Samostoyatelynaya rabota 1

setwd('C:/Users/VerkhovodTS/Desktop/R')  
f <- read.csv2('AGE.csv', header = TRUE, encoding = 'UNICOD')  
f <- f[,-1]   
library(dplyr)

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

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

head(f)

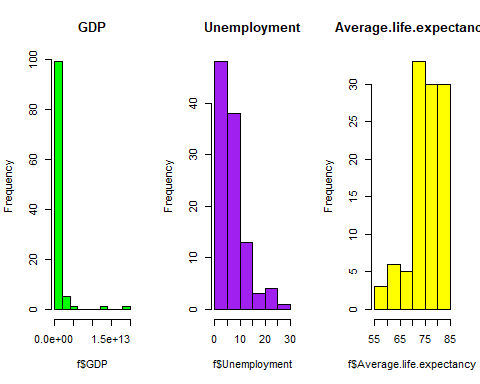
## GDP Unemployment Average.life.expectancy Constitutional.form  
## 1 2.019176e+10 11 64 Republic  
## 2 1.301969e+10 14 78 Republic  
## 3 6.430000e+11 8 76 Republic  
## 4 1.152746e+10 18 75 Republic  
## 5 1.330000e+12 6 83 Constitutional monarchy  
## 6 4.180000e+11 5 82 Republic  
## GDP.per.capita Ec.active.population Birth.rate  
## 1 556 27.40746 33  
## 2 4531 45.82874 12  
## 3 14592 45.48837 17  
## 4 3915 44.26953 14  
## 5 54066 52.79712 13  
## 6 47549 51.90916 10

Висновок: для виявлення факторів, що впливають на середню тривалість життя ми вибрали такі показники як: ВВП, рівень безробіття, форма правліяння в країні, ВВП на душу населення, частка економічно активного населення, рінень народжуваності. Кількість спостережень 107. Кількість змінних 7.

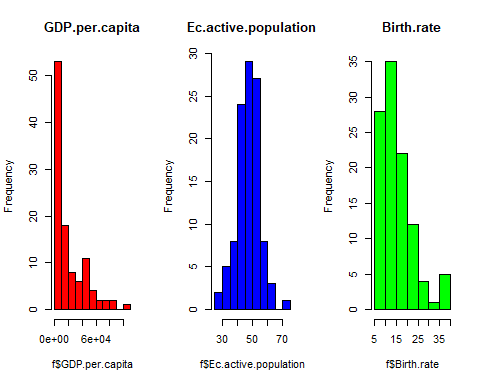
library(ggplot2)

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

par(mfrow = c(1,3))  
hist(f$GDP, col='green', main = 'GDP')  
hist(f$Unemployment,col = 'purple', main = 'Unemployment')  
hist(f$Average.life.expectancy,col = 'yellow', main = 'Average.life.expectancy')



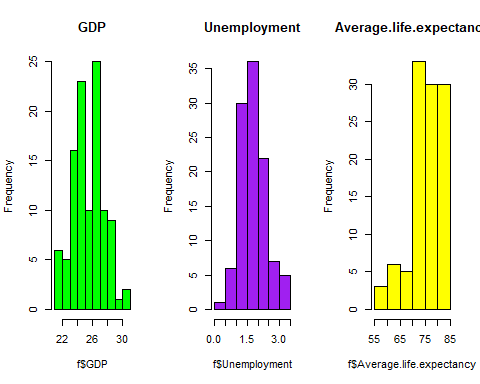
hist(f$GDP.per.capita,col = 'red', main = 'GDP.per.capita')  
hist(f$Ec.active.population,col = 'Blue', main = 'Ec.active.population')  
hist(f$Birth.rate,col = 'green', main = 'Birth.rate')



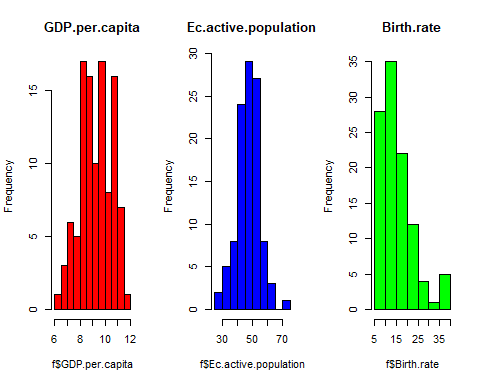
Висновок: Розподіл змінних “GDP”, “Unemployment”, “GDP.per.capita” мають довгий хвіст.

