Лабораторна робота №2

Виконали: Кузьменко Юрій, Болотов Єгор

Завдання 1: Побудова однофакторної моделі

Виконайте чистку dataset одним зі способів: • видалення порожніх, заміною на середнє чи медіанне значення, прогнозування пропущених даних, • заміна неправильних типів у dataset.

```
colnames(df)[2]='Artist_name'
colnames(df)[3]='Song_name'
colnames(df)[5]='Top_ten_times'
colnames(df)[6]='Peak_position'
colnames(df)[7]='Peak_position_times'
colnames(df)[8]='Peak_streams'
colnames(df)[9]='Total_streams'


df$Peak_position_times <- gsub("[(x)]","" , df$Peak_position_times ,ignore.case
= TRUE)
df$Peak_position_times <- as.numeric(df$Peak_position_times)
```

Використовуючи комп'ютерне програмне забезпечення виконайте оцінку моделі залежності = 0 + 1 1 + .

Використовуємо X - Peak Position times Y - Top Ten Times

```
x1 <- df$Peak_position_times
Y <- df$Top_ten_times
x1_scaled <- scale(x1)
mod1 <- lm(Y ~ x1_scaled)
```

(A) Знайдіть оцінене значення

```
b1 <- cov(x1_scaled, Y)/var(x1_scaled)
b1 <- b1[1,1]
b1
```

```
## [1] 10.97885
```

(B) Знайдіть оцінене значення

```
b0 <- mean(Y) - b1*mean(x1_scaled)
b0
```

```
## [1] 2.713641
```

(C) За допомогою вбудованої функції lm () запишіть значення оцінених коефіцієнтів (mod1 <- lm(Y ~ X, data)).

```
mod1 <- lm(Y ~ x1_scaled)
coefs <- mod1$coefficients
coefs

## (Intercept)   x1_scaled
##    2.713641   10.978847
```

(D) Зробіть висновки про збіг коефіцієнтів. Запишіть аналітичний вигляд моделі з відомими значеннями коефіцієнтів;

```
smod1 <- summary(mod1)
smod1

##
## Call:
## lm(formula = Y ~ x1_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -193.265   -1.535   -1.535   -1.535  212.125
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7136     0.1053   25.76   <2e-16 ***
## x1_scaled    10.9788     0.1053  104.23   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.09 on 11082 degrees of freedom
## Multiple R-squared:  0.495,  Adjusted R-squared:  0.495
## F-statistic: 1.086e+04 on 1 and 11082 DF,  p-value: < 2.2e-16
```

Коефіцієнт детермінації рівен 0.495, що робить цю модель неприйнятною

```
anltc_mdl <- b0 + b1*x1_scaled
anltc_mdl

##                [,1]
##     [1,]  89.874905
##     [2,]  62.459041
##     [3,]  13.719727
##
## [11082,]   1.534899
## [11083,]   1.534899
## [11084,]   1.534899
## attr(,"scaled:center")
## [1] 0.3869542
## attr(,"scaled:scale")
## [1] 3.604104
```

Завдання 2: Побудова 2-х факторної моделі

Використовуючи комп'ютерне програмне забезпечення виконайте регресійні розрахунки та побудуйте модель, коли = 2, тобто за 2-ма незалежними змінними.

Використовуємо X1 - Peak Position times X2 - Peak Streams Y - Top Ten Times

(A) Знайдіть оцінене значення вектора коефіцієнтів за допомогою аналітичної формули (2.1) для 2-х факторної моделі;

```
x2 = df$Peak_streams
x2_scaled <- scale(x2)

mod2 <- lm(Y ~ x1_scaled + x2_scaled)
mod2$coefficients

## (Intercept)   x1_scaled   x2_scaled
##    2.713641   10.090035    2.708297

X <- cbind(1,x1_scaled, x2_scaled)
beta <- solve(t(X) %*% X) %*% t(X) %*% Y
beta

##           [,1]
## [1,]  2.713641
## [2,] 10.090035
## [3,]  2.708297
```

(B) За допомогою вбудованої функції lm () обчисліть значення оцінених коефіцієнтів (mod2 <- lm(Y ~ x1 + x2, data)).

