Лабораторна робота №3 (частина 1)

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

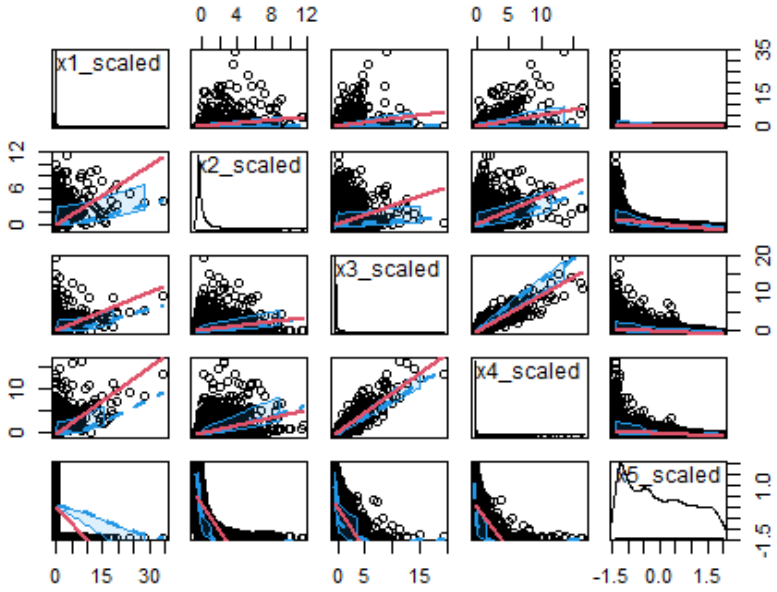
Total Streams - Total song streams

```
mod2 <- lm(Y ~ x1_scaled + x2_scaled);
mod3 <- lm(Y ~ x1_scaled + x2_scaled + x3_scaled);
modAll <- lm(Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled);
modInter <- lm(Y ~ x1_scaled + x2_scaled + x3_scaled, )
modZero <- lm(Y ~ 1)
```

Завдання 1: Інформаційний критерій міри продуктивності моделі (BIC / AIC)

(A) Оберіть 5-ть змінних, які на вашу думку мають нелінійність, та побудуйте скатерплот; car::scatterplotMatrix(~ x1 + x2 + x3 + x4 + x5, regLine = list(col = 2), col = 1, smooth = list(col.smooth = 4, col.spread = 4), data = data)

```
car::scatterplotMatrix( ~ x1_scaled + x2_scaled + x3_scaled
                        + x4_scaled + x5_scaled, regLine = list(col = 2),
 col = 1, smooth = list(col.smooth = 4, col.spread = 4),
 data = df)
```



(B) Побудуйте BIC та AIC для моделей mod2 та mod3;

```
#BIC
BIC(mod2)

## [1] 84219.08

BIC(mod3)

## [1] 81555.71

#AIC
AIC(mod2)

## [1] 84189.82

AIC(mod3)

## [1] 81519.14
```

(C) Яка модель mod2 чи mod3 краща за критерієм AIC;

За критерієм AIC краща модель - mod3

(D) Перевірте (C) за допомогою summary(_);

```
summary(mod2)

##
## 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

summary(mod3)

##
## Call:
## lm(formula = Y ~ x1_scaled + x2_scaled + x3_scaled)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -166.960   -1.038    0.602    1.089  166.111
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.71364    0.09085   29.87   <2e-16 ***
## x1_scaled    8.61413    0.09986   86.26   <2e-16 ***
## x2_scaled    1.51552    0.09860   15.37   <2e-16 ***
## x3_scaled    5.44961    0.09914   54.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.565 on 11080 degrees of freedom
## Multiple R-squared:  0.6243, Adjusted R-squared:  0.6242
## F-statistic:  6138 on 3 and 11080 DF,  p-value: < 2.2e-16
```

