Лабораторна робота №2 (частина 3)

Побудова та статистичний аналіз лінійної множинної регресії.

Опис dataset

Назва dataset: Spotify Top 10000 Streamed Songs

Link на dataset: https://www.kaggle.com/datasets/rakkesharv/spotify-top-10000-streamed-songs

Опис dataset та постановку задачі: Це набір даних, зібраний з веб-сайту Spotify, котрий містить потоки виконавця та кількість прослiховувань (було взято саме топ-10000) Основна мета: вплив факторів на популярність пісні й дізнатись найпопулярніших виконавців та треки.

Змінні та їх опис: Position - Spotify Ranking Artist Name - Artist Name Song Name - Song Name Days - No of days since the release of the song (Updated) Top 10 (xTimes) - No of times inside top 10 Peak Position - Peak position attained Peak Position (xTimes) - No of times Peak position attained Peak Streams - Total no of streams during Peak position Total Streams - Total song streams

Завдання 7: ANOVA

(A) Використовуючи функцію simpleAnova, виведіть спрощену таблицю ANOVA для:

a. моделі mod2

```
simpleAnova <- function(object, ...) {

  # Обчислити таблицю anova
  tab <- anova(object, ...)

  # Отримати кількість предикторів
  p <- nrow(tab) - 1

  # Додайте рядок предикторів
  predictorsRow <- colSums(tab[1:p, 1:2])
  predictorsRow <- c(predictorsRow, predictorsRow[2] / predictorsRow[1])

  # F-значення
  Fval <- predictorsRow[3] / tab[p + 1, 3]
  pval <- pf(Fval, df1 = p, df2 = tab$Df[p + 1], lower.tail = FALSE)
  predictorsRow <- c(predictorsRow, Fval, pval)

  # Спрощена таблиця
  tab <- rbind(predictorsRow, tab[p + 1, ])
  row.names(tab)[1] <- "Predictors"
  return(tab)

}
simpleAnova(mod2)
```

```
## Analysis of Variance Table
##
## Response: Y
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Predictors     2 1408427  704214  6048.4 < 2.2e-16 ***
## Residuals  11081 1290160     116
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

b. моделі mod3

```
simpleAnova(mod3)
```

```
## Analysis of Variance Table
##
## Response: Y
##               Df  Sum Sq Mean Sq F value    Pr(>F)
## Predictors     5 1950885  390177  5780.9 < 2.2e-16 ***
## Residuals  11078  747702      67
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

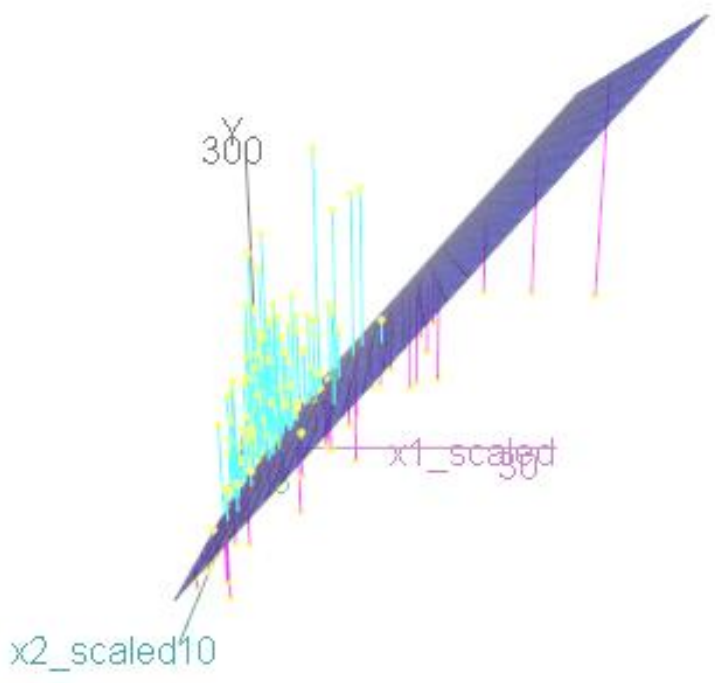
(B) Зробіть аналіз таблиць для моделей mod2 та mod3:
Значення відрізняються, але не відрізняється від значень в summary()

Завдання 8: Визначення адекватності моделі за коефіцієнтами R^2та RAdj^2

(A) Виконайте 3D візуалізацію для mod2: car::scatter3d(y ~ x1 + x2, fit = “linear”) rgl::rglwidget()

```
car::scatter3d(Y ~ x1_scaled + x2_scaled, fit = "linear")

rgl::rglwidget()
```



(B) Обчисліть значення R^2 за допомогою формул для mod1 та mod0; R^2 = 1 – (SSE/SST)

```
mod0 <- lm(Y ~ 0 + x1_scaled)
smod0 <- summary(mod0)

Rsquared <- function(object, ...) {
sse <- sum((fitted(object) - Y)^2)
ssr <- sum((fitted(object) - mean(Y))^2)
sst <- ssr + sse
Rsq <- 1 - (sse/sst)
return(Rsq)
}

Rsquared(mod0)

## [1] 0.4953166

Rsquared(mod1)

## [1] 0.4950333
```

(C) Обчисліть RAdj^2 для моделі mod1 та mod0 та порівняйте з відповідними значеннями коефіцієнта R^2

```
AdjRsquared <- function(object, param, ...) {
sse <- sum((fitted(object) - Y)^2)
ssr <- sum((fitted(object) - mean(Y))^2)
sst <- ssr + sse
ARsq <- 1 - (sse/sst)*((length(Y)-1)/((length(Y)-param-1)))
return(ARsq)
}

AdjRsquared(mod0,2)

## [1] 0.4805001

AdjRsquared(mod1,1)

## [1] 0.4949877
```

Порівняти з:

```
Rsquared(mod0)

## [1] 0.4805001

Rsquared(mod1)

## [1] 0.4950333
```

(D) Вкажіть значення коефіцієнта R^2 та RAdj^2 для mod1, mod2 та mod3. Зробіть висновок,яка модель, у вашому випадку, є найкращою.