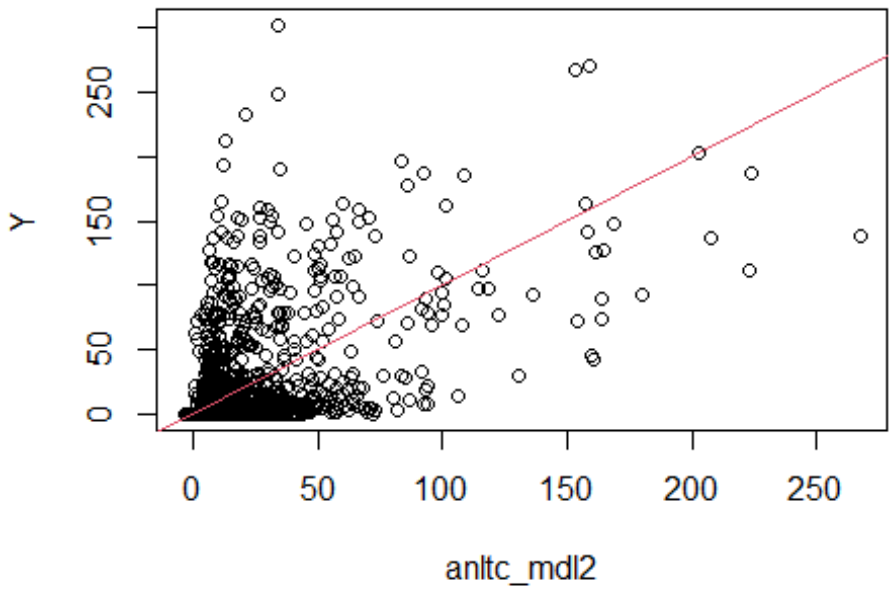
```
mod2

##
## Call:
## lm(formula = Y ~ x1_scaled + x2_scaled)
##
## Coefficients:
## (Intercept)    x1_scaled    x2_scaled
##       2.714       10.090        2.708
```

(C) Зробіть висновки про збіг коефіцієнтів. Запишіть аналітичний вигляд моделі з відомими значеннями коефіцієнтів;

```
smod2 <- summary(mod2)
smod2

##
## Call:
## lm(formula = Y ~ x1_scaled + x2_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -172.922   -1.505   -0.656   -0.224  212.434
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.7136     0.1025   26.48   <2e-16 ***
## x1_scaled    10.0900     0.1085   92.99   <2e-16 ***
## x2_scaled     2.7083     0.1085   24.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.79 on 11081 degrees of freedom
## Multiple R-squared:  0.5219, Adjusted R-squared:  0.5218
## F-statistic:  6048 on 2 and 11081 DF,  p-value: < 2.2e-16
```

Коефіцієнт детермінації знаходиться на значення 0.5219, ця модель є прийнятною

```
anltc_mdl2 <- X[, 2:3]*beta[1:2,1]+ beta[1,1]
anltc_mdl2 <- t(anltc_mdl2[,1] + anltc_mdl2[, 2])
plot(anltc_mdl2, Y)
abline(a = 0, b = 1, col=2)
```



(D) Видрукуйте відповідні значення для
mod2$fitted.values та похибок ε mod2 residuals;

```
smod2 <- summary(mod2)
smod2$residuals
```

```
##             1             2             3             4             5
##  2.124341e+02  1.135898e+02  1.943971e+02  3.910948e+00 -1.729218e+02
##
##         11081         11082         11083         11084
##  5.397603e-01  5.525412e-01  5.593734e-01  5.621278e-01
```

```
smod2$fitted.values
```

```
## NULL
```

Завдання 3: Побудова математичної моделі за всіма параметрами

(A) Використовуючи комп'ютерне програмне забезпечення виконайте регресійні розрахунки та побудуйте модель за не більше ніж 5-ма незалежними змінними. mod3 <- lm(Y ~ ., data)

