LR3

Kovel

29 РѕРєС‚СЏР±СЂСЏ 2020 Рі

# Download the data

#Download the files  
f\_train <- read.csv2('indeks\_train.csv', header = TRUE, encoding = 'UNICOD')  
f\_train <- f\_train[,-1]  
f\_test <- read.csv2('indeks\_test.csv', header = TRUE, encoding = 'UNICOD')  
f\_test <- f\_test[,-1]

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

# Simple Linear Regression (one factor – X1)

## Fitting Simple Linear Regression to the Training set

model\_sr <- lm(Y ~ X1, f\_train)  
summary(model\_sr)

##   
## Call:  
## lm(formula = Y ~ X1, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33990 -0.06710 0.00767 0.08116 0.23314   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.006489 0.091194 -0.071 0.943   
## X1 1.000355 0.088633 11.286 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1354 on 76 degrees of freedom  
## Multiple R-squared: 0.6263, Adjusted R-squared: 0.6214   
## F-statistic: 127.4 on 1 and 76 DF, p-value: < 2.2e-16

## Висновок: обрана змінна X1 значуща, коефіцієнт детермінації 0,62 - не великий,тобто змінна Х1 не дуже гарно описує змінну У.

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

## [1] 0.6263216

R2\_sr

## [1] 0.6263216

train\_mse\_sr

## [1] 0.01785301

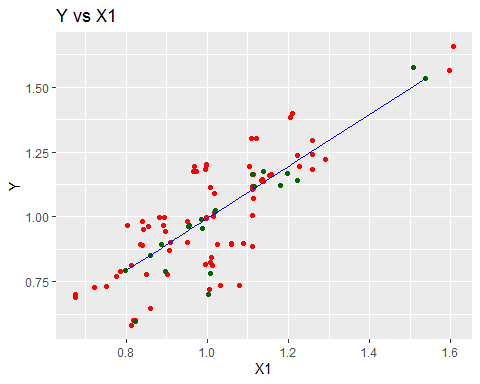
test\_mse\_sr

## [1] 0.01066023

## Висновок: вручну розраховані коефіцієнти детермінації,вони рівні. Значення середньоквадратичної похибки на навчальній вибірці – 0.01785301, на тестовій вибірці – 0.01066023, тобто перенавчання немає.

## Visualising

library(ggplot2)  
ggplot() +  
 geom\_point(aes(f\_train$X1, f\_train$Y),colour = 'red') +  
 geom\_point(aes(f\_test$X1, f\_test$Y),colour = 'dark green') +  
 geom\_line(aes(f\_test$X1, p\_sr),colour = 'blue') +  
 ggtitle('Y vs X1') +  
 xlab('X1') +  
 ylab('Y')

 ##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(Y ~ ., f\_train)   
summary(model\_mr)

##   
## Call:  
## lm(formula = Y ~ ., data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.25531 -0.08736 0.01094 0.09805 0.18070   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.0657249 0.2299379 4.635 1.55e-05 \*\*\*  
## t 0.0016052 0.0005507 2.915 0.00474 \*\*   
## X1 0.7099223 0.1064627 6.668 4.49e-09 \*\*\*  
## X2 -0.0216675 0.0110687 -1.958 0.05416 .   
## X3 -0.0077312 0.0023636 -3.271 0.00165 \*\*   
## X4 -0.6874064 0.2544285 -2.702 0.00860 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1172 on 72 degrees of freedom  
## Multiple R-squared: 0.7348, Adjusted R-squared: 0.7163   
## F-statistic: 39.89 on 5 and 72 DF, p-value: < 2.2e-16

## Висновок:змінна X2 найменш значуща, а найбільш, як і визначились раніше,Х1; коефіцієнт детермінації дорівнює 0,73, він збільшився, що говорить про те,що включення усіх змінних до моделі призвело до того,що модель стала кращою, змінна У описується заданними змінними.

## Optimized model

#as p-value, Pr(>|t|) of variable "type" is higher than significance level (5%), let's exclude this variable from the model  
mr\_opt <- lm(Y ~ t + X1 + X3 + X4, f\_train)   
summary(mr\_opt)

##   
## Call:  
## lm(formula = Y ~ t + X1 + X3 + X4, data = f\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.28634 -0.08996 0.01857 0.10766 0.18446   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.9193594 0.2216194 4.148 8.94e-05 \*\*\*  
## t 0.0016528 0.0005607 2.947 0.00430 \*\*   
## X1 0.7097647 0.1085081 6.541 7.30e-09 \*\*\*  
## X3 -0.0078172 0.0024086 -3.245 0.00177 \*\*   
## X4 -0.5950552 0.2548199 -2.335 0.02228 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1194 on 73 degrees of freedom  
## Multiple R-squared: 0.7206, Adjusted R-squared: 0.7053   
## F-statistic: 47.08 on 4 and 73 DF, p-value: < 2.2e-16