# Проведемо логарифмування для позбавлення від хвостів

f\_log <- f  
f\_log$GDP <- log(f$GDP, )  
f\_log$Unemployment <- log(f$Unemployment, )  
f\_log$GDP.per.capita <- log(f$GDP.per.capita, )  
  
f<- f\_log   
par(mfrow = c(1,3))  
hist(f$GDP, col='green', main = 'GDP')  
hist(f$Unemployment,col = 'purple', main = 'Unemployment')  
hist(f$Average.life.expectancy,col = 'yellow', main = 'Average.life.expectancy')



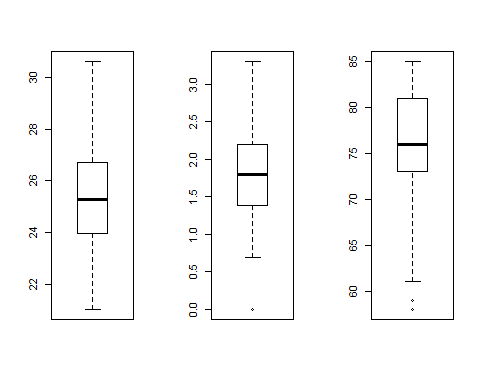
hist(f$GDP.per.capita,col = 'red', main = 'GDP.per.capita')  
hist(f$Ec.active.population,col = 'Blue', main = 'Ec.active.population')  
hist(f$Birth.rate,col = 'green', main = 'Birth.rate')



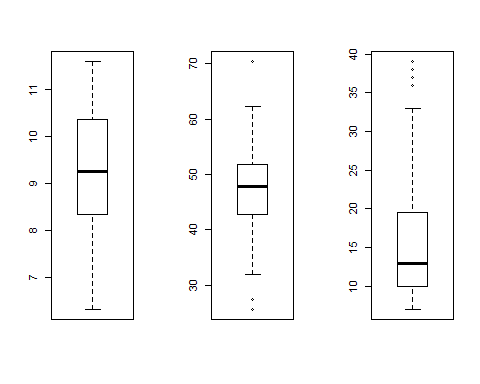
Висновок: ми позбавились від довгих хвостів, тепер всі змінні мають нормальний розподіл.

# Перевіримо дані на наявність викидів

par(mfrow = c(1,3))  
boxplot(f$GDP)  
boxplot(f$Unemployment)  
boxplot(f$Average.life.expectancy)



boxplot(f$GDP.per.capita)  
boxplot(f$Ec.active.population)  
boxplot(f$Birth.rate)



Висновок: викиди мають всі знінні, окрім “GDP” та “GDP.per.capita”.

# Descriptive statistics

library(psych)

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

##   
## Attaching package: 'psych'

## The following objects are masked from 'package:ggplot2':  
##   
## %+%, alpha

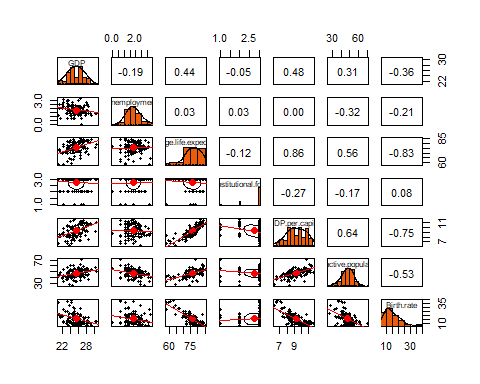
describe(f)

