***Лабораторна робота №2. Підготовка даних***

***Лабораторна робота №3. Побудова моделей лінійної регресії***

***Лабораторна робота №4. Побудова моделей нелінійної регресії***

***Лабораторна робота. Побудова нейронної мережі***

студентки групи ЕП-61

Рибакової Анастасії Сергіївни

Мета лабораторних робіт – підготовка даних (збір, кодування, обчислення статистик, заповнення пропусків, видалення викидів, шкалювання та/чи логарифмування даних), побудова моделі простої лінійної регресії, побудова моделі багатофакторної лінійної регресії з усіма змінними та оптимізованої багатофакторної моделі (за виключенням декількох незначущих змінних), побудова поліноміальної моделі, дерева рішень, випадкового лісу та нейронної мережі.

Для виконання лабораторних робіт було обрано датасет на медичну тему. Розглянемо як смертність від різних хвороб, освіта та витрати на медицину впливають на тривалість життя в різних країнах. Дані зібрані по країнах за 2015 рік.

Обрані показники:

1. Тривалість життя (Life)
2. Смертність дорослих (AM)
3. Дитяча смертність (ID)
4. Гепатит Б (HepB)
5. Кір(Measles)
6. ІМТ (BMI)
7. Смертність дітей до 5 років (U5D)
8. Поліомієліт (Polio)
9. Загальні витрати на медицину (TotalExpend)
10. Дифтерія (Diphth)
11. ВІЛ / СНІД,% (HIV)
12. ВВП (GDP)
13. Чисельність населення (P)
14. Виснаження 1-19 років (TH119)
15. Виснаження 5-9 років (TH59)
16. Навчання в школі(S)

Rybakova1,2,3,4

## “DATA PREPARATION”

# Download the data

#Download file to the table. Source file is 'D.csv'  
f <- read.csv2('D.csv', header = TRUE, encoding = 'UNICOD')  
f <- f[,-1]  
#Connect library  
library (dplyr)

##   
## 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

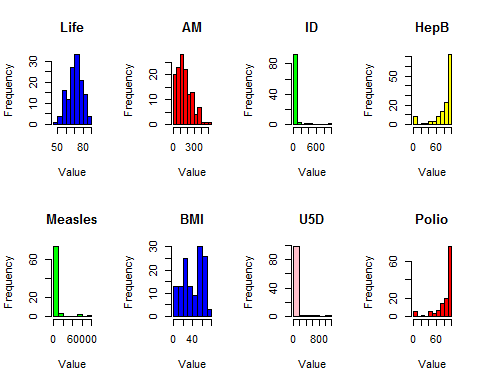
#Have a look at the data  
glimpse(f)

## Observations: 132  
## Variables: 16  
## $ Life <dbl> 59.9, 77.5, 75.4, 51.7, 76.2, 74.6, 73.5, 82.7, 81...  
## $ AM <dbl> 271, 8, 11, 348, 118, 12, 132, 6, 66, 119, 132, 19...  
## $ ID <dbl> 64, NA, 21, 67, 8, 1, 1, 1, NA, 5, 98, NA, NA, NA,...  
## $ HepB <dbl> 62, 98, 95, 64, 94, 93, 94, 91, 98, 94, 97, 97, 98...  
## $ Measles <dbl> 492, NA, NA, 11699, 1, 13, 2, 340, 117, NA, 289, 6...  
## $ BMI <dbl> 18.6, 57.2, 58.4, 22.7, 62.2, 54.1, 51.2, 66.1, 57...  
## $ U5D <dbl> 86, 1, 24, 101, 9, 1, 1, 1, NA, 6, 121, NA, 1, NA,...  
## $ Polio <dbl> 58, 98, 95, 68, 92, 95, 96, 92, 98, 97, 97, 97, 99...  
## $ TotalExpend <dbl> 8.18, 5.88, 7.21, 3.31, 4.79, 4.48, 4.56, 9.42, 11...  
## $ Diphth <dbl> 62, 98, 95, 64, 94, 93, 94, 92, 98, 94, 97, 97, 99...  
## $ HIV <dbl> 0.1, 0.1, 0.1, 2.0, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, ...  
## $ GDP <dbl> 613, 4576, 548, 479, 12245, 3995, 3218, 62215, 513...  
## $ P <dbl> 327582, 288914, 39113313, 2692466, 42981515, 29622...  
## $ TH119 <dbl> 17.5, 1.3, 5.8, 8.3, 0.9, 2.1, 2.1, 0.6, 2.0, 2.9,...  
## $ TH59 <dbl> 0.5, 0.8, 0.7, 0.5, 0.8, 0.7, 0.7, 0.9, 0.9, 0.8, ...  
## $ S <dbl> 10.0, 14.2, 14.4, 11.4, 17.3, 12.7, 12.3, 20.4, 15...

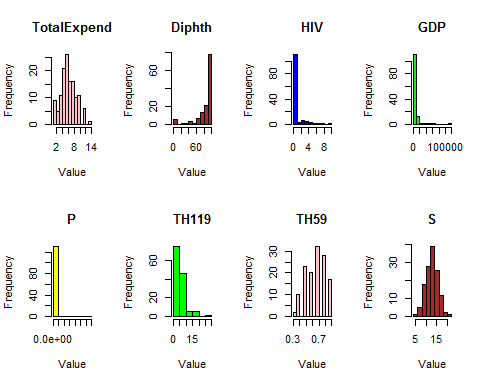
#Висновок: 132 спостережень та 16 змінних.