З аналізу можна зрозуміти що mod3 буде кращою

Завдання 2: Пошук накращої підмножини факторів

(A) Зробити покращення за допомогою MASS::stepAIC моделі за всіма параметрами modAll, які є в dataset.

```
modAIC <- MASS::stepAIC(modAll, k = 2)

## Start:  AIC=46692.29
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled
##
##             Df Sum of Sq     RSS   AIC
## <none>                    747702 46692
## - x5_scaled  1       639  748341 46700
## - x2_scaled  1      4731  752433 46760
## - x3_scaled  1     83558  831260 47865
## - x1_scaled  1    217243  964946 49517
## - x4_scaled  1    261191 1008893 50011
```

(B) Побудувати критерій AIC

o modAIC <- MASS::stepAIC(mod, k = 2).

```
modAIC <- MASS::stepAIC(modAll, k = 2)

## Start:  AIC=46692.29
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled
##
##             Df Sum of Sq     RSS   AIC
## <none>                    747702 46692
## - x5_scaled  1       639  748341 46700
## - x2_scaled  1      4731  752433 46760
## - x3_scaled  1     83558  831260 47865
## - x1_scaled  1    217243  964946 49517
## - x4_scaled  1    261191 1008893 50011
```

o Описати значення AIC на кожному етапі і вказати яка змінна видаляється;

Етапів немає, значення AIC рівне = 46692.29

o Записати фінальне аналітичне представлення модель;

```
summary(modAIC)

##
## 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
```

(C) Побудувати критерій BIC

o MASS::stepAIC(*, trace = 0, k = log(n))).

```
modBIC <- MASS::stepAIC(modAll, k = log(nrow(df)))

## Start:  AIC=46736.17
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled
##
##             Df Sum of Sq     RSS   AIC
## <none>                    747702 46736
## - x5_scaled  1       639  748341 46736
## - x2_scaled  1      4731  752433 46797
## - x3_scaled  1     83558  831260 47901
## - x1_scaled  1    217243  964946 49554
## - x4_scaled  1    261191 1008893 50048
```

Етапів немає, значення BIC = 46736.17

o Записати фінальне аналітичне представлення модель;

```
summary(modBIC)

##
## 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
```

Завдання 3: Пошук накращої підмножини використовуючи direction = “forward”,“backward” або “both”.

(A) Побудувати MASS::stepAIC за direction = backward:

o MASS::stepAIC (modAll, direction = “backward”, k = log(n));

```
modBack <- MASS::stepAIC (modAll, direction = "backward", k = log(nrow(df)))
```

```
## Start:  AIC=46736.17
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled
##
##             Df Sum of Sq     RSS   AIC
## <none>                     747702 46736
## - x5_scaled  1       639  748341 46736
## - x2_scaled  1      4731  752433 46797
## - x3_scaled  1     83558  831260 47901
## - x1_scaled  1    217243  964946 49554
## - x4_scaled  1    261191 1008893 50048
```

Етапів немає, значення рівне 46736.17

o Аналітично записати фінальну модель;

```
summary(modBack)
```

```
##
## 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
```

(B) Побудувати MASS::stepAIC за direction = forward:

o MASS::stepAIC(modZero, direction = “forward”, scope = list(lower = modZero, upper = modAll), k = log(n));

```
modFwd1 <- MASS::stepAIC(modZero, direction = "forward", scope = list(lower =
   modZero, upper = modAll), k = log(nrow(df)))

## Start:  AIC=60915.68
## Y ~ 1
##
##             Df Sum of Sq     RSS   AIC
## + x4_scaled  1   1468005 1230582 52221
## + x1_scaled  1   1335890 1362697 53352
## + x3_scaled  1    871527 1827060 56602
## + x2_scaled  1    401606 2296981 59139
## + x5_scaled  1    152653 2545934 60280
## <none>                   2698587 60916
##
## Step:  AIC=52221.39
## Y ~ x4_scaled
##
##             Df Sum of Sq     RSS   AIC
## + x1_scaled  1    395861  834721 47928
## + x3_scaled  1    261609  968973 49582
## + x2_scaled  1     10810 1219773 52133
## + x5_scaled  1      4754 1225828 52188
## <none>                   1230582 52221
##
## Step:  AIC=47928.5
## Y ~ x4_scaled + x1_scaled
##
##             Df Sum of Sq    RSS   AIC
## + x3_scaled  1     82035 752686 46791
## + x5_scaled  1      1344 833377 47920
## <none>                   834721 47928
## + x2_scaled  1       429 834292 47932
##
## Step:  AIC=46791.18
```

```
## Y ~ x4_scaled + x1_scaled + x3_scaled
##
##             Df Sum of Sq    RSS   AIC
## + x2_scaled  1    4345.4 748341 46736
## <none>                   752686 46791
## + x5_scaled  1     253.8 752433 46797
##
## Step:  AIC=46736.32
## Y ~ x4_scaled + x1_scaled + x3_scaled + x2_scaled
##
##             Df Sum of Sq    RSS   AIC
## + x5_scaled  1    638.96 747702 46736
## <none>                   748341 46736
##
## Step:  AIC=46736.17
## Y ~ x4_scaled + x1_scaled + x3_scaled + x2_scaled + x5_scaled
```

