla_cv <- cv.tree(RT_corolla, FUN = prune.tree)

# Plot the pruning result
plot(RT_corolla_cv$size, RT_corolla_cv$dev, type = "b")
RT_corolla_cv
```



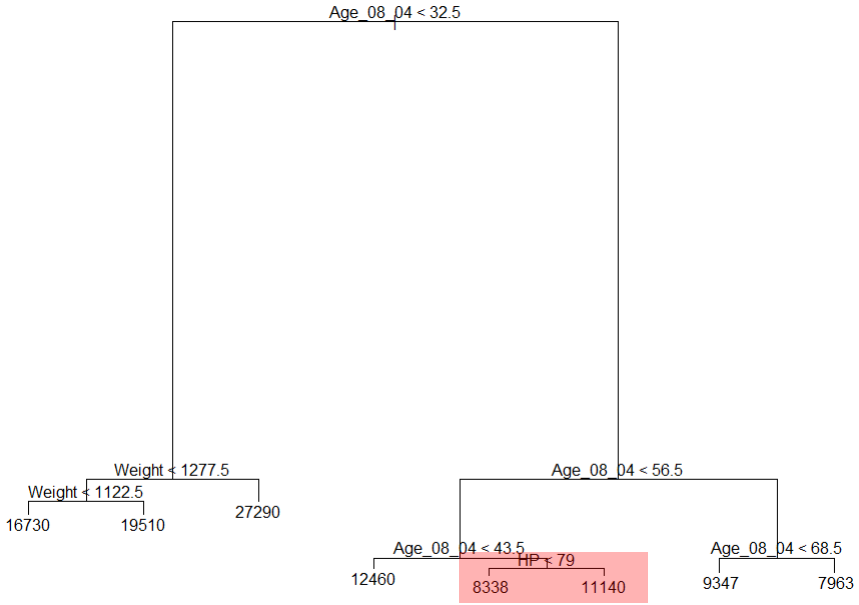

```
> RT_corolla_cv
$size
[1] 8 7 6 5 4 3 2 1

$dev
[1]  2253837543 2336834116 2523293483 2845773000 3014774171 3286331819 4713013726
[8] 13285022655

$k
[1]       -Inf  143453026  198196171  246070712  279463566  423633456 1496873688
[8] 8716995661

$method
[1] "deviance"

attr(,"class")
[1] "prune"         "tree.sequence"
```

R Exercise: Regression Tree

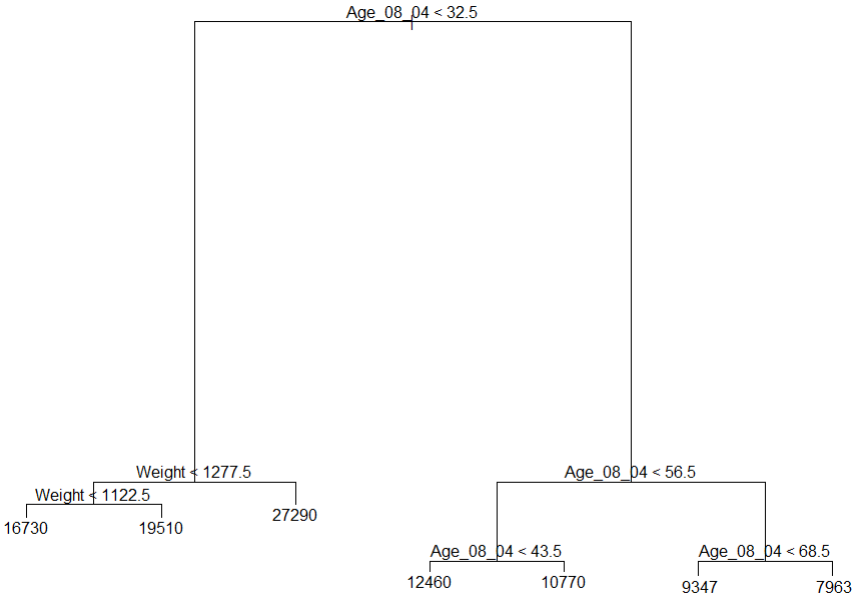
- Prune the tree based on 10-fold cross-validation

```
# Select the final model
RT_corolla_pruned <- prune.tree(RT_corolla, best = 7)
plot(RT_corolla_pruned)
text(RT_corolla_pruned, pretty = 1)
```

Before pruning



After pruning



R Exercise: Regression Tree

- Make prediction for the test dataset and evaluate the performance

```
# Prediction
RT_corolla_prey <- predict(RT_corolla_pruned, RT_tst, type = "vector")

# Compare the regression performance
Perf_table[2,] <- perf_eval_reg(RT_tst$Price, RT_corolla_prey)
Perf_table
```

```
> RT_corolla_prey[1:10]
       2        3        8       11       16       17       20       22       23
19514.55 19514.55 19514.55 19514.55 19514.55 19514.55 16729.77 19514.55 16729.77
      35
16729.77
```

	RMSE	MAE	MAPE
MLR	1244.73	882.77	8.92
Regression Tree	1416.66	1023.72	10.18


ANY
questions?