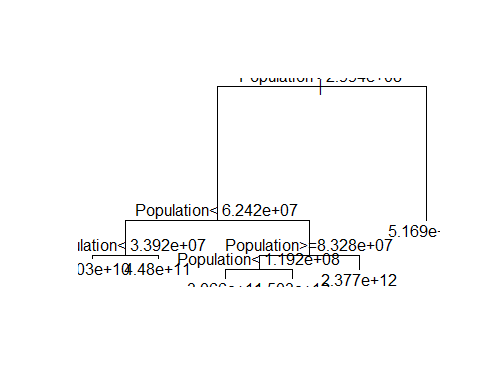
LR4\_Verkhovod

# Decision Tree Regression

setwd('C:/Users/VerkhovodTS/Desktop/R/')  
f\_train <- read.csv2('gdp\_train2.csv', header = TRUE, encoding = 'UNICOD')  
f\_train <- f\_train[,-1]   
f\_train <- f\_train[,-1]   
f\_test <- read.csv2('gdp\_test2.csv', header = TRUE, encoding = 'UNICOD')  
f\_test <- f\_test[,-1]   
f\_test <- f\_test[,-1]

Висновок: Ми завантажили навчальну і тестову вибірки. Видалили перші 2 стовпчики в кожній з вибірок.

library(rpart)  
model\_dt <- rpart(GDP ~ Population, f\_train, control = rpart.control(minsplit = 9))  
plot(model\_dt)  
text(model\_dt)



Висновок: побудовано дерево рішень, екзогенна змінна – Population

# Predicting

p\_dt <- predict(model\_dt, f\_test)  
train\_mse\_dt <- sum((f\_train$GDP-predict(model\_dt, f\_train))^2) /length(f\_train$GDP)  
test\_mse\_dt <- sum((f\_test$GDP-p\_dt)^2)/length(p\_dt)  
train\_mse\_dt

## [1] 2.590712e+23

test\_mse\_dt

## [1] 2.201086e+23

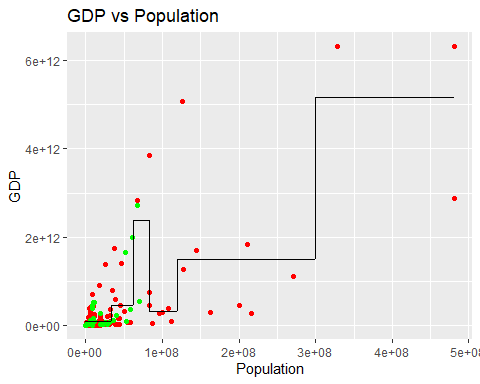
Висновок: значення середньоквадратичної помилки і на навчальній вибірці – 2.590712e+23, і на тестовій вибірці – 2.201086e+23 покращилися, перенавчання немає.

# Visualising

library(ggplot2)

## Warning: package 'ggplot2' was built under R version 3.6.3

x\_grid <- seq(min(f\_train$Population), max(f\_train$Population), 100)  
ggplot() +  
 geom\_point(aes(f\_train$Population, f\_train$GDP),colour = 'red') +  
 geom\_point(aes(f\_test$Population, f\_test$GDP),colour = 'green') +  
 geom\_line(aes(x\_grid, predict(model\_dt, data.frame(Population = x\_grid))),colour = 'black') +  
 ggtitle('GDP vs Population') +  
 xlab('Population') +  
 ylab('GDP')



Висновок: на графіку червоним позначені точки навчальної вибірки, зеленим – точки тестової вибірки, чорна лінія – модельні значення.

# Random forest

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.6.3

## randomForest 4.6-14

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

set.seed(1234)  
model\_rf = randomForest(x = f\_train['Population'],  
 y = f\_train$GDP,  
 ntree = 2)

# Predicting

p\_rf <- predict(model\_rf, f\_test)  
  
train\_mse\_rf <- sum((f\_train$GDP-predict(model\_rf, f\_train))^2)/length(f\_train$GDP)  
test\_mse\_rf <- sum((f\_test$GDP-p\_rf)^2)/length(p\_rf)  
  
train\_mse\_rf

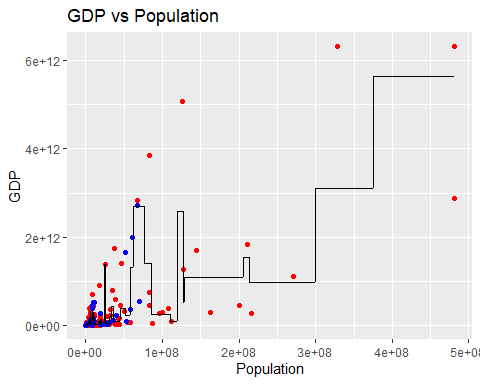
## [1] 2.649926e+23

test\_mse\_rf

## [1] 1.965748e+23

Висновок: значення середньоквадратичної помилки на навчальній вибірці погіршилась – 2.649926e+23, на тестовій вибірці – 1.965748e+23 покращилися, немає перенавчання. # Visualising

ggplot() +  
 geom\_point(aes(f\_train$Population, f\_train$GDP),colour = 'red') +  
 geom\_point(aes(f\_test$Population, f\_test$GDP),colour = 'blue') +  
 geom\_line(aes(x\_grid, predict(model\_rf, data.frame(Population = x\_grid))),colour = 'black') +  
 ggtitle('GDP vs Population') +  
 xlab('Population') +  
 ylab('GDP')



Висновок: на графіку червоним позначені точки навчальної вибірки, синім – точки тестової вибірки, чорна лінія – модельні значення. # Saving results

fit <- read.csv2('GDP\_fit.csv', header = TRUE, encoding = 'UNICOD')  
fit$p\_dt <- p\_dt  
fit$p\_rf <- p\_rf  
head(fit)

## X p\_sr p\_mr p\_pr p\_dt p\_rf  
## 1 1 46436244085 98042633247 136800418159 84027302363 15327125860  
## 2 2 103784685305 275624979933 141028028181 84027302363 80138440858  
## 3 3 129042028034 281828223536 144205207894 84027302363 31747431133  
## 4 4 33680994776 125878688709 136431871990 84027302363 18646311719  
## 5 5 21554217315 -43838422475 136276759010 84027302363 5940473558  
## 6 6 109498751849 138411791713 141677081440 84027302363 80138440858

write.csv2(fit[-1], file = "GDP\_fit.csv")

Висновок: результати моделювання збережені у файлі.