## vars n mean sd median trimmed mad min max  
## GDP 1 107 25.41 2.01 25.28 25.41 2.02 21.02 30.60  
## Unemployment 2 107 1.79 0.62 1.79 1.78 0.60 0.00 3.30  
## Average.life.expectancy 3 107 75.79 6.13 76.00 76.53 5.93 58.00 85.00  
## Constitutional.form\* 4 104 2.73 0.53 3.00 2.83 0.00 1.00 3.00  
## GDP.per.capita 5 107 9.24 1.25 9.26 9.29 1.45 6.32 11.59  
## Ec.active.population 6 107 47.13 7.36 47.77 47.34 6.23 25.63 70.37  
## Birth.rate 7 107 15.71 7.42 13.00 14.57 5.93 7.00 39.00  
## range skew kurtosis se  
## GDP 9.58 0.07 -0.43 0.19  
## Unemployment 3.30 0.09 -0.02 0.06  
## Average.life.expectancy 27.00 -0.97 0.70 0.59  
## Constitutional.form\* 2.00 -1.79 2.29 0.05  
## GDP.per.capita 5.27 -0.20 -0.78 0.12  
## Ec.active.population 44.74 -0.20 0.68 0.71  
## Birth.rate 32.00 1.33 1.32 0.72

Висновок: Пропущені значення має змінна Constitutional.form Викиди мають усі змінні, окрім “GDP” та “GDP.per.capita”.

# Correlations

pairs.panels(f, lm=TRUE, method = "pearson", hist.col = "#ed5d09")



Висновок: найбільший вплив на Середню тривалість життя має показник ВВП на душу населення.

# Fill n/a

f$Constitutional.form <- as.numeric(as.factor(f$Constitutional.form))-1  
library(tidyr)

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

f\_fill <- f  
f\_fill$Constitutional.form <-ifelse(is.na(f$Constitutional.form), round(mean(f$Constitutional.form, na.rm = TRUE)),f$Constitutional.form)  
f<- f\_fill

Висновок: Пропуски були заповнені середніми значеннями.

# Ejections (outside the three sigma)

f\_ej <- f  
f\_ej$GDP<- ifelse(f$GDP < mean(f$GDP)-sd(f$GDP)\*3, mean(f$GDP)-sd(f$GDP)\*3, f$GDP)  
f\_ej$Unemployment<- ifelse(f$Unemployment < mean(f$Unemployment)-sd(f$Unemployment)\*3, mean(f$Unemployment)-sd(f$Unemployment)\*3, f$Unemployment)  
f\_ej$Average.life.expectancy<- ifelse(f$Average.life.expectancy < mean(f$Average.life.expectancy)-sd(f$Average.life.expectancy)\*3, mean(f$Average.life.expectancy)+sd(f$Average.life.expectancy)\*3, f$Average.life.expectancy)  
f\_ej$Ec.active.population<- ifelse(f$Ec.active.population < mean(f$Ec.active.population)-sd(f$Ec.active.population)\*3, mean(f$Ec.active.population)-sd(f$Ec.active.population)\*3, f$Ec.active.population)  
  
f\_ej$Ec.active.population<- ifelse(f$Ec.active.population < mean(f$Ec.active.population)+sd(f$Ec.active.population)\*3, f$Ec.active.population, mean(f$Ec.active.population)+sd(f$Ec.active.population)\*3)  
f\_ej$Birth.rate<- ifelse(f$Birth.rate < mean(f$Birth.rate)+sd(f$Birth.rate)\*3, f$Birth.rate, mean(f$Birth.rate)+sd(f$Birth.rate)\*3)  
describe(f\_ej)

## vars n mean sd median trimmed mad min max  
## GDP 1 107 25.41 2.01 25.28 25.41 2.02 21.02 30.60  
## Unemployment 2 107 1.79 0.62 1.79 1.78 0.60 0.00 3.30  
## Average.life.expectancy 3 107 75.79 6.13 76.00 76.53 5.93 58.00 85.00  
## Constitutional.form 4 107 1.74 0.52 2.00 1.84 0.00 0.00 2.00  
## GDP.per.capita 5 107 9.24 1.25 9.26 9.29 1.45 6.32 11.59  
## Ec.active.population 6 107 47.12 7.32 47.77 47.34 6.23 25.63 69.20  
## Birth.rate 7 107 15.70 7.39 13.00 14.57 5.93 7.00 37.98  
## range skew kurtosis se  
## GDP 9.58 0.07 -0.43 0.19  
## Unemployment 3.30 0.09 -0.02 0.06  
## Average.life.expectancy 27.00 -0.97 0.70 0.59  
## Constitutional.form 2.00 -1.83 2.47 0.05  
## GDP.per.capita 5.27 -0.20 -0.78 0.12  
## Ec.active.population 43.57 -0.24 0.57 0.71  
## Birth.rate 30.98 1.31 1.25 0.71

f <- f\_ej

Висновок: Для корекції викидів в усіх випадках обраний варіант заповнення граничними значеннями.