# Visualising

library(ggplot2)  
par(mfrow = c(2, 4))   
hist(f$Life, col = 'blue', main = 'Life', xlab = 'Value')  
hist(f$AM, col = 'red', main = 'AM', xlab = 'Value')  
hist(f$ID, col = 'green', main = 'ID', xlab = 'Value')  
hist(f$HepB, col = 'yellow', main = 'HepB', xlab = 'Value')  
hist(f$Measles, col = 'green', main = 'Measles', xlab = 'Value')  
hist(f$BMI, col = 'blue', main = 'BMI', xlab = 'Value')  
hist(f$U5D, col = 'pink', main = 'U5D', xlab = 'Value')  
hist(f$Polio, col = 'red', main = 'Polio', xlab = 'Value')

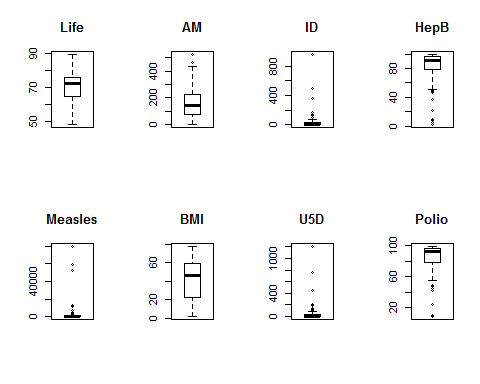


hist(f$TotalExpend, col = 'pink', main = 'TotalExpend', xlab = 'Value')  
hist(f$Diphth, col = 'brown', main = 'Diphth', xlab = 'Value')  
hist(f$HIV, col = 'blue', main = 'HIV', xlab = 'Value')  
hist(f$GDP, col = 'green', main = 'GDP', xlab = 'Value')  
hist(f$P, col = 'yellow', main = 'P', xlab = 'Value')  
hist(f$TH119, col = 'green', main = 'TH119', xlab = 'Value')  
hist(f$TH59, col = 'pink', main = 'TH59', xlab = 'Value')  
hist(f$S, col = 'brown', main = 'S', xlab = 'Value')

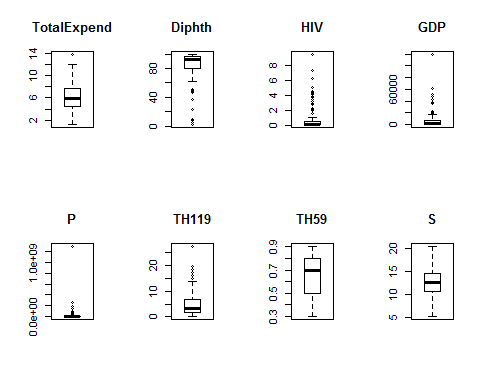


#Висновок: розподіли змінних Life, AM, BMI, TotalExpend, TH59, S є близькими до нормального.

#Box-plot  
par(mfrow = c(2, 4))  
boxplot(f$Life, main = 'Life')  
boxplot(f$AM, main = 'AM')  
boxplot(f$ID, main = 'ID')  
boxplot(f$HepB, main = 'HepB')  
boxplot(f$Measles, main = 'Measles')  
boxplot(f$BMI, main = 'BMI')  
boxplot(f$U5D, main = 'U5D')  
boxplot(f$Polio, main = 'Polio')



boxplot(f$TotalExpend, main = 'TotalExpend')  
boxplot(f$Diphth, main = 'Diphth')  
boxplot(f$HIV, main = 'HIV')  
boxplot(f$GDP, main = 'GDP')  
boxplot(f$P, main = 'P')  
boxplot(f$TH119, main = 'TH119')  
boxplot(f$TH59, main = 'TH59')  
boxplot(f$S, main = 'S')



#Висновок: змінні AM, ID, HepB, Measles, U5D, Polio, TotalExpend, Diphth, HIV, GDP, P, TH119 мають викиди.

# Descriptive statistics

library (psych)

##   
## Attaching package: 'psych'

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

d=describe(f) #Інакше describe не запускається  
#Висновок: Основні показники описової статистики за кожною змінною вказують на те, що є пропущені значення в змінних ID, Measles, U5D. Змінні AM, ID, HepB, Measles, U5D, Polio, TotalExpend, Diphth, HIV, GDP, P, TH119 мають викиди.

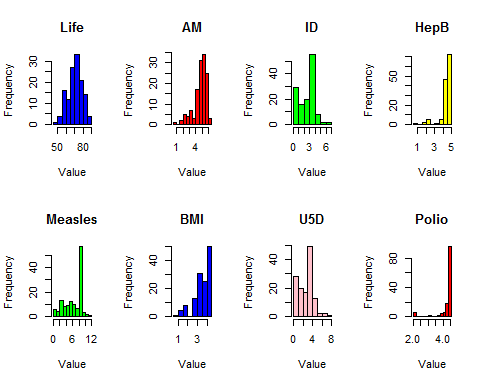
# Fill n/a with average

f\_fill <- f  
f\_fill$ID <- ifelse(is.na(f$ID),round(mean(f$ID,na.rm = TRUE)),f$ID)  
f\_fill$Measles <- ifelse(is.na(f$Measles),round(mean(f$Measles,na.rm = TRUE)),f$Measles)  
f\_fill$U5D <- ifelse(is.na(f$U5D),round(mean(f$U5D,na.rm = TRUE)),f$U5D)  
f <- f\_fill  
h=describe(f\_fill)  
#Висновок: пропуски заповнені середніми значеннями змінних.

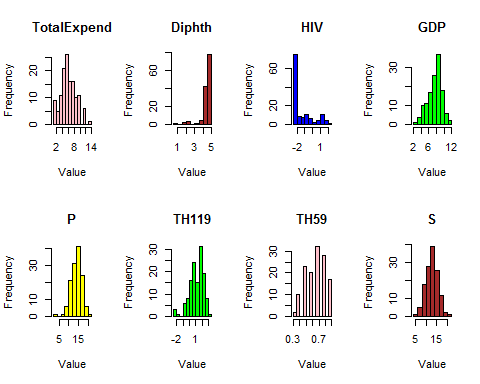
# Log

f\_log <- f  
f\_log$AM <- log(f$AM, )  
f\_log$ID <- log(f$ID, )  
f\_log$HepB <- log(f$HepB, )  
f\_log$Measles <- log(f$Measles, )  
f\_log$BMI <- log(f$BMI, )  
f\_log$U5D <- log(f$U5D, )  
f\_log$Polio <- log(f$Polio, )  
f\_log$Diphth<- log(f$Diphth, )  
f\_log$HIV <- log(f$HIV, )  
f\_log$GDP <- log(f$GDP, )  
f\_log$P <- log(f$P, )  
f\_log$TH119 <- log(f$TH119, )  
f<- f\_log