6 етапів, значення

1. 60915.68 (аналіз Y ~ 1)

2. 52221.39 (аналіз з використанням Y ~ Total_streams)

3. 47928.5 (аналіз з використанням Y ~ Total_streams + Peak_position_times)

4. 46791.18 (аналіз з використанням Y ~ Total_streams + Peak_position_times + Days)

5. 46736.32 (аналіз з використанням Y ~ Total_streams + Peak_position_times + Days + Peak_streams)

Фінальне = 46736.17 (Використано всі змінні)

o Аналітично записати фінальну модель;

```
summary(modFwd1)

##
## Call:
## lm(formula = Y ~ x4_scaled + x1_scaled + x3_scaled + x2_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 ***
## x4_scaled   17.05671    0.27419  62.208   <2e-16 ***
## x1_scaled    5.58899    0.09851  56.734   <2e-16 ***
## x3_scaled   -8.70695    0.24746 -35.185   <2e-16 ***
## x2_scaled   -0.92927    0.11100  -8.372   <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
```

(C) Побудувати MASS::stepAIC за direction = forward:

o MASS::stepAIC(modInter, direction = “forward”, scope = list(lower = modZero, upper = modAll), k = log(n));

```
modFwd2 <- MASS::stepAIC(modInter, direction = "forward", scope = list(lower =
  modZero, upper = modAll), k = log(nrow(df)))
```

```
## Start:  AIC=50091.37
## Y ~ x1_scaled + x2_scaled + x3_scaled
##
##             Df Sum of Sq     RSS   AIC
## + x4_scaled  1    265389  748341 46736
## + x5_scaled  1      4836 1008893 50048
## <none>                   1013730 50091
##
## Step:  AIC=46736.32
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled
##
##             Df Sum of Sq    RSS   AIC
## + x5_scaled  1    638.96 747702 46736
## <none>                   748341 46736
##
## Step:  AIC=46736.17
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled
```

3 етапи, значення

1. 50091.37 (аналіз Y ~ Peak_Position_times + Peak_Streams + Days)

2. 46736.32 (аналіз з використанням Y ~ Peak_Position_times + Peak_Streams + Days + Total_streams)

Фінальне = 46736.17 (Використано всі змінні)

o Аналітично записати фінальну модель;

```
summary(modFwd2)
```

```
##
## 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
```

(D) Побудувати MASS::stepAIC за direction = both:

o MASS::stepAIC(modInter, direction = “both”, scope = list(lower = modZero, upper = modAll), k = log(n));

```
modBoth <- MASS::stepAIC(modInter, direction = "both", scope = list(lower =
   modZero, upper = modAll), k = log(nrow(df)))
```

```
## Start:  AIC=50091.37
## Y ~ x1_scaled + x2_scaled + x3_scaled
##
##             Df Sum of Sq     RSS   AIC
## + x4_scaled  1    265389  748341 46736
## + x5_scaled  1      4836 1008893 50048
## <none>                   1013730 50091
## - x2_scaled  1     21614 1035344 50316
## - x3_scaled  1    276430 1290160 52755
## - x1_scaled  1    680769 1694499 55776
##
## Step:  AIC=46736.32
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled
##
##             Df Sum of Sq     RSS   AIC
## + x5_scaled  1       639  747702 46736
## <none>                    748341 46736
## - x2_scaled  1      4345  752686 46791
## - x3_scaled  1     85951  834292 47932
## - x1_scaled  1    216756  965097 49546
## - x4_scaled  1    265389 1013730 50091
##
## Step:  AIC=46736.17
## Y ~ x1_scaled + x2_scaled + x3_scaled + x4_scaled + x5_scaled
##
```

```
##             Df Sum of Sq     RSS   AIC
## <none>                     747702 46736
## - x5_scaled  1       639  748341 46736
## - x2_scaled  1      4731  752433 46797
## - x3_scaled  1     83558  831260 47901
## - x1_scaled  1    217243  964946 49554
## - x4_scaled  1    261191 1008893 50048
```