set.seed(123)  
library(caTools)

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

split = sample.split(f$Average.life.expectancy, SplitRatio = 0.8)  
f\_train = subset(f, split == TRUE)  
f\_test = subset(f, split == FALSE)  
write.csv2(f\_train, file = "AGE\_train.csv")  
write.csv2(f\_test, file = "AGE\_test.csv")

Висновок: Датасет розподілений на навчальну та тестову вибірки. Результати збережені в окремих файлах.

# Regression

# Simple Linear Regression (one factor – GDP.per.capita)

# Fitting Simple Linear Regression to the Training set

model\_sr <- lm(Average.life.expectancy~GDP.per.capita, f\_train)  
summary(model\_sr)

##   
## Call:  
## lm(formula = Average.life.expectancy ~ GDP.per.capita, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.4594 -1.1630 0.2774 2.2540 5.9363   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 34.781 2.541 13.69 <2e-16 \*\*\*  
## GDP.per.capita 4.429 0.275 16.10 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.219 on 85 degrees of freedom  
## Multiple R-squared: 0.7531, Adjusted R-squared: 0.7502   
## F-statistic: 259.3 on 1 and 85 DF, p-value: < 2.2e-16

Висновок: Обрана змінна значуща, коефіцієнт детермінації =0,75

# Predicting

p\_sr <- predict(model\_sr, f\_test)  
r2\_sr <- 1-sum((f\_train$Average.life.expectancy - predict(model\_sr, f\_train))^2)/sum((f\_train$Average.life.expectancy - mean(f\_train$Average.life.expectancy))^2)   
R2\_sr <- cor(f\_train$Average.life.expectancy, fitted(model\_sr))^2  
  
train\_mse\_sr <- sum((f\_train$Average.life.expectancy-predict(model\_sr, f\_train))^2)/length(f\_train$Average.life.expectancy)   
test\_mse\_sr <- sum((f\_test$Average.life.expectancy-p\_sr)^2)/length(p\_sr)   
train\_mse\_sr

## [1] 10.12092

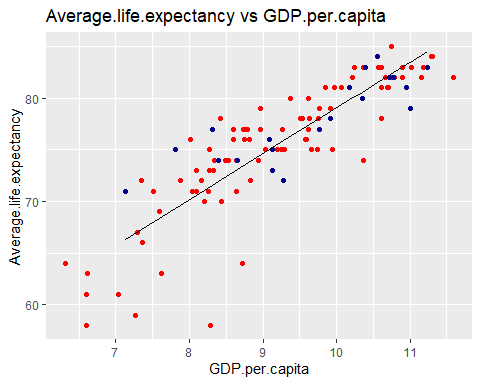
test\_mse\_sr

## [1] 7.557154

Висновок: Значення середньоквадратичної похибки на навчальній вибірці – 10.12092, на тестовій вибірці – 7.557154, тобтонемає перенавчання.

# Visualising

library(ggplot2)   
ggplot()+  
 geom\_point(aes(f\_train$GDP.per.capita, f\_train$Average.life.expectancy), color='red')+  
 geom\_point(aes(f\_test$GDP.per.capita, f\_test$Average.life.expectancy), color='dark blue')+  
 geom\_line(aes(f\_test$GDP.per.capita, p\_sr), color='black')+  
 ggtitle('Average.life.expectancy vs GDP.per.capita')+  
 xlab('GDP.per.capita')+  
 ylab('Average.life.expectancy')