#Visualising  
par(mfrow = c(2, 4))   
hist(f$Life, col = 'blue', main = 'Life', xlab = 'Value')  
hist(f$AM, col = 'red', main = 'AM', xlab = 'Value')  
hist(f$ID, col = 'green', main = 'ID', xlab = 'Value')  
hist(f$HepB, col = 'yellow', main = 'HepB', xlab = 'Value')  
hist(f$Measles, col = 'green', main = 'Measles', xlab = 'Value')  
hist(f$BMI, col = 'blue', main = 'BMI', xlab = 'Value')  
hist(f$U5D, col = 'pink', main = 'U5D', xlab = 'Value')  
hist(f$Polio, col = 'red', main = 'Polio', xlab = 'Value')

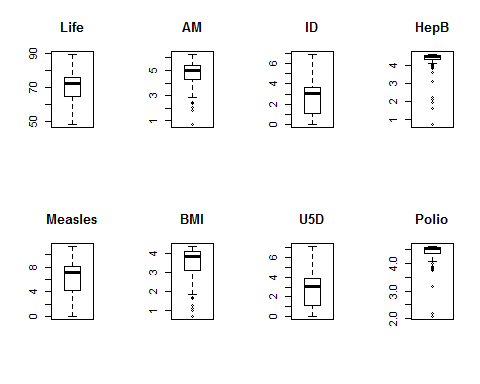


hist(f$TotalExpend, col = 'pink', main = 'TotalExpend', xlab = 'Value')  
hist(f$Diphth, col = 'brown', main = 'Diphth', xlab = 'Value')  
hist(f$HIV, col = 'blue', main = 'HIV', xlab = 'Value')  
hist(f$GDP, col = 'green', main = 'GDP', xlab = 'Value')  
hist(f$P, col = 'yellow', main = 'P', xlab = 'Value')  
hist(f$TH119, col = 'green', main = 'TH119', xlab = 'Value')  
hist(f$TH59, col = 'pink', main = 'TH59', xlab = 'Value')  
hist(f$S, col = 'brown', main = 'S', xlab = 'Value')

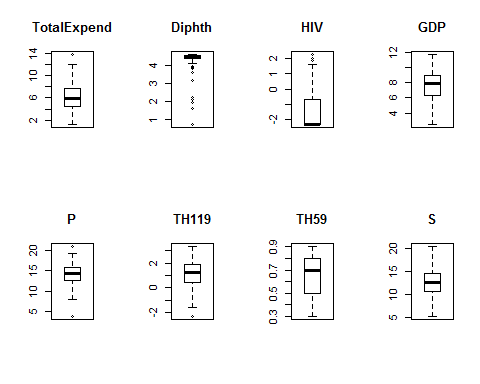


#Висновок: розподіли змінних Life, AM, BMI, TotalExpend, TH59, S є близькими до нормального.

#Box-plot  
par(mfrow = c(2, 4))  
boxplot(f$Life, main = 'Life')  
boxplot(f$AM, main = 'AM')  
boxplot(f$ID, main = 'ID')  
boxplot(f$HepB, main = 'HepB')  
boxplot(f$Measles, main = 'Measles')  
boxplot(f$BMI, main = 'BMI')  
boxplot(f$U5D, main = 'U5D')  
boxplot(f$Polio, main = 'Polio')



boxplot(f$TotalExpend, main = 'TotalExpend')  
boxplot(f$Diphth, main = 'Diphth')  
boxplot(f$HIV, main = 'HIV')  
boxplot(f$GDP, main = 'GDP')  
boxplot(f$P, main = 'P')  
boxplot(f$TH119, main = 'TH119')  
boxplot(f$TH59, main = 'TH59')  
boxplot(f$S, main = 'S')



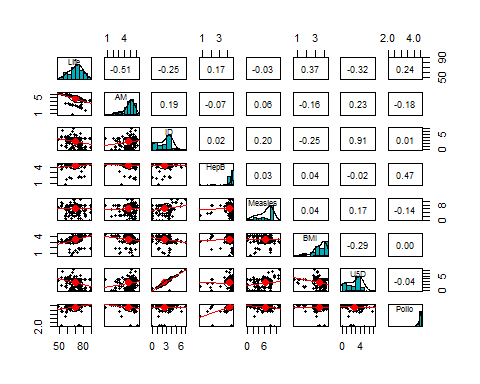
#Висновок: змінні AM, HepB, BMI, Polio, TotalExpend, Diphth, HIV, P, TH119 мають викиди. Змінні ID, Measles, U5D та GDP після логарифмування звільнилися від викидів.

# Replace ejections (outside the three sigma) with boundary values.

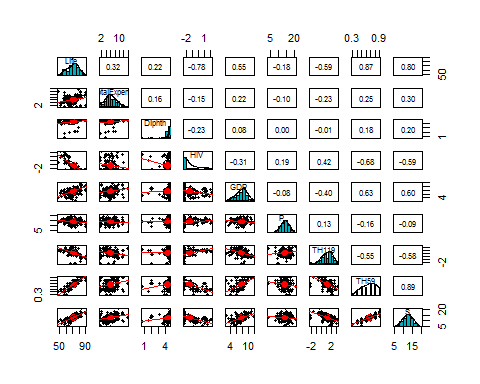
f\_ej <- f  
f\_ej$AM <- ifelse(f$AM < mean(f$AM) + sd(f$AM)\*3, f$AM,mean(f$AM) + sd(f$AM)\*3)  
#f\_ej$ID <- ifelse(f$ID < mean(f$ID) + sd(f$ID)\*3, f$ID,mean(f$ID) + sd(f$ID)\*3)  
f\_ej$HepB <- ifelse(f$HepB < mean(f$HepB) + sd(f$HepB)\*3, f$HepB,mean(f$HepB) + sd(f$HepB)\*3)  
#f\_ej$Measles <- ifelse(f$Measles < mean(f$Measles) + sd(f$Measles)\*3, f$Measles,mean(f$Measles) + sd(f$Measles)\*3)  
#f\_ej$U5D <- ifelse(f$U5D < mean(f$U5D) + sd(f$U5D)\*3, f$U5D,mean(f$U5D) + sd(f$U5D)\*3)  
f\_ej$BMI <- ifelse(f$BMI < mean(f$BMI) + sd(f$BMI)\*3, f$BMI,mean(f$BMI) + sd(f$BMI)\*3)  
f\_ej$Polio <- ifelse(f$Polio < mean(f$Polio) + sd(f$Polio)\*3, f$Polio,mean(f$Polio) + sd(f$Polio)\*3)  
f\_ej$TotalExpend <- ifelse(f$TotalExpend < mean(f$TotalExpend) + sd(f$TotalExpend)\*3, f$TotalExpend,mean(f$TotalExpend) + sd(f$TotalExpend)\*3)  
f\_ej$Diphth <- ifelse(f$Diphth < mean(f$Diphth) + sd(f$Diphth)\*3, f$Diphth,mean(f$Diphth) + sd(f$Diphth)\*3)  
f\_ej$HIV <- ifelse(f$HIV < mean(f$HIV) + sd(f$HIV)\*3, f$HIV,mean(f$HIV) + sd(f$HIV)\*3)  
#f\_ej$GDP <- ifelse(f$GDP < mean(f$GDP) + sd(f$GDP)\*3, f$GDP,mean(f$GDP) + sd(f$GDP)\*3)  
f\_ej$P <- ifelse(f$P < mean(f$P) + sd(f$P)\*3, f$P,mean(f$P) + sd(f$P)\*3)  
f\_ej$TH119 <- ifelse(f$TH119 < mean(f$TH119) + sd(f$TH119)\*3, f$TH119,mean(f$TH119) + sd(f$TH119)\*3)  
f <- f\_ej  
#Висновок: для корекції викидів здійснене заповнення граничними значеннями.