3 етапи, значення

1. 50091.37 (аналіз Y ~ Peak_Position_times + Peak_Streams + Days)

2. 46736.32 (аналіз з використанням Y ~ Peak_Position_times + Peak_Streams + Days + Total_streams)

Фінальне = 46736.17 (Використано всі змінні) Не відрізняється від modFwd2

o Аналітично записати фінальну модель;

```
summary(modBoth)
```

```
##
## 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
```

Загальний висновок: серед усих варіацій множин, накйкращою виявилась з усіма змінними при собі, но можемо побачити, що модель Y ~ Peak_Position_times + Peak_Streams + Days + Total_streams максимально схожа за значенням за повної моделі, тому можна сказати, що Peak_position не сильно впливає на наш аналіз

Завдання 4: Використання якісних предикторів.

(A) Визначення категоріальних змінних str(data);

```
df$Peak_position_times <- as.factor(df$Peak_position_times)
df = subset(df, select = -c(Artist_name, Song_name))
str(df)
```

```
## tibble [11,084 × 7] (S3: tbl_df/tbl/data.frame)
##  $ Position           : num [1:11084] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Days               : num [1:11084] 1506 1673 1853 2547 1223 ...
##  $ Top_ten_times      : num [1:11084] 302 178 212 6 186 4 233 44 133 2 ...
##  $ Peak_position      : num [1:11084] 1 1 1 7 1 8 1 2 1 5 ...
##  $ Peak_position_times: Factor w/ 57 levels "0","1","2","3",..: 28 19 5 1 57
1 12 2 7 1 ...
##  $ Peak_streams       : num [1:11084] 2118242 2127668 1660502 659366 2905678
...
##  $ Total_streams      : num [1:11084] 8.83e+08 8.65e+08 7.81e+08 7.35e+08
7.19e+08 ...
```

```
# Короткий зміст лінійної моделі
mod1 <- lm(Top_ten_times ~ ., data = df)
summary(mod1)
```

```
##
## Call:
## lm(formula = Top_ten_times ~ ., data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.836  -0.435  -0.058   0.396 126.446
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.456e+00  2.470e-01   5.896 3.82e-09 ***
## Position                 1.707e-04  4.038e-05   4.228 2.38e-05 ***
## Days                    -5.459e-02  1.817e-03 -30.046  < 2e-16 ***
## Peak_position           -1.480e-02  2.024e-03  -7.313 2.79e-13 ***
## Peak_position_times1     1.127e+01  5.507e-01  20.460  < 2e-16 ***
## Peak_position_times2     2.295e+01  1.044e+00  21.990  < 2e-16 ***
## Peak_position_times3     2.533e+01  1.184e+00  21.391  < 2e-16 ***
## Peak_position_times4     2.176e+01  1.343e+00  16.199  < 2e-16 ***
## Peak_position_times5     2.082e+01  1.702e+00  12.232  < 2e-16 ***
## Peak_position_times6     3.144e+01  1.531e+00  20.531  < 2e-16 ***
## Peak_position_times7     3.741e+01  1.713e+00  21.840  < 2e-16 ***
## Peak_position_times8     4.256e+01  2.474e+00  17.206  < 2e-16 ***
## Peak_position_times9     2.985e+01  3.485e+00   8.566  < 2e-16 ***
## Peak_position_times10    3.757e+01  2.231e+00  16.842  < 2e-16 ***
## Peak_position_times11    7.128e+01  2.543e+00  28.029  < 2e-16 ***
## Peak_position_times12    5.705e+01  2.337e+00  24.412  < 2e-16 ***
## Peak_position_times13    4.335e+01  3.479e+00  12.462  < 2e-16 ***
## Peak_position_times14    4.712e+01  2.676e+00  17.606  < 2e-16 ***
```