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

# Multiple Linear Regression (many factors)

# All factors

model\_mr <- lm(data = f\_train, Average.life.expectancy ~ .)  
summary(model\_mr)

##   
## Call:  
## lm(formula = Average.life.expectancy ~ ., data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.5882 -1.4386 0.3019 1.8463 5.2034   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 53.34250 6.38151 8.359 1.60e-12 \*\*\*  
## GDP 0.13515 0.16976 0.796 0.4283   
## Unemployment -0.75110 0.54927 -1.367 0.1753   
## Constitutional.form 0.91508 0.54579 1.677 0.0975 .   
## GDP.per.capita 2.97588 0.41290 7.207 2.83e-10 \*\*\*  
## Ec.active.population -0.06019 0.05986 -1.006 0.3176   
## Birth.rate -0.37483 0.06156 -6.089 3.74e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.612 on 80 degrees of freedom  
## Multiple R-squared: 0.847, Adjusted R-squared: 0.8355   
## F-statistic: 73.81 on 6 and 80 DF, p-value: < 2.2e-16

Висновок: Значущими є Birth.rate та GDP.per.capita, коефіцієнт детермінації дорівнює 0,84.

# Optimized model

# Оптимізуємо модель за допомогою виключення з неї змінних GDP, Unemployment, та Constitutional.form та Ec.active.population

model\_opt <- lm(data=f\_train, Average.life.expectancy~Birth.rate+GDP.per.capita)  
summary(model\_opt)

##   
## Call:  
## lm(formula = Average.life.expectancy ~ Birth.rate + GDP.per.capita,   
## data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.0081 -1.1806 0.3742 1.6632 5.2946   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 55.2794 3.8201 14.471 < 2e-16 \*\*\*  
## Birth.rate -0.3561 0.0555 -6.416 7.87e-09 \*\*\*  
## GDP.per.capita 2.8184 0.3382 8.335 1.29e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.652 on 84 degrees of freedom  
## Multiple R-squared: 0.8343, Adjusted R-squared: 0.8304   
## F-statistic: 211.5 on 2 and 84 DF, p-value: < 2.2e-16

Висновок: усі змінні значущі, коефіцієнт детермінації = 0,83.

# Prediction

p\_mr <- predict(model\_opt,f\_test)  
train\_mse\_opt <- sum((f\_train$Average.life.expectancy-predict(model\_opt,f\_train))^2)/length(f\_train$Average.life.expectancy)  
tast\_mse\_opt <- sum((f\_test$Average.life.expectancy-p\_mr)^2)/length(p\_mr)  
train\_mse\_opt

## [1] 6.792481

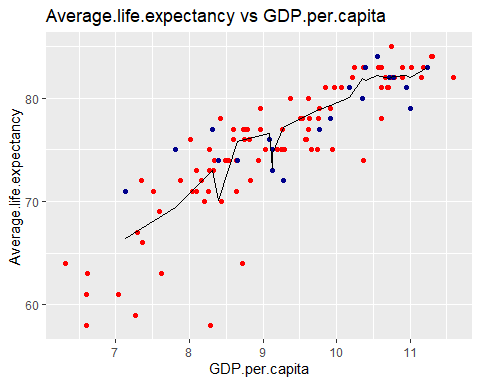
tast\_mse\_opt

## [1] 6.948426

Висновок: значення середньоквадратичної помилки на навчальній вибірці – 6.792481, на тестовій вибірці – 6.948426, є перенавчання.

# Visualising

ggplot()+  
 geom\_point(aes(f\_train$GDP.per.capita, f\_train$Average.life.expectancy), color='red')+  
 geom\_point(aes(f\_test$GDP.per.capita, f\_test$Average.life.expectancy), color='dark blue')+  
 geom\_line(aes(f\_test$GDP.per.capita, p\_mr), color='black')+  
 ggtitle('Average.life.expectancy vs GDP.per.capita')+  
 xlab('GDP.per.capita')+  
 ylab('Average.life.expectancy')