Використовуємо X1 - Peak Position times X2 - Peak Streams X3 - Days X4 - Total Steams X5 - Peak Position Y - Top Ten Times

```
x3 <- df$Days
x4 <- df$Total_streams
x5 <- df$Peak_position

x3_scaled <- scale(x3)
x4_scaled <- scale(x4)
x5_scaled <- scale(x5)

Y <- df$Top_ten_times
mod3 <- lm(Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled)
```

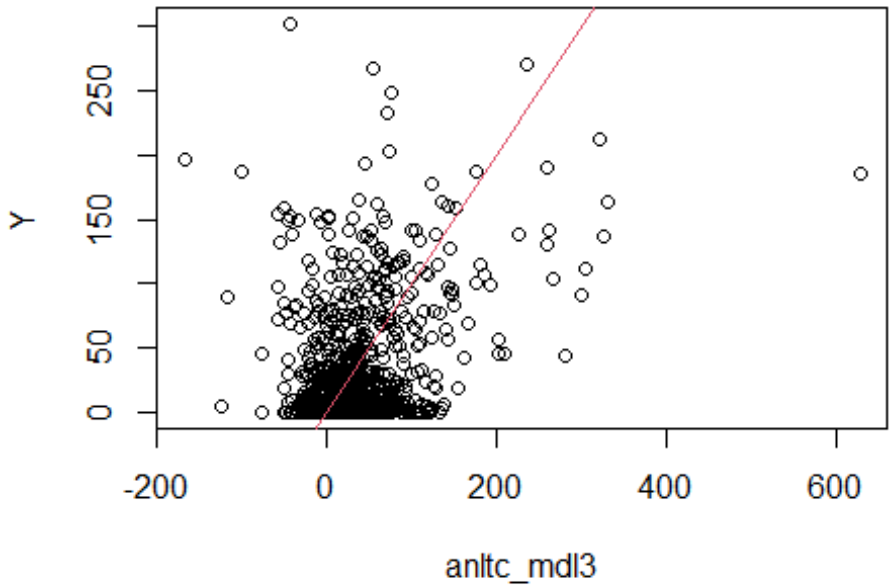
(B) Запишіть аналітичний вигляд моделі з відомими значеннями коефіцієнтів;

```
smod3 <- summary(mod3)
smod3
```

```
##
## Call:
## lm(formula = Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled +
##     x5_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -155.812   -0.269   -0.015    0.238  140.727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.71364    0.07803  34.775   <2e-16 ***
## x1_scaled    5.58899    0.09851  56.734   <2e-16 ***
```

```
## x2_scaled   -0.92927    0.11100  -8.372   <2e-16 ***
## x3_scaled   -8.70695    0.24746 -35.185   <2e-16 ***
## x4_scaled   17.05671    0.27419  62.208   <2e-16 ***
## x5_scaled   -0.31377    0.10198  -3.077   0.0021 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.215 on 11078 degrees of freedom
## Multiple R-squared:  0.7229, Adjusted R-squared:  0.7228
## F-statistic:  5781 on 5 and 11078 DF,  p-value: < 2.2e-16
```

```
X <- cbind(1,x1_scaled, x2_scaled, x3_scaled, x4_scaled, x5_scaled)
beta <- solve(t(X) %*% X) %*% t(X) %*% Y

anltc_mdl3 <- X[, 2:6]*beta[1:5,1]+ beta[1,1]
anltc_mdl3 <- t(anltc_mdl3[,1] + anltc_mdl3[, 2] + anltc_mdl3[, 3] +
anltc_mdl3[, 4]
                + anltc_mdl3[, 5])
plot(anltc_mdl3, Y)
abline(a = 0, b = 1, col=2)
```



(C) Визначте значення коефіцієнта детермінації R^2 за допомогою вбудованої функції mod*$r.squared для моделей mod1, mod2 та mod3;