```
## Peak_position_times15   6.467e+01  3.140e+00  20.597  < 2e-16 ***
## Peak_position_times16   6.893e+01  2.652e+00  25.989  < 2e-16 ***
## Peak_position_times19   9.703e+01  4.054e+00  23.933  < 2e-16 ***
## Peak_position_times20   5.452e+01  2.725e+00  20.006  < 2e-16 ***
## Peak_position_times21   5.351e+01  3.488e+00  15.344  < 2e-16 ***
## Peak_position_times22   4.759e+01  4.920e+00   9.672  < 2e-16 ***
## Peak_position_times23   7.104e+01  6.967e+00  10.197  < 2e-16 ***
## Peak_position_times24   1.202e+02  4.912e+00  24.466  < 2e-16 ***
## Peak_position_times25   4.833e+01  6.943e+00   6.961 3.56e-12 ***
## Peak_position_times26   5.988e+01  4.022e+00  14.887  < 2e-16 ***
## Peak_position_times27   1.131e+02  6.964e+00  16.235  < 2e-16 ***
## Peak_position_times28   6.542e+01  4.026e+00  16.252  < 2e-16 ***
## Peak_position_times29   8.658e+01  3.213e+00  26.943  < 2e-16 ***
## Peak_position_times30   6.806e+01  4.058e+00  16.770  < 2e-16 ***
## Peak_position_times31   9.181e+01  4.906e+00  18.714  < 2e-16 ***
## Peak_position_times32   5.060e+01  4.945e+00  10.233  < 2e-16 ***
## Peak_position_times33   1.514e+01  6.959e+00   2.176   0.0296 *
## Peak_position_times34   8.338e+01  3.518e+00  23.697  < 2e-16 ***
## Peak_position_times37   3.855e+01  6.989e+00   5.515 3.56e-08 ***
## Peak_position_times38   7.477e+01  7.032e+00  10.632  < 2e-16 ***
## Peak_position_times40   1.421e+02  7.045e+00  20.165  < 2e-16 ***
## Peak_position_times41   4.172e+01  6.933e+00   6.017 1.83e-09 ***
## Peak_position_times42   1.216e+02  4.987e+00  24.385  < 2e-16 ***
## Peak_position_times45   1.907e+02  7.035e+00  27.105  < 2e-16 ***
## Peak_position_times46   4.848e+01  7.012e+00   6.915 4.94e-12 ***
## Peak_position_times47   7.879e+01  6.967e+00  11.309  < 2e-16 ***
## Peak_position_times48   4.108e+01  6.971e+00   5.893 3.91e-09 ***
## Peak_position_times51   6.383e+01  6.943e+00   9.193  < 2e-16 ***
## Peak_position_times52   1.210e+02  6.938e+00  17.436  < 2e-16 ***
## Peak_position_times53   1.048e+02  4.973e+00  21.077  < 2e-16 ***
## Peak_position_times55   8.096e+01  6.941e+00  11.664  < 2e-16 ***
## Peak_position_times56   1.200e+02  6.940e+00  17.287  < 2e-16 ***
## Peak_position_times58   1.415e+02  6.947e+00  20.363  < 2e-16 ***
## Peak_position_times62   6.379e+01  6.942e+00   9.189  < 2e-16 ***
## Peak_position_times64   6.309e+01  6.974e+00   9.046  < 2e-16 ***
## Peak_position_times65   9.788e+01  6.995e+00  13.993  < 2e-16 ***
## Peak_position_times85   1.070e+02  7.066e+00  15.147  < 2e-16 ***
## Peak_position_times103  1.138e+02  7.004e+00  16.247  < 2e-16 ***
## Peak_position_times124  6.606e+01  7.115e+00   9.284  < 2e-16 ***
## Peak_streams           -3.405e-06  1.648e-07 -20.666  < 2e-16 ***
## Total_streams           2.715e-07  4.916e-09  55.227  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.931 on 11022 degrees of freedom
## Multiple R-squared:  0.8038, Adjusted R-squared:  0.8027
## F-statistic: 740.2 on 61 and 11022 DF,  p-value: < 2.2e-16
```