# Correlations

f1 <- c(1,2,3,4,5,6,7,8)  
c\_f1 <- f[f1]  
pairs.panels(c\_f1, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



f2 <- c(1,9,10,11,12,13,14,15,16)  
c\_f2 <- f[f2]  
pairs.panels(c\_f2, lm=TRUE, # linear fit  
 method = "pearson", # correlation method  
 hist.col = "#00AFBB"  
 )



#Висновок: Розбито змінні на дві групи (в кожну включено залежну змінну Life). Найбільший взаємозв'язок (модуль>0,7) мають змінні Life та HIV, Life та TH59, Life та S, ID та U5D, TH59 та S.

# Splitting the dataset into the TRAIN set and TEST set

set.seed (104)  
library (caTools)  
split = sample.split (f$Life, SplitRatio = 0.8)  
f\_train = subset (f, split == TRUE)  
f\_test = subset (f, split == FALSE)  
#Write prepared data to the file  
write.csv2 (f\_train, file = "D\_train.csv")  
write.csv2 (f\_test, file = "D\_test.csv")  
#Висновок: датасет розподілений на навчальну та тестову вибірки (105 та 27 спостережень). Результати є в окремих файлах.

## “LINEAR REGRESSION”

# Simple Linear Regression (one factor TH59/S/AM)

#Fitting Simple Linear Regression to the Training set  
model\_sr1 <- lm(Life~TH59, f\_train)  
summary(model\_sr1)

##   
## Call:  
## lm(formula = Life ~ TH59, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.9117 -2.8037 0.3003 2.5963 8.0043   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 40.532 1.759 23.04 <2e-16 \*\*\*  
## TH59 44.960 2.548 17.64 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.968 on 103 degrees of freedom  
## Multiple R-squared: 0.7514, Adjusted R-squared: 0.749   
## F-statistic: 311.3 on 1 and 103 DF, p-value: < 2.2e-16

model\_sr2 <- lm(Life~S, f\_train)  
summary(model\_sr2)

##   
## Call:  
## lm(formula = Life ~ S, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.4161 -3.2157 0.4096 3.6083 10.1066   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.212 2.186 19.31 <2e-16 \*\*\*  
## S 2.250 0.168 13.39 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.807 on 103 degrees of freedom  
## Multiple R-squared: 0.6352, Adjusted R-squared: 0.6317   
## F-statistic: 179.4 on 1 and 103 DF, p-value: < 2.2e-16

model\_sr3 <- lm(Life~HIV, f\_train)  
summary(model\_sr3)

##   
## Call:  
## lm(formula = Life ~ HIV, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -15.1216 -3.2683 -0.2513 2.7784 13.9784   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 64.3324 0.7024 91.59 <2e-16 \*\*\*  
## HIV -4.6423 0.3649 -12.72 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.963 on 103 degrees of freedom  
## Multiple R-squared: 0.6111, Adjusted R-squared: 0.6074   
## F-statistic: 161.9 on 1 and 103 DF, p-value: < 2.2e-16

#Висновок: розглянуто три моделі. Обрана змінна найбільш значуща у першій моделі (ТН59), коефіцієнт детермінації 0,75. Тому далі працюємо саме з нею.

#Predicting  
p\_sr <- predict(model\_sr1, f\_test)  
r2\_sr <- 1-sum((f\_train$Life - predict(model\_sr1, f\_train))^2)/sum((f\_train$Life - mean(f\_train$Life))^2)  
R2\_sr <- cor(f\_train$Life, fitted(model\_sr1))^2  
train\_mse\_sr <- sum((f\_train$Life-predict(model\_sr1, f\_train))^2)/length(f\_train$Life)  
test\_mse\_sr <- sum((f\_test$Life-p\_sr)^2)/length(p\_sr)  
r2\_sr

## [1] 0.751409

R2\_sr

## [1] 0.751409

train\_mse\_sr

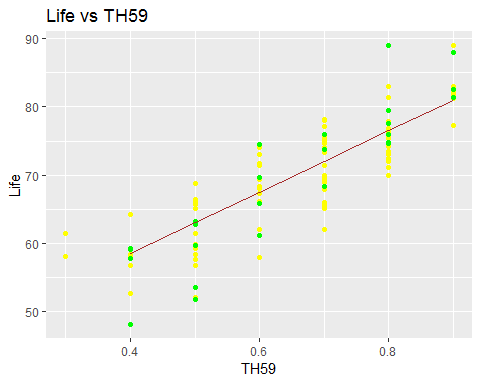
## [1] 15.44708

test\_mse\_sr

## [1] 27.5074

#Висновок: Розраховані коефіцієнти детермінації. Значення середньоквадратичної похибки на навчальній вибірці – 15.44708, на тестовій вибірці – 27.50741, тобто є перенавчання.