R-squared

```
smod1$r.squared
```

```
## [1] 0.4950333
```

```
smod2$r.squared
```

```
## [1] 0.5219129
```

```
smod3$r.squared
```

```
## [1] 0.7229283
```

Adj. R-squared

```
smod1$adj.r.squared
```

```
## [1] 0.4949877
```

```
smod2$adj.r.squared
```

```
## [1] 0.5218266
```

```
smod3$adj.r.squared
```

```
## [1] 0.7228033
```

За даними значеннями можна встановити що найкраща модель - mod3

Завдання 9: Модель без вільного коефіцієнта та центровані моделі

(A) Побудуйте модель без вільного коефіцієнта для 1-но факторної моделі (mod0) $Y = 0 + \beta 1X1 + \varepsilon$. Це можна виконати за допомогою mod0 <- lm(y ~ 0 + x1, data):

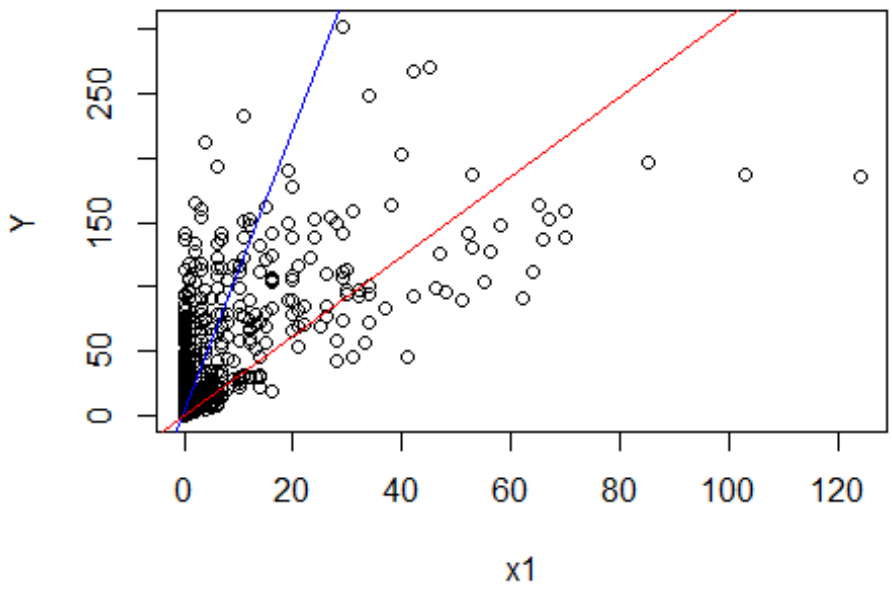
```
mod0 <- lm(Y ~ 0 + x1)
summary(mod0)
```

```
##
## Call:
## lm(formula = Y ~ 0 + x1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -197.3    0.0    0.0    0.0  212.3
##
## Coefficients:
```

```
##    Estimate Std. Error t value Pr(>|t|)
## x1  3.09141    0.02933   105.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 11083 degrees of freedom
## Multiple R-squared:  0.5006, Adjusted R-squared:  0.5005
## F-statistic: 1.111e+04 on 1 and 11083 DF,  p-value: < 2.2e-16
```

(B) Побудуйте розсіювання та накладіть відповідні прямі регресія для моделей mod1 та mod0;

```
plot(x1, Y)
abline(mod1, col="blue")
abline(mod0, col="red")
```



(C) Виконаємо аналогічні обчислення для центрованого dataset (dataCen)

a. Побудуйте mod0 та mod1 для dataCen;

```
x1_cen <- dataCen$Peak_position_times
mod1_cen <- lm(Y ~  x1_cen)
mod0_cen <- lm(Y ~ 0 + x1_cen)
summary(mod0_cen)
```

```
## Call:
## lm(formula = Y ~ 0 + x1_cen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -190.551    1.179    1.179    1.179  214.839
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## x1_cen  3.04621    0.03009   101.2   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Residual standard error: 11.42 on 11083 degrees of freedom
## Multiple R-squared:  0.4805, Adjusted R-squared:  0.4805
## F-statistic: 1.025e+04 on 1 and 11083 DF,  p-value: < 2.2e-16
```

b. Обчисліть та порівняйте значення коефіцієнтів R^2 та RAdj^2 для моделей mod0 та mod1 з пункту (a);

```
summary(mod0_cen)$r.squared

## [1] 0.4805001

summary(mod1_cen)$r.squared

## [1] 0.4950333

summary(mod0_cen)$adj.r.squared

## [1] 0.4804533

summary(mod1_cen)$adj.r.squared

## [1] 0.4949877

# Значення збігаються

# Побудова розсіювання та результатів mod0 та mod1
plot(x1_cen, Y)
abline(mod1_cen, col="blue")
abline(mod0_cen, col="red")
```