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

# Polynomial Linear Regression (one factor - GDP.per.capita)

# Features extending

f\_train\_poly <-f\_train[,c('Average.life.expectancy','GDP.per.capita')]  
f\_test\_poly <-f\_test[,c('Average.life.expectancy','GDP.per.capita')]  
f\_train\_poly$GDP.per.capita2 <-f\_train\_poly$GDP.per.capita^2  
f\_train\_poly$GDP.per.capita3 <-f\_train\_poly$GDP.per.capita^3  
f\_test\_poly$GDP.per.capita2 <-f\_test\_poly$GDP.per.capita^2  
f\_test\_poly$GDP.per.capita3 <-f\_test\_poly$GDP.per.capita^3

Висновок: Ми додали значення GDP.per.capita^2 та GDP.per.capita^3

# 3 powers

model\_pr <-lm(data = f\_train\_poly, Average.life.expectancy~GDP.per.capita2+GDP.per.capita3)  
summary(model\_pr)

##   
## Call:  
## lm(formula = Average.life.expectancy ~ GDP.per.capita2 + GDP.per.capita3,   
## data = f\_train\_poly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.811 -1.149 0.375 1.771 5.627   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.46388 5.54936 6.030 4.23e-08 \*\*\*  
## GDP.per.capita2 1.03092 0.20000 5.155 1.66e-06 \*\*\*  
## GDP.per.capita3 -0.05698 0.01436 -3.968 0.000152 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.109 on 84 degrees of freedom  
## Multiple R-squared: 0.7724, Adjusted R-squared: 0.767   
## F-statistic: 142.6 on 2 and 84 DF, p-value: < 2.2e-16

Висновок: змінні GDP.per.capita^2 та GDP.per.capita^3 значущі, коефіцієнт детермінації = 0.77.

# Predicting

p\_pr<- predict(model\_pr, f\_test\_poly)  
  
train\_mse\_poly <-sum((f\_train\_poly$Average.life.expectancy-predict(model\_pr,f\_train\_poly))^2)/  
 length(f\_train\_poly$Average.life.expectancy)  
test\_mse\_poly <-sum((f\_test\_poly$Average.life.expectancy-p\_pr)^2)/length(p\_pr)  
train\_mse\_poly

## [1] 9.330371

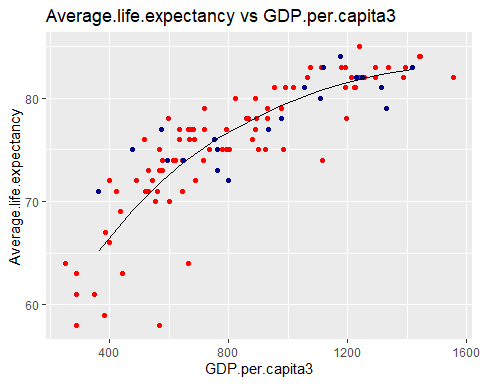
test\_mse\_poly

## [1] 7.944063

Висновок: значення середньоквадратичної помилки на навчальній вибірці – 9.330371, на тестовій вибірці – 7.944063, тобто немає перенавчання.

# Visualising

ggplot()+  
 geom\_point(aes(f\_train\_poly$GDP.per.capita3, f\_train\_poly$Average.life.expectancy), color='red')+  
 geom\_point(aes(f\_test\_poly$GDP.per.capita3, f\_test\_poly$Average.life.expectancy), color='dark blue')+  
 geom\_line(aes(f\_test\_poly$GDP.per.capita3, p\_pr), color='black')+  
 ggtitle('Average.life.expectancy vs GDP.per.capita3')+  
 xlab('GDP.per.capita3')+  
 ylab('Average.life.expectancy')



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

#Saving results

fit <-data.frame(p\_sr,p\_mr,p\_pr)  
write.csv2(fit, file = "AGE\_fit.csv")

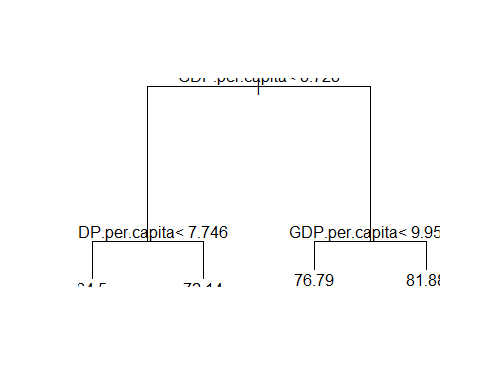
Висновок: результати моделювання збережені у файлі. Порівнюючі коефіцієнти детермінації та середньоквадратичнку помилку можемо побачити, що краща модель - поліноміальна. Але всї побудовані нами моделі перенавчені.

