Лабораторна робота №2. Лінійна регресія

студентки групи ЕП-61

Лузанової О.С.

Мета роботи – побудова лінійної регресії, а саме: використовуюси данні (Лабораторна робота №1) побудувати модель однофакторної лінійної регресії, багатофакторної лінійної регресії, поліноміальної лінійної регресії на тренувальній виборці та тестовій; зробити прогнози та проаналізувати середньоквадратичні помилки.

Після виконання роботи студент повинен:

ЗНАТИ сутність та принципи побудови моделей регресій.

УМІТИ організувати та провести моделювання регресійних моделей та прогнозування.

Завдання:

1. Побудувати модель однофакторної лінійної регресії.
2. Побудувати модель багатофакторної лінійної регресії.
3. Побудувати модель поліноміальної лінійної регресії.
4. Виконати розрахунки на тренувальній та тестовій вибірці.
5. На базі побудованих моделей розробити прогноз.
6. Зробити висновки щодо якості моделей та прогнозу.

lab2\_luzanova\_ep61

# Download the data

#Source files are here  
#setwd('D:/luzanova')  
  
##Features scaling is included in the packages we will work with  
  
#Download the files  
f\_train <- read.csv2('crime\_train.csv', header = TRUE, encoding = 'UNICOD')  
f\_train <- f\_train[,-c(1,2,18)]  
f\_test <- read.csv2('crime\_test.csv', header = TRUE, encoding = 'UNICOD')  
f\_test <- f\_test[,-c(1,2,18)]

#Висновок: окремо задані навчальна і тестова вибірки, видалені перші стовпчики з індексами об’єктів до кожної з підвибірок та стовпці Territory та Subregion

# Simple Linear Regression (one factor - Population )

## Fitting Simple Linear Regression to the Training set

model\_sr <- lm(Homicide ~ Population , f\_train)  
summary(model\_sr)

##   
## Call:  
## lm(formula = Homicide ~ Population, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.6722 -0.9117 -0.1677 0.9044 3.0695   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11.49723 1.22669 -9.373 7.01e-15 \*\*\*  
## Population 1.06619 0.07703 13.841 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.317 on 88 degrees of freedom  
## Multiple R-squared: 0.6852, Adjusted R-squared: 0.6816   
## F-statistic: 191.6 on 1 and 88 DF, p-value: < 2.2e-16

#Висновок: коефіцієнт детермінації 0,68, змінна Population - значуща

## Predicting

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

## [1] 0.685223

R2\_sr

## [1] 0.685223

train\_mse\_sr

## [1] 1.696169

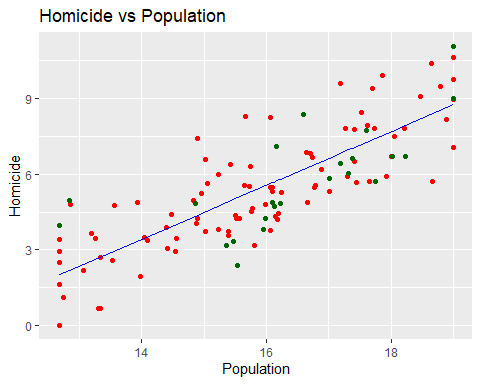
test\_mse\_sr

## [1] 2.231931

#Висновок: вручну розраховані коефіцієнти детермінації. Значення середньоквадратичної похибки на навчальній вибірці – 1,696, на тестовій вибірці – 2,232, тобто є перенавчання.

## Visualising

library(ggplot2)  
ggplot() +  
 geom\_point(aes(f\_train$Population, f\_train$Homicide),colour = 'red') +  
 geom\_point(aes(f\_test$Population, f\_test$Homicide),colour = 'darkgreen') +  
 geom\_line(aes(f\_test$Population, p\_sr),colour = 'blue') +  
 ggtitle('Homicide vs Population') +  
 xlab('Population') +  
 ylab('Homicide')



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

# Multiple Linear Regression (many factors)

## All factors

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

##   
## Call:  
## lm(formula = Homicide ~ ., data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.90996 -0.50855 0.00492 0.55005 1.88916   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 14.965191 13.899629 1.077 0.28523   
## Population 1.110230 0.168511 6.588 6.28e-09 \*\*\*  
## GDP 0.056021 0.232602 0.241 0.81036   
## Unempl 0.235468 0.184265 1.278 0.20540   
## Agri\_fore\_fish\_GDP -0.037672 0.138978 -0.271 0.78712   
## Life\_expect -8.425755 3.071960 -2.743 0.00768 \*\*   
## Death\_rate 0.076808 0.512951 0.150 0.88139   
## Birth\_rate 0.413830 0.547025 0.757 0.45181   
## Empl\_serv 1.684325 0.864070 1.949 0.05516 .   
## Empl\_agri 0.255452 0.187945 1.359 0.17833   
## Empl\_indu -0.037172 0.379595 -0.098 0.92226   
## educ\_expen 0.121604 0.176409 0.689 0.49283   
## Mil\_expen -0.185844 0.104021 -1.787 0.07821 .   
## Arm\_forc\_person 0.020996 0.097386 0.216 0.82992   
## RegionAmericas 2.032480 0.423679 4.797 8.45e-06 \*\*\*  
## RegionAsia 0.245297 0.449664 0.546 0.58709   
## RegionEurope 0.275544 0.477713 0.577 0.56587   
## RegionOceania 0.003262 0.751360 0.004 0.99655   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.8069 on 72 degrees of freedom  
## Multiple R-squared: 0.9033, Adjusted R-squared: 0.8805   
## F-statistic: 39.57 on 17 and 72 DF, p-value: < 2.2e-16