#Visualising  
library(ggplot2)  
ggplot() +  
 geom\_point(aes(f\_train$TH59, f\_train$Life),colour = 'yellow') +  
 geom\_point(aes(f\_test$TH59, f\_test$Life),colour = 'green') +  
 geom\_line(aes(f\_test$TH59, p\_sr),colour = 'brown') +  
 ggtitle('Life vs TH59') +  
 xlab('TH59') +  
 ylab('Life')



#Висновок: на графіку жовтим позначені точки навчальної вибірки, зеленим – точки тестової вибірки, а коричневим – модельні значення.

# Multiple Linear Regression

#All factors  
model\_mr <- lm(data = f\_train, Life ~ .)  
summary(model\_mr)

##   
## Call:  
## lm(formula = Life ~ ., data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.1129 -1.6732 0.0162 1.7782 7.7309   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 50.8744319 4.4133047 11.528 < 2e-16 \*\*\*  
## AM -0.4814735 0.3452805 -1.394 0.166655   
## ID 1.1120040 0.5302865 2.097 0.038832 \*   
## HepB -0.1885569 1.0078723 -0.187 0.852021   
## Measles -0.0006517 0.1234352 -0.005 0.995799   
## BMI -0.3319442 0.4547591 -0.730 0.467347   
## U5D -1.3628767 0.4890457 -2.787 0.006506 \*\*   
## Polio 0.1546820 0.6951637 0.223 0.824426   
## TotalExpend 0.4223919 0.1352501 3.123 0.002415 \*\*   
## Diphth 0.4104275 1.1405039 0.360 0.719801   
## HIV -2.2803890 0.3436646 -6.636 2.45e-09 \*\*\*  
## GDP 0.2673381 0.2261126 1.182 0.240227   
## P 0.0710293 0.1191216 0.596 0.552504   
## TH119 -0.9182097 0.3953297 -2.323 0.022479 \*   
## TH59 21.3049763 5.2908751 4.027 0.000119 \*\*\*  
## S 0.0364136 0.2582347 0.141 0.888181   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.094 on 89 degrees of freedom  
## Multiple R-squared: 0.8694, Adjusted R-squared: 0.8474   
## F-statistic: 39.51 on 15 and 89 DF, p-value: < 2.2e-16

#Висновок: зміннi HepB, Measles, Polio, Diphth, P, S не значущі, виключимо їх з моделі (оскільки їх p-value, Pr(>|t|) вищі за 5%). Коефіцієнт детермінації помітно зріс, порівняно з лінійною регресією (0,87).

#Optimized model  
model\_opt <- lm(data = f\_train, Life ~ AM + ID + BMI + U5D + TotalExpend + HIV + GDP + TH119 + TH59)  
summary(model\_opt)

##   
## Call:  
## lm(formula = Life ~ AM + ID + BMI + U5D + TotalExpend + HIV +   
## GDP + TH119 + TH59, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.3431 -1.7114 -0.0529 1.7937 7.5983   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 53.2177 3.0627 17.376 < 2e-16 \*\*\*  
## AM -0.4675 0.3297 -1.418 0.159519   
## ID 1.1531 0.5047 2.285 0.024560 \*   
## BMI -0.3701 0.4337 -0.853 0.395666   
## U5D -1.3662 0.4692 -2.912 0.004477 \*\*   
## TotalExpend 0.4314 0.1253 3.443 0.000856 \*\*\*  
## HIV -2.2691 0.3182 -7.132 1.94e-10 \*\*\*  
## GDP 0.2587 0.2149 1.204 0.231625   
## TH119 -0.9130 0.3683 -2.479 0.014950 \*   
## TH59 22.4294 3.7826 5.930 4.89e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.01 on 95 degrees of freedom  
## Multiple R-squared: 0.8681, Adjusted R-squared: 0.8556   
## F-statistic: 69.46 on 9 and 95 DF, p-value: < 2.2e-16

#Висновок: побудовано дев'ятифакторну модель. Найбільш значущі змінні U5D, TotalExpend, HIV, TH59 коефіцієнт детермінації 0,87 (скорегований коефіцієнт детермінації трохи зріс з 0,847 (при включенні всіх факторів) до 0,856).

#Prediction  
p\_mr <- predict(model\_opt, f\_test)  
train\_mse\_opt <- sum((f\_train$Life-predict(model\_opt, f\_train))^2)/length(f\_train$Life)  
test\_mse\_opt <- sum((f\_test$Life-p\_mr)^2)/length(p\_mr)  
train\_mse\_opt

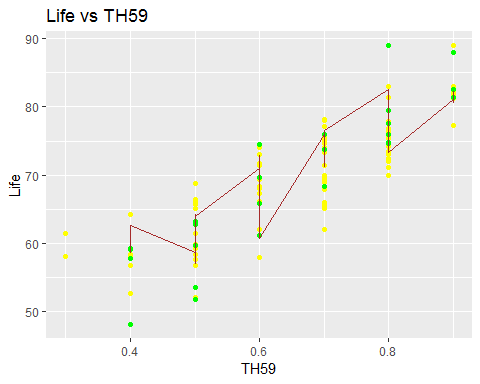
## [1] 8.196973

test\_mse\_opt

## [1] 18.3941

#Висновок: значення середньоквадратичних помилок покращились: на навчальній вибірці – 8.196973, на тестовій вибірці – 18.3941, тобто є перенавчання.

#Visualising  
ggplot() +  
 geom\_point(aes(f\_train$TH59, f\_train$Life),colour = 'yellow') +  
 geom\_point(aes(f\_test$TH59, f\_test$Life),colour = 'green') +  
 geom\_line(aes(f\_test$TH59, p\_mr),colour = 'brown') +  
 ggtitle('Life vs TH59') +  
 xlab('TH59') +  
 ylab('Life')



#Висновок: на графіку жовтим позначені точки навчальної вибірки, зеленим – точки тестової вибірки, а коричневим – модельні значення.