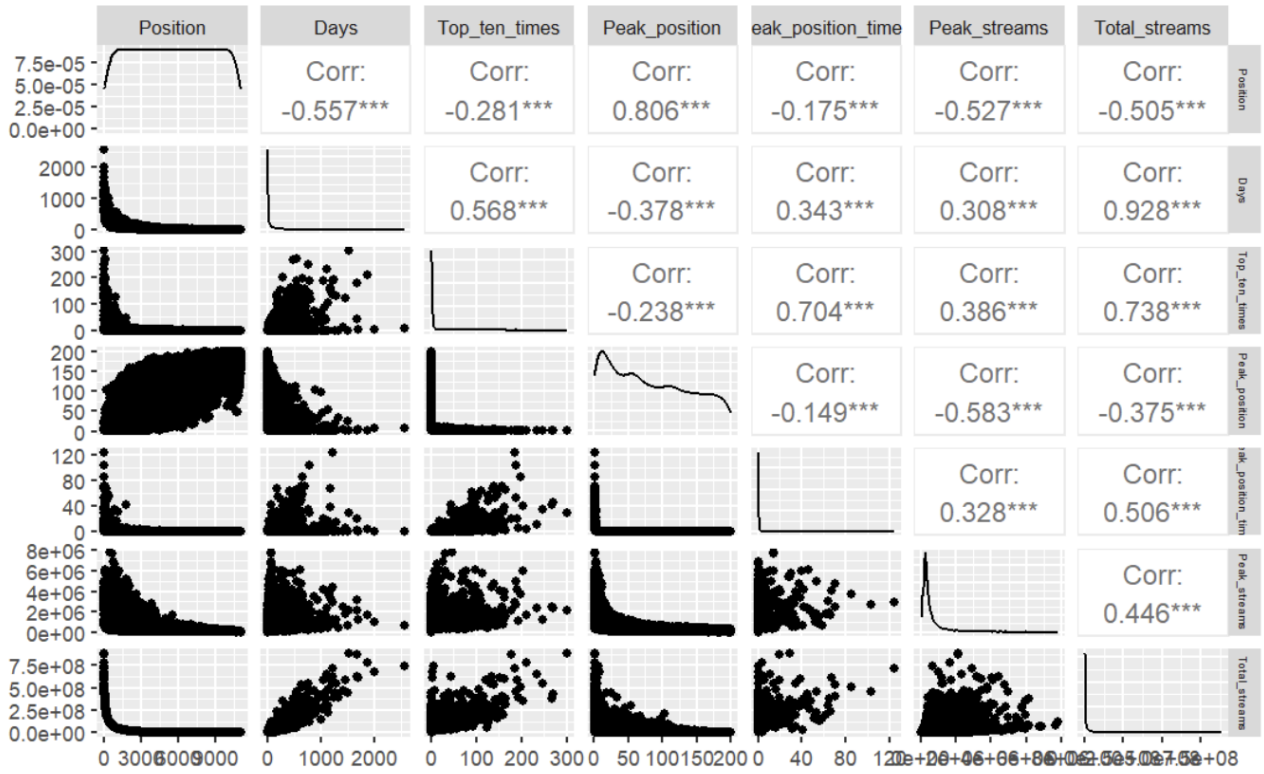
```
smod1$r.squared

## [1] 0.4950333

smod2$r.squared

## [1] 0.5219129
```

```
smod3$r.squared

## [1] 0.7229283
```

Третя модель з найбільшим коефіцієнтом детермінації

(D) Видрукуйте car::scatterplotMatrix (ggpairs(df, columns = c(1,4:9))) та вкажіть, які змінні на ваш погляд мають більш лінійний зв'язок:

```
p <- ggpairs(df, columns = c(1,4:9))

p + theme(strip.text.x = element_text(size = 8),
          strip.text.y = element_text(size = 5))
```



Дивлячись на дані графіки можна відмітити такі змінні як: Position та Peak Position Top Ten times та Total Streams

Є ще інші чудові графіки, але вони мають нелінійний зв'язок