#видалемо з моделі всі змінні, окрім Population, Life\_expect, Empl\_serv, RegionAmericas

## Optimized model

model\_opt <- lm(data = f\_train, Homicide ~ Population + Life\_expect + Empl\_serv)   
summary(model\_opt)

##   
## Call:  
## lm(formula = Homicide ~ Population + Life\_expect + Empl\_serv,   
## data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.4108 -0.7369 -0.1821 0.6582 3.0446   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 31.44422 7.38829 4.256 5.28e-05 \*\*\*  
## Population 1.08551 0.06776 16.020 < 2e-16 \*\*\*  
## Life\_expect -11.88586 2.04846 -5.802 1.07e-07 \*\*\*  
## Empl\_serv 1.98828 0.70564 2.818 0.006 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.123 on 86 degrees of freedom  
## Multiple R-squared: 0.7765, Adjusted R-squared: 0.7687   
## F-statistic: 99.6 on 3 and 86 DF, p-value: < 2.2e-16

## Prediction

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

## [1] 1.204248

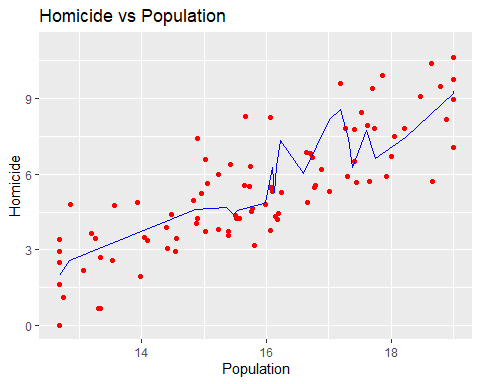
test\_mse\_opt

## [1] 2.227331

#Висновок. Значення середньоквадратичної похибки на навчальній вибірці – 1,204, на тестовій вибірці – 2,227, тобто є перенавчання. До того ж значення середньоквадратичних похибок зменшилось порівнюючи з попередніми значеннями.

## Visualising

ggplot() +  
 geom\_point(aes(f\_train$Population, f\_train$Homicide),colour = 'red') +  
 geom\_point(aes(f\_test$Population, f\_test$Homicide),colour = 'dark green') +  
 geom\_line(aes(f\_test$Population, p\_mr),colour = 'blue') +  
 ggtitle('Homicide vs Population') +  
 xlab('Population') +  
 ylab('Homicide')



# Polynomial Linear Regression (one factor - Population)

## Features extending

f\_train\_poly <- f\_train[,c('Homicide', 'Population')]  
f\_test\_poly <- f\_test[,c('Homicide', 'Population')]  
f\_train\_poly$Population2 <- f\_train\_poly$Population^2  
f\_train\_poly$Population3 <- f\_train\_poly$Population^3  
f\_test\_poly$Population2 <- f\_test\_poly$Population^2  
f\_test\_poly$Population3 <- f\_test\_poly$Population^3

## 3 powers

model\_pr <- lm(data = f\_train\_poly, Homicide ~ Population2 + Population3)   
summary(model\_pr)

##   
## Call:  
## lm(formula = Homicide ~ Population2 + Population3, data = f\_train\_poly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8293 -0.8285 -0.1621 0.7751 3.1716   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -3.375e+00 3.459e+00 -0.976 0.332  
## Population2 3.603e-02 4.154e-02 0.867 0.388  
## Population3 -9.455e-05 1.729e-03 -0.055 0.957  
##   
## Residual standard error: 1.32 on 87 degrees of freedom  
## Multiple R-squared: 0.6872, Adjusted R-squared: 0.68   
## F-statistic: 95.57 on 2 and 87 DF, p-value: < 2.2e-16

#Висновок. Значення R-squared покращилось в порівнянні з попередніми значеннями.

## Predicting

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

## [1] 1.685496

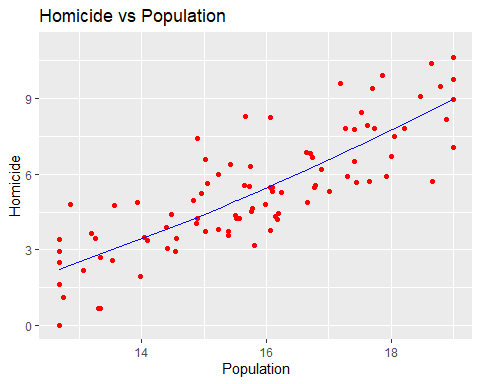
test\_mse\_poly

## [1] 2.056081

#Висновок. Значення середньоквадратичної похибки на навчальній вибірці – 1,685, на тестовій вибірці – 2,056, тобто є перенавчання. До того ж значення середньоквадратичних похибок зменшилось порівнюючи з попередніми значеннями.

## Visualising

ggplot() +  
 geom\_point(aes(f\_train\_poly$Population, f\_train\_poly$Homicide),colour = 'red') +  
 geom\_point(aes(f\_test\_poly$Population, f\_test\_poly$Homicide),colour = 'dark green') +  
 geom\_line(aes(f\_test\_poly$Population, p\_pr),colour = 'blue') +  
 ggtitle('Homicide vs Population') +  
 xlab('Population') +  
 ylab('Homicide')



# Saving results

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

Висновки

В ході виконання Лабораторної роботи №2 було виконано усі поставлені завдання. Поліноміальна лінійна регресія 3-го поряду дала найкращі результати прогнозування, але, варто зазначити, що як і в лінійній регресії, так і в поліноміальній лінійній регресії наявне перенавчання