Лабораторна робота №2 (частина 2)

Виконали: Кузьменко Юрій, Болотов Єгор

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

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 4: Статистичні оцінки моделі за всіма параметрами

(A) Вказати отримані значення

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

```
##               Estimate Std. Error    t value      Pr(>|t|)
## (Intercept)  2.7136413 0.07803427  34.774995 1.428494e-251
## x1_scaled    5.5889867 0.09851295  56.733520  0.000000e+00
## x2_scaled   -0.9292734 0.11099889  -8.371916  6.350423e-17
## x3_scaled   -8.7069453 0.24745974 -35.185301 3.444178e-257
## x4_scaled   17.0567078 0.27418873  62.207910  0.000000e+00
## x5_scaled   -0.3137653 0.10197718  -3.076818  2.097332e-03
```

a. Estimate: оцінка найменших квадратів βj

```
           Estimate
```

(Intercept) 2.7136413
x1_scaled 5.5889867
x2_scaled -0.9292734
x3_scaled -8.7069453
x4_scaled 17.0567078
x5_scaled -0.3137653

b. Std. Error: Оцінка стандартних помилок $\hat{SE}(\beta j)$

```
          Std. Error
```

(Intercept) 0.07803427
x1_scaled 0.09851295
x2_scaled 0.11099889
x3_scaled 0.24745974
x4_scaled 0.27418873
x5_scaled 0.10197718

c. t value: t -статистика$\beta j\hat{SE}(\beta j)$

```
          t value
```

(Intercept) 34.774995
x1_scaled 56.733520
x2_scaled -8.371916
x3_scaled -35.185301
x4_scaled 62.207910
x5_scaled -3.076818

d. Pr(>|t|): p – значення t-тесту

```
          Pr(>|t|)
```

(Intercept) 1.428494e-251
x1_scaled 0.000000e+00
x2_scaled 6.350423e-17
x3_scaled 3.444178e-257
x4_scaled 0.000000e+00
x5_scaled 2.097332e-03

(B) Значення середньо квадратичним відхиленням (Residual standard error)

a. Визначити коефіцієнт середнього квадратичного відхилення за допомогою наступних функцій sqrt(sum(mod3$residuals^2)/mod3$df.residual)

```
RSE <- sqrt(sum(mod3$residuals^2)/ mod3$df.residual)
RSE
```

```
## [1] 8.215493
```

b. Порівняти результат отриманий в (a) з відповідним Residual standard error в об'єкті summary(mod3) або за допомогою функції mod3$sigma.

```
RSE
```

```
## [1] 8.215493
```

```
smod3$sigma
```

```
## [1] 8.215493
```

Значення ідентичні

(C) Порівняти моделі mod2 та mod3 за допомогою car::compareCoefs(mod2, mod3)

```
car::compareCoefs(mod2, mod3)
```

```
## Calls:
## 1: lm(formula = Y ~ x1_scaled + x2_scaled)
## 2: lm(formula = Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled +
##   x5_scaled)
##
##             Model 1 Model 2
## (Intercept)   2.714   2.714
## SE            0.102   0.078
##
## x1_scaled   10.0900  5.5890
## SE           0.1085  0.0985
##
## x2_scaled     2.708  -0.929
```

```
## SE            0.109   0.111
##
## x3_scaled            -8.707
## SE                    0.247
##
## x4_scaled            17.057
## SE                    0.274
##
## x5_scaled            -0.314
## SE                    0.102
##
```

Завдання 5: Довірчі інтервали коефіцієнтів моделі та центрування dataset

(A) Вказати довірчі інтервали для коефіцієнтів βj з різними інтервалами довіри

a. 0.95% інтервал довіри для βj

```
confint(mod3, level = .95)
```

```
##                   2.5 %     97.5 %
## (Intercept)  2.5606802  2.8666024
## x1_scaled    5.3958837  5.7820896
## x2_scaled   -1.1468509 -0.7116958
## x3_scaled   -9.1920105 -8.2218802
## x4_scaled   16.5192491 17.5941666
## x5_scaled   -0.5136587 -0.1138718
```

b. 0.90% інтервал довіри для βj

```
confint(mod3, level = .90)
```

```
##                     5 %       95 %
## (Intercept)  2.5852756  2.8420070
## x1_scaled    5.4269337  5.7510396
## x2_scaled   -1.1118655 -0.7466812
## x3_scaled   -9.1140144 -8.2998762
## x4_scaled   16.6056698 17.5077459
## x5_scaled   -0.4815168 -0.1460137
```

c. 0.99% інтервал довіри для βj

```
confint(mod3, level = .99)
```

```
##                   0.5 %      99.5 %
## (Intercept)  2.5126037  2.91467888
## x1_scaled    5.3351904  5.84278294
## x2_scaled   -1.2152368 -0.64330990
```

```
## x3_scaled   -9.3444692 -8.06942145
## x4_scaled   16.3503228 17.76309292
## x5_scaled   -0.5764864 -0.05104419
```