## “NONLINEAR REGRESSION”

# Polynomial Linear Regression (one factor - TH59)

#Features extending  
f\_train\_poly <- f\_train[,c('Life','TH59')]  
f\_test\_poly <- f\_test[,c('Life','TH59')]  
f\_train\_poly$TH592 <- f\_train\_poly$TH59^2  
f\_train\_poly$TH593 <- f\_train\_poly$TH59^3  
f\_test\_poly$TH592 <- f\_test\_poly$TH59^2  
f\_test\_poly$TH593 <- f\_test\_poly$TH59^3  
#Висновок: додано змінні TH59^2 та TH59^3.  
#3 powers  
model\_pr <- lm(data = f\_train\_poly, Life ~ TH592 + TH593)  
summary(model\_pr)

##   
## Call:  
## lm(formula = Life ~ TH592 + TH593, data = f\_train\_poly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.7907 -2.5699 0.2093 2.9703 7.2974   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 52.097 2.524 20.642 < 2e-16 \*\*\*  
## TH592 51.459 17.237 2.985 0.00355 \*\*   
## TH593 -16.565 16.659 -0.994 0.32242   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.947 on 102 degrees of freedom  
## Multiple R-squared: 0.7564, Adjusted R-squared: 0.7516   
## F-statistic: 158.4 on 2 and 102 DF, p-value: < 2.2e-16

#Висновок: змінна TH59^2 значуща, а TH59^3 - не значуща, коефіцієнт детермінації зменшився – 0,75.

#Predicting  
p\_pr <- predict(model\_pr, f\_test\_poly)  
train\_mse\_poly <- sum((f\_train\_poly$Life-predict(model\_pr, f\_train\_poly))^2)/length(f\_train\_poly$Life)  
test\_mse\_poly <- sum((f\_test\_poly$Life-p\_pr)^2)/length(p\_pr)  
train\_mse\_poly

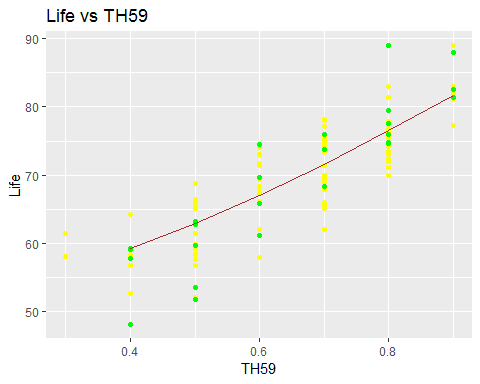
## [1] 15.13735

test\_mse\_poly

## [1] 27.86672

#Висновок: значення середньоквадратичних помилок зросли на навчальній вибірці – 15.13735, на тестовій вибірці – 27.86672, тобто є перенавчання.

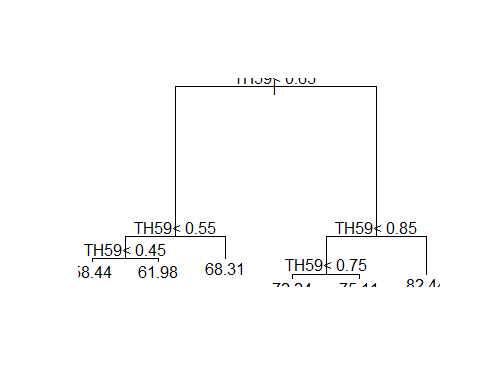
#Visualising  
ggplot() +  
 geom\_point(aes(f\_train\_poly$TH59, f\_train\_poly$Life),colour = 'yellow') +  
 geom\_point(aes(f\_test\_poly$TH59, f\_test\_poly$Life),colour = 'green') +  
 geom\_line(aes(f\_test\_poly$TH59, p\_pr),colour = 'brown') +  
 ggtitle('Life vs TH59') +  
 xlab('TH59') +  
 ylab('Life')



#Висновок: на графіку жовтим позначені точки навчальної вибірки, зеленим – точки тестової вибірки, а коричневим – модельні значення.

# Decision Tree Regression

library(rpart)  
model\_dt <- rpart(Life ~ TH59, f\_train, control = rpart.control(minsplit = 10))  
plot(model\_dt)  
text(model\_dt)



#Висновок: побудовано дерево рішень, екзогенна змінна – TH59.

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

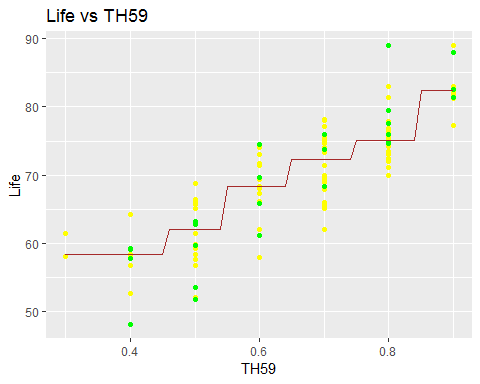
## [1] 13.90709

test\_mse\_dt

## [1] 26.28905

#Висновок: значення середньоквадратичної похибки на навчальній вибірці – 13.90709, на тестовій вибірці – 26.28905. Модель перенавчено.

#Visualising  
library(ggplot2)  
x\_grid <- seq(min(f\_train$TH59), max(f\_train$TH59), 0.01)  
ggplot() +  
 geom\_point(aes(f\_train$TH59, f\_train$Life),colour = 'yellow') +  
 geom\_point(aes(f\_test$TH59, f\_test$Life),colour = 'green') +  
 geom\_line(aes(x\_grid, predict(model\_dt, data.frame(TH59 = x\_grid))),colour = 'brown') +  
 ggtitle('Life vs TH59') +  
 xlab('TH59') +  
 ylab('Life')



#Висновок: на графіку жовтим позначені точки навчальної вибірки, зеленим – точки тестової вибірки, а коричневим – модельні значення.

# Random forest

library(randomForest)

## 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(1047)  
model\_rf = randomForest(x = f\_train['TH59'],  
 y = f\_train$Life,  
 ntree = 26)  
#Висновок: побудовано віпадковий ліс із 26 дерев, екзогенна змінна – TH59.  
  