# Decision Tree Regression

library(rpart)

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

model\_dt <- rpart(Average.life.expectancy ~ GDP.per.capita, f\_train, control = rpart.control(minsplit = 30))  
plot(model\_dt)  
text(model\_dt)



Висновок: побудовано дерево рішень, екзогенна змінна – GDP.per.capita

# Predicting

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

## [1] 9.712351

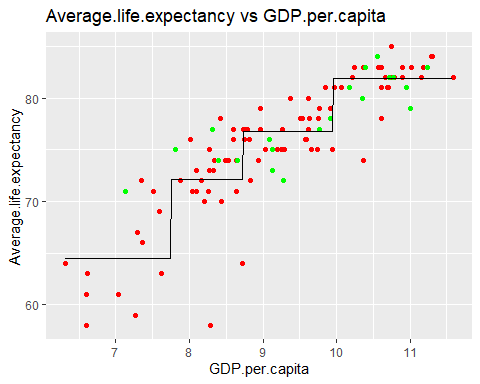
test\_mse\_dt

## [1] 7.20737

Висновок: значення середньоквадратичної помилки на навчальній вибірці – 9.712351, на тестовій вибірці – 7.20737, перенавчання немає.

# Visualising

library(ggplot2)  
x\_grid <- seq(min(f\_train$GDP.per.capita), max(f\_train$GDP.per.capita), 0.01)  
ggplot() +  
 geom\_point(aes(f\_train$GDP.per.capita, f\_train$Average.life.expectancy),colour = 'red') +  
 geom\_point(aes(f\_test$GDP.per.capita, f\_test$Average.life.expectancy),colour = 'green') +  
 geom\_line(aes(x\_grid, predict(model\_dt, data.frame(GDP.per.capita = x\_grid))),colour = 'black') +  
 ggtitle('Average.life.expectancy vs GDP.per.capita') +  
 xlab('GDP.per.capita') +  
 ylab('Average.life.expectancy')



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:psych':  
##   
## outlier

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

## The following object is masked from 'package:dplyr':  
##   
## combine

set.seed(1234)  
model\_rf = randomForest(x = f\_train['GDP.per.capita'],  
 y = f\_train$Average.life.expectancy,  
 ntree = 2)

# Predicting

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

## [1] 5.251832

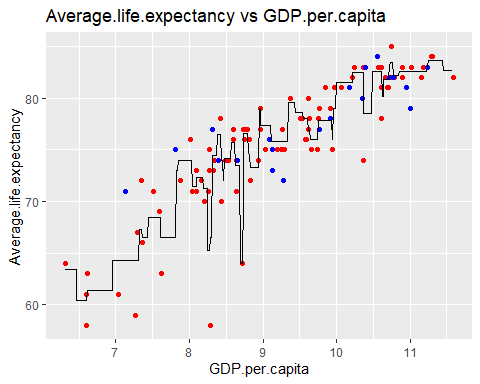
test\_mse\_rf

## [1] 9.949549

Висновок: значення середньоквадратичної помилки на навчальній вибірці – 5.251832, на тестовій вибірці – 9.949549 покращилися, є перенавчання.

# Visualising

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



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

# Saving results

fit <- read.csv2('AGE\_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 6 82.47448 82.07160 81.86583 81.87500 82.16667  
## 2 11 73.12811 75.76733 73.77011 72.13636 73.45000  
## 3 22 82.24511 81.92563 81.74037 81.87500 81.79167  
## 4 24 75.04304 76.62995 75.85982 76.79310 77.40000  
## 5 30 79.85848 80.05066 80.18878 81.87500 81.50000  
## 6 32 83.28827 82.23344 82.27557 81.87500 82.56667

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

Висновок:результати моделювання збережені у файлі.