## Висновок: усі змінні значущі, коефіцієнт детермінації трохи зменшився – 0,72.

## Prediction

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

## [1] 0.01334668

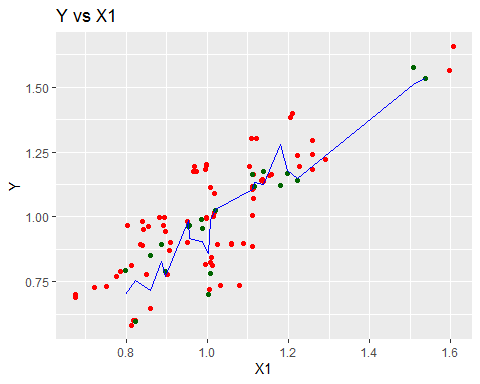
test\_mse\_opt

## [1] 0.008486525

## Висновок: значення середньоквадратичної помилки покращилися – на навчальній вибірці – 0.01334668, на тестовій вибірці – 0.008486525, тобто перенавчання немає.

## Visualising

ggplot() +  
 geom\_point(aes(f\_train$X1, f\_train$Y),colour = 'red') +  
 geom\_point(aes(f\_test$X1, f\_test$Y),colour = 'dark green') +  
 geom\_line(aes(f\_test$X1, p\_mr),colour = 'blue') +  
 ggtitle('Y vs X1') +  
 xlab('X1') +  
 ylab('Y')

 ##Висновок: на графіку червоним позначені точки навчальної вибірки, зеленим – точки тестової вибірки, синім – модельні значення. # Polynomial Linear Regression (one factor - X1)

## Features extending

f\_train\_poly <- f\_train[,c('Y', 'X1')]  
f\_test\_poly <- f\_test[,c('Y', 'X1')]  
f\_train\_poly$X12 <- f\_train\_poly$X1^2  
f\_train\_poly$X13 <- f\_train\_poly$X1^3  
f\_test\_poly$X12 <- f\_test\_poly$X1^2  
f\_test\_poly$X13 <- f\_test\_poly$X1^3

## Висновок:додано змінні Х1^2 та Х1^3.

## 3 powers

model\_pr <- lm(Y ~ X12 + X13, f\_train\_poly)   
summary(model\_pr)

##   
## Call:  
## lm(formula = Y ~ X12 + X13, data = f\_train\_poly)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.33676 -0.06865 0.00686 0.08494 0.23744   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3811 0.1222 3.120 0.00257 \*\*  
## X12 0.8221 0.2910 2.825 0.00605 \*\*  
## X13 -0.2137 0.1685 -1.268 0.20856   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1362 on 75 degrees of freedom  
## Multiple R-squared: 0.6267, Adjusted R-squared: 0.6167   
## F-statistic: 62.95 on 2 and 75 DF, p-value: < 2.2e-16

## Висновок:змінна Х1^2 значуща,та Х1^3 стала не значущою, коефіцієнт детермінації зменшився – 0,62

## Predicting

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

## [1] 0.01783671

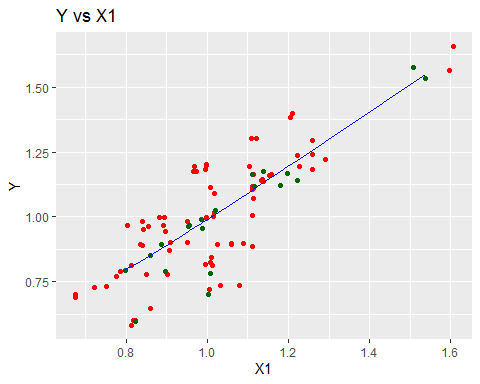
test\_mse\_poly

## [1] 0.01043797

## Висновок:значення середньоквадратичної помилки зросли на навчальній вибірці – 0.01783671, на тестовій вибірці – 0.01043797, тобто перенавчання немає

## Visualising

ggplot() +  
 geom\_point(aes(f\_train\_poly$X1, f\_train\_poly$Y),colour = 'red') +  
 geom\_point(aes(f\_test\_poly$X1, f\_test\_poly$Y),colour = 'dark green') +  
 geom\_line(aes(f\_test\_poly$X1, p\_pr),colour = 'blue') +  
 ggtitle('Y vs X1') +  
 xlab('X1') +  
 ylab('Y')

 ##Висновок:на графіку червоним позначені точки навчальної вибірки, зеленим – точки тестової вибірки, синім – модельні значення. ##Головний висновок:попередня модель була кращою, тобто зупиняемося на неї - множинної регресії із виключенням змінної Х2 # Saving results

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

## Висновок:результати моделювання збережені у файлі.