Як ми бачимо очевидних категоріальних змінних в нашому датасеті немає, але ми можемо представити кількість раз в топі як категоріальну змінну Найдорожча змінна - Peak_position_times19

(B) Визначення кодування фіктивної змінної contrasts(data$x);

```
contrasts(df$Peak_position_times)
```

```
    1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34
0   0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
1   1 0 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
2   0 1 0 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
3   0 0 1 0 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
4   0 0 0 1 0 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
5   0 0 0 0 1 0 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
6   0 0 0 0 0 1 0 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
7   0 0 0 0 0 0 1 0 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
8   0 0 0 0 0 0 0 1 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
9   0 0 0 0 0 0 0 0 1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
10  0 0 0 0 0 0 0 0 0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
11  0 0 0 0 0 0 0 0 0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
12  0 0 0 0 0 0 0 0 0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
13  0 0 0 0 0 0 0 0 0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
14  0 0 0 0 0 0 0 0 0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
15  0 0 0 0 0 0 0 0 0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
16  0 0 0 0 0 0 0 0 0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
    37 38 40 41 42 45 46 47 48 51 52 53 55 56 58 62 64 65 66 67 70 85 103 124
0    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
1    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
2    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
3    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
4    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
5    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
6    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
7    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
8    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
9    0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
10   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
11   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
12   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
13   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
14   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
15   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
16   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0    0   0
 [ reached getOption("max.print") -- omitted 40 rows ]
```

Як ми бачимо,зараз еталонна змінна Peak_position_times0

(C) Зміна еталонної змінної за допомогою relevel(*)

```
df$Peak_position_times <- relevel(df$Peak_position_times, ref='19')
```

(D) Побудова моделі з фіктивною змінною $Y = \beta 0 + \beta 1x1 + \beta 2x2 + \beta 3D$;

```
mod2 <- lm(Top_ten_times ~ Total_streams + Peak_streams+Peak_position_times,
data = df )
summary(mod2)

##
## Call:
## lm(formula = Top_ten_times ~ Total_streams + Peak_streams +
Peak_position_times,
##     data = df)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.049   0.032   0.441   0.646 112.925
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.088e+02  4.248e+00  25.604  < 2e-16 ***
## Total_streams            1.229e-07  1.743e-09  70.525  < 2e-16 ***
## Peak_streams            -1.846e-06  1.383e-07 -13.353  < 2e-16 ***
## Peak_position_times0    -1.088e+02  4.243e+00 -25.646  < 2e-16 ***
## Peak_position_times1    -9.701e+01  4.261e+00 -22.766  < 2e-16 ***
## Peak_position_times2    -8.540e+01  4.352e+00 -19.620  < 2e-16 ***
## Peak_position_times3    -8.114e+01  4.390e+00 -18.484  < 2e-16 ***
## Peak_position_times4    -8.422e+01  4.436e+00 -18.986  < 2e-16 ***
## Peak_position_times5    -8.894e+01  4.578e+00 -19.428  < 2e-16 ***
## Peak_position_times6    -7.375e+01  4.487e+00 -16.435  < 2e-16 ***
## Peak_position_times7    -6.764e+01  4.571e+00 -14.797  < 2e-16 ***
## Peak_position_times8    -6.179e+01  4.937e+00 -12.514  < 2e-16 ***
## Peak_position_times9    -7.555e+01  5.568e+00 -13.570  < 2e-16 ***
## Peak_position_times10   -6.423e+01  4.800e+00 -13.382  < 2e-16 ***
## Peak_position_times11   -2.229e+01  4.932e+00  -4.520 6.24e-06 ***
## Peak_position_times12   -4.660e+01  4.862e+00  -9.584  < 2e-16 ***
## Peak_position_times13   -6.482e+01  5.572e+00 -11.634  < 2e-16 ***
## Peak_position_times14   -5.981e+01  5.038e+00 -11.870  < 2e-16 ***
## Peak_position_times15   -3.547e+01  5.322e+00  -6.666 2.76e-11 ***
## Peak_position_times16   -3.299e+01  5.031e+00  -6.558 5.69e-11 ***
## Peak_position_times20   -3.842e+01  5.027e+00  -7.641 2.33e-14 ***
## Peak_position_times21   -4.961e+01  5.567e+00  -8.911  < 2e-16 ***
## Peak_position_times22   -5.214e+01  6.653e+00  -7.836 5.06e-15 ***
## Peak_position_times23   -2.136e+01  8.413e+00  -2.539  0.01113 *
## Peak_position_times24    1.559e+01  6.655e+00   2.343  0.01916 *
## Peak_position_times25   -5.582e+01  8.416e+00  -6.633 3.44e-11 ***
## Peak_position_times26   -4.172e+01  5.951e+00  -7.011 2.51e-12 ***
## Peak_position_times27    2.210e+01  8.414e+00   2.626  0.00864 **
## Peak_position_times28   -3.755e+01  5.956e+00  -6.305 3.00e-10 ***
## Peak_position_times52    1.129e+01  8.416e+00   1.342  0.17978
## Peak_position_times66    9.769e-01  8.413e+00   0.116  0.90756
## Peak_position_times67   -2.378e+01  8.426e+00  -2.822  0.00478 **
## Peak_position_times70   -1.089e+01  6.655e+00  -1.636  0.10182
## Peak_position_times85    3.325e+01  8.418e+00   3.950 7.85e-05 ***
## Peak_position_times103   2.752e+01  8.415e+00   3.270  0.00108 **
## Peak_position_times124  -5.742e+00  8.440e+00  -0.680  0.49631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.285 on 11025 degrees of freedom
## Multiple R-squared:  0.7832, Adjusted R-squared:  0.7821
## F-statistic: 686.7 on 58 and 11025 DF,  p-value: < 2.2e-16
```