#Predicting  
p\_rf <- predict(model\_rf, f\_test)  
train\_mse\_rf <- sum((f\_train$Life-predict(model\_rf, f\_train))^2)/length(f\_train$Life)  
test\_mse\_rf <- sum((f\_test$Life-p\_rf)^2)/length(p\_rf)  
train\_mse\_rf

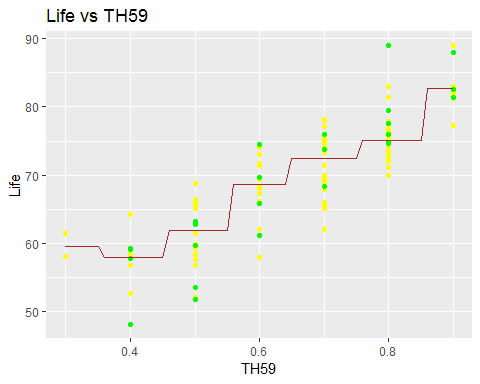
## [1] 13.90755

test\_mse\_rf

## [1] 25.59485

#Висновок: значення середньоквадратичної похибки на навчальній вибірці – 13.90755, на тестовій вибірці – 25.59485. Модель перенавчено.

#Visualising  
ggplot() +  
 geom\_point(aes(f\_train$TH59, f\_train$Life),colour = 'yellow') +  
 geom\_point(aes(f\_test$TH59, f\_test$Life),colour = 'green') +  
 geom\_line(aes(x\_grid, predict(model\_rf, data.frame(TH59 = x\_grid))),colour = 'brown') +  
 ggtitle('Life vs TH59') +  
 xlab('TH59') +  
 ylab('Life')



#Висновок: на графіку жовтим позначені точки навчальної вибірки, зеленим – точки тестової вибірки, а коричневим – модельні значення.

## “NN”

# Features Scaling

f\_train\_sc <- as.data.frame(scale(f\_train))  
f\_test\_sc <- as.data.frame(scale(f\_test))  
head (f\_train\_sc)

## Life AM ID HepB Measles BMI  
## 1 -1.3767581 0.87598825 1.0447812 -0.1662164 0.05949332 -0.6826452  
## 3 0.5801623 -2.17133541 0.3172438 0.4139459 0.78545409 0.6997304  
## 5 0.6811646 0.08526655 -0.3128326 0.3995594 -2.27262084 0.7758947  
## 6 0.4791599 -2.08858468 -1.6704465 0.3850190 -1.30758490 0.6073247  
## 7 0.3402817 0.19189385 -1.6704465 0.3995594 -2.01183131 0.5407589  
## 8 1.5018087 -2.74779081 -1.6704465 0.3554632 -0.07953972 0.8493704  
## U5D Polio TotalExpend Diphth HIV GDP  
## 1 1.0788917 -0.5070852 0.9183325 -0.2663054 -0.6810082 -0.6414007  
## 3 0.3151975 0.4062316 0.5334480 0.3661269 -0.6810082 -0.7020198  
## 5 -0.2717001 0.3468381 -0.4267792 0.3504442 -0.6810082 0.9780563  
## 6 -1.5864508 0.4062316 -0.5497835 0.3345938 -0.6810082 0.3723101  
## 7 -1.5864508 0.4256134 -0.5180405 0.3504442 -0.6810082 0.2553420  
## 8 -1.5864508 0.3468381 1.4103498 0.3185720 -0.6810082 1.8571297  
## P TH119 TH59 S  
## 1 -0.5600770 1.6671740 -1.1350449 -0.964543926  
## 3 1.1819626 0.6080863 0.1746223 0.603985795  
## 5 1.2163144 -1.1787788 0.8294559 1.637789475  
## 6 -1.4354610 -0.3662026 0.1746223 -0.002037052  
## 7 0.2314017 -0.3662026 0.1746223 -0.144630663  
## 8 0.1571489 -1.5676282 1.4842895 2.742889960

#Висновок: виконано шкалювання

# Fitting the NN

library(nnet)  
ff\_ap <- nnet(Life ~ AM + ID + U5D + TotalExpend + HIV + GDP + TH59, f\_train\_sc, linout = TRUE ,size = 5, maxit = 10000)  
library(graphics)  
source(file = 'plot.nnet.R')  
plot.nnet(ff\_ap)

## Loading required package: scales

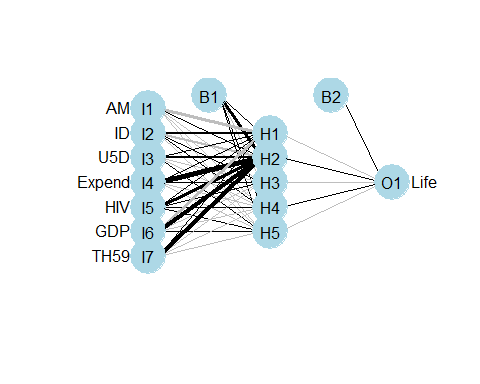
##   
## Attaching package: 'scales'

## The following objects are masked from 'package:psych':  
##   
## alpha, rescale

## Loading required package: reshape

##   
## Attaching package: 'reshape'

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



# Prediction  
p\_y\_train <- predict(ff\_ap, f\_train\_sc)  
p\_y\_test <- predict(ff\_ap, f\_test\_sc)  
#Висновок: побудовано двошарову нейронну мережу для прогнозування Life.

# Invert the effect of the scale function  
sc\_y\_train <- scale(f\_train$Life)  
y\_train <- DMwR::unscale(p\_y\_train, sc\_y\_train)  
sc\_y\_test <- scale(f\_test$Life)  
y\_test <- DMwR::unscale(p\_y\_test, sc\_y\_test)  
# MSE  
train\_mse\_nn <- sum((f\_train$Life-y\_train)^2)/length(f\_train$Life)  
test\_mse\_nn <- sum((f\_test$Life-y\_test)^2)/length(f\_test$Life)  
train\_mse\_nn

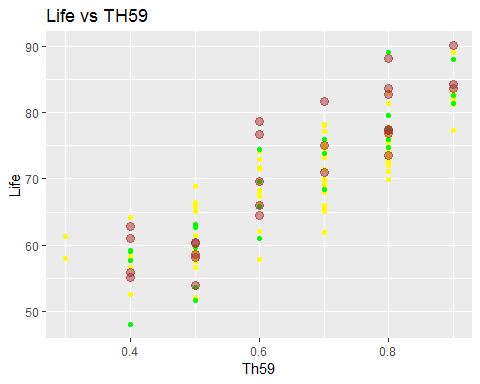
## [1] 1.233277

test\_mse\_nn

## [1] 21.02051

#Висновок: середньоквадратична помилка на навчальній вибірці –1.244452, на тестовій вибірці – 38.51775. Модель перенавчено.