(B) Вказати довірчі інтервали для коефіцієнтів βj з різними інтервалами довіри

b. 0.95% інтервал довіри для βj

```
confint(mod3, level = .95)
```

```
##                  2.5 %     97.5 %
## (Intercept)  2.5606802  2.8666024
## x1_scaled    5.3958837  5.7820896
## x2_scaled   -1.1468509 -0.7116958
## x3_scaled   -9.1920105 -8.2218802
## x4_scaled   16.5192491 17.5941666
## x5_scaled   -0.5136587 -0.1138718
```

c. 0.90% інтервал довіри для βj

```
confint(mod3, level = .90)
```

```
##                    5 %       95 %
## (Intercept)  2.5852756  2.8420070
## x1_scaled    5.4269337  5.7510396
## x2_scaled   -1.1118655 -0.7466812
## x3_scaled   -9.1140144 -8.2998762
## x4_scaled   16.6056698 17.5077459
## x5_scaled   -0.4815168 -0.1460137
```

d. 0.99% інтервал довіри для βj

```
confint(mod3, level = .99)
```

```
##                  0.5 %      99.5 %
## (Intercept)  2.5126037  2.91467888
## x1_scaled    5.3351904  5.84278294
## x2_scaled   -1.2152368 -0.64330990
## x3_scaled   -9.3444692 -8.06942145
## x4_scaled   16.3503228 17.76309292
## x5_scaled   -0.5764864 -0.05104419
```

(C) Виконайте центрування dataset

c. створити центрований dataCen;

```
dataCen <- data.frame(scale(df[c(1,4:9)] , center = TRUE,scale = FALSE))
x1_cen <- dataCen$Peak_position_times
x2_cen <- dataCen$Peak_streams
x3_cen <- dataCen$Days
x4_cen <- dataCen$Total_streams
x5_cen <- dataCen$Peak_position
```

d. побудуйте аналогічне моделі mod3 модель modCen для dataCen

```
mod3 <- lm(Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled)
summary(mod3)
```

```
##
## 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
```

```
mod_cen <- lm(Y ~ x1_cen + x2_cen + x3_cen + x4_cen + x5_cen)
summary(mod_cen)
```

```
##
## Call:
## lm(formula = Y ~ x1_cen + x2_cen + x3_cen + x4_cen + x5_cen)
##
## 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)  2.714e+00  7.803e-02  34.775   <2e-16 ***
## x1_cen       1.551e+00  2.733e-02  56.734   <2e-16 ***
## x2_cen      -1.479e-06  1.767e-07  -8.372   <2e-16 ***
## x3_cen      -6.711e-02  1.907e-03 -35.185   <2e-16 ***
## x4_cen       3.268e-07  5.253e-09  62.208   <2e-16 ***
## x5_cen      -5.324e-03  1.730e-03  -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
```

c. порівняйте чи збігаються статистичні оцінки в обох моделях.
Статистичні оцінки збігаються

Завдання 6: Прогноз

(A) Задайте значення $X1 = x1, \ldots, Xp = xp$ для прогнозування, з розрахунку для кожного i-го фактора як $max(Xi) + 10\%$;

```
topPrediction <- data.frame(x1 = max(x1) + 0.1 * max(x1),
                            x2 = max(x2) + 0.1 * max(x2),
                            x3 = max(x3) + 0.1 * max(x3),
                            x4 = max(x4) + 0.1 * max(x4),
                            x5 = max(x5) + 0.1 * max(x5))
topPrediction
```

```
##      x1      x2     x3        x4  x5
## 1 136.4 8564706 2801.7 971706712 220
```

```
top_predict <- lm(Y ~ x1 + x2 + x3 + x4 + x5)

predict(top_predict, newdata = topPrediction)
```

```
##        1
## 328.1931
```

(B) Прогнозування серединного значення $\hat{Y}$ та його довірчого інтервалу:

d. 95% довірчий інтервал;

```
predict(top_predict, newdata = topPrediction, interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 328.1931 320.4767 335.9095
```

b. 90% довірчий інтервал;

```
predict(top_predict, newdata = topPrediction, interval = "confidence", level = 0.90)
```

```
##        fit      lwr      upr
## 1 328.1931 321.7175 334.6687
```

c. 99% довірчий інтервал;

```
predict(top_predict, newdata = topPrediction, interval = "confidence", level =
0.99)
```

```
##        fit      lwr      upr
## 1 328.1931 318.0515 338.3348
```

(D) Прогнозування значення Y та його довірчого інтервалу при умові, що це невідоме значення лежить за межами досліджуваного діапазону:

a. 95% довірчий інтервал;

```
predict(top_predict, newdata = topPrediction, interval = "prediction", level =
0.95)
```

```
##        fit     lwr      upr
## 1 328.1931 310.336 346.0502
```

b. 90% довірчий інтервал;

```
predict(top_predict, newdata = topPrediction, interval = "prediction", level =
0.90)
```

```
##        fit      lwr      upr
## 1 328.1931 313.2074 343.1789
```

c. 99% довірчий інтервал;

```
predict(top_predict, newdata = topPrediction, interval = "prediction", level =
0.99)
```

```
##        fit      lwr      upr
## 1 328.1931 304.7234 351.6628
```