# Visualising  
library(ggplot2)  
ggplot() +  
 geom\_point(aes(f\_train$TH59, f\_train$Life),colour = 'yellow') +  
 geom\_point(aes(f\_test$TH59, f\_test$Life),colour = 'green') +  
 geom\_point(aes(f\_test$TH59, y\_test),colour = 'brown', size = 3, alpha=0.5) +  
 ggtitle('Life vs TH59') +  
 xlab('Th59') +  
 ylab('Life')



#Висновок: на графіку жовтим позначені точки навчальної вибірки, зеленим – точки тестової вибірки, а коричневим – модельні значення.Видно, що прогноз є досить чітким.

#sr  
train\_mse\_sr

## [1] 15.44708

test\_mse\_sr

## [1] 27.5074

#opt  
train\_mse\_opt

## [1] 8.196973

test\_mse\_opt

## [1] 18.3941

#polynom  
train\_mse\_poly

## [1] 15.13735

test\_mse\_poly

## [1] 27.86672

#dt  
train\_mse\_dt

## [1] 13.90709

test\_mse\_dt

## [1] 26.28905

#rf  
train\_mse\_rf

## [1] 13.90755

test\_mse\_rf

## [1] 25.59485

#nn  
train\_mse\_nn

## [1] 1.233277

test\_mse\_nn

## [1] 21.02051

# Saving results

fin <- data.frame(p\_sr, p\_mr, p\_pr, p\_dt, p\_rf)  
write.csv2(fin, file = "D\_fin.csv")  
#Висновок: результати моделювання збережені у файлі.

# Compare models

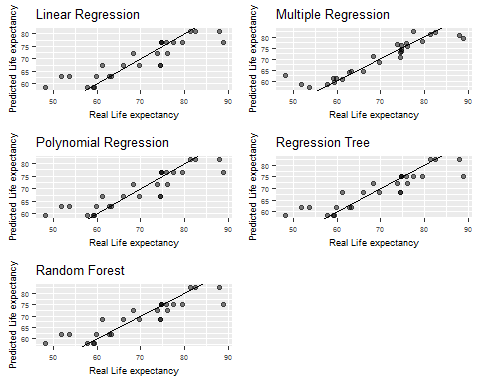
g\_sr <- ggplot(fin, aes(x=f\_test$Life, y=p\_sr)) +   
 geom\_abline(intercept=0, slope=1) +  
 geom\_point(alpha=0.5) + labs(title="Linear Regression", x="Real Life expectancy", y="Predicted Life expectancy") +   
 theme(plot.title=element\_text(size=10), axis.title.x=element\_text(size=7), axis.title.y=element\_text(size=7), axis.text.x=element\_text(size=5), axis.text.y=element\_text(size=5)) + theme(legend.position="none")  
g\_mr <- ggplot(fin, aes(x=f\_test$Life, y=p\_mr)) +   
 geom\_abline(intercept=0, slope=1) +  
 geom\_point(alpha=0.5) + labs(title="Multiple Regression", x="Real Life expectancy", y="Predicted Life expectancy") +   
 theme(plot.title=element\_text(size=10), axis.title.x=element\_text(size=7), axis.title.y=element\_text(size=7), axis.text.x=element\_text(size=5), axis.text.y=element\_text(size=5)) + theme(legend.position="none")  
g\_pr <- ggplot(fin, aes(x=f\_test$Life, y=p\_pr)) +   
 geom\_abline(intercept=0, slope=1) +  
 geom\_point(alpha=0.5) + labs(title="Polynomial Regression", x="Real Life expectancy", y="Predicted Life expectancy") +   
 theme(plot.title=element\_text(size=10), axis.title.x=element\_text(size=7), axis.title.y=element\_text(size=7), axis.text.x=element\_text(size=5), axis.text.y=element\_text(size=5)) + theme(legend.position="none")  
g\_dt <- ggplot(fin, aes(x=f\_test$Life, y=p\_dt)) +   
 geom\_abline(intercept=0, slope=1) +  
 geom\_point(alpha=0.5) + labs(title="Regression Tree", x="Real Life expectancy", y="Predicted Life expectancy") +   
 theme(plot.title=element\_text(size=10), axis.title.x=element\_text(size=7), axis.title.y=element\_text(size=7), axis.text.x=element\_text(size=5), axis.text.y=element\_text(size=5)) + theme(legend.position="none")  
g\_rf <- ggplot(fin, aes(x=f\_test$Life, y=p\_rf)) +   
 geom\_abline(intercept=0, slope=1) +  
 geom\_point(alpha=0.5) + labs(title="Random Forest", x="Real Life expectancy", y="Predicted Life expectancy") +   
 theme(plot.title=element\_text(size=10), axis.title.x=element\_text(size=7), axis.title.y=element\_text(size=7), axis.text.x=element\_text(size=5), axis.text.y=element\_text(size=5)) + theme(legend.position="none")  
library(gridExtra)

##   
## Attaching package: 'gridExtra'

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

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

gridExtra::grid.arrange(g\_sr,g\_mr,g\_pr,g\_dt,g\_rf,ncol=2)



# Висновки

#Отже,за даними без пропусків та викидів (частину змінних, розподіл яких є не нормальним, прологарифмовано) було побудовано моделі простої лінійної регресії, багатофакторної лінійної регресії з усіма змінними та дев'ятифакторну (за виключенням декількох незначущих змінних), поліноміальну модель, дерево рішень та випадковий ліс. Усі розглянуті моделі є перенавченими (зміна SplitRatio в межах 0,7-0,8 значно не змінила ситуації).   
#Найменша середньоквадратична помилка на тренувальній вибірці була у нейронній мережі (1,24), а найбільша - при простій лінійній регресії (15,45).  
#Найменша середньоквадратична помилка на тестовій вибірці була у дев'ятифакторній моделі (18,39), а найбільша - у нейронній мережі (38,52).   
#Коефіцієнт детермінації був найбільшим у багатофакторній лінійній регресійній моделі зі всімома змінними (0,88).  
#Хоча в усіх моделях присутнє перенавчання, серед розглянутих варіантів найякіснішою є побудована проста лінійна регресія. Отримані прогнозні значення є досить близькими до реальних.