Як ми бачимо, коли Peak_position_times19 стала головною, то усі інші змінні дуже сильно втратили в цінності

(E) Висновки: на скільки зміниться приріст середнього значення Y, який пов'язаний зі зміною $D = 0$ на $D = 1$.

Для нашого випадку важко прописати кожен коєфіцієнт, оскільки у нас 31 категорія, але за допомогою summary(mod2) можна легко побачити на скільки змінився кожний коєфіцієнт

```
summary(mod2)

##
## Call:
## lm(formula = Top_ten_times ~ Total_streams + Peak_streams +
Peak_position_times,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -83.049   0.032   0.441   0.646 112.925
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)              1.088e+02  4.248e+00  25.604  < 2e-16 ***
## Total_streams            1.229e-07  1.743e-09  70.525  < 2e-16 ***
## Peak_streams            -1.846e-06  1.383e-07 -13.353  < 2e-16 ***
## Peak_position_times0    -1.088e+02  4.243e+00 -25.646  < 2e-16 ***
## Peak_position_times1    -9.701e+01  4.261e+00 -22.766  < 2e-16 ***
## Peak_position_times2    -8.540e+01  4.352e+00 -19.620  < 2e-16 ***
## Peak_position_times3    -8.114e+01  4.390e+00 -18.484  < 2e-16 ***
## Peak_position_times4    -8.422e+01  4.436e+00 -18.986  < 2e-16 ***
## Peak_position_times5    -8.894e+01  4.578e+00 -19.428  < 2e-16 ***
## Peak_position_times6    -7.375e+01  4.487e+00 -16.435  < 2e-16 ***
## ...
## Peak_position_times52    1.129e+01  8.416e+00   1.342  0.17978
## Peak_position_times66    9.769e-01  8.413e+00   0.116  0.90756
## Peak_position_times67   -2.378e+01  8.426e+00  -2.822  0.00478 **
## Peak_position_times70   -1.089e+01  6.655e+00  -1.636  0.10182
## Peak_position_times85    3.325e+01  8.418e+00   3.950 7.85e-05 ***
## Peak_position_times103   2.752e+01  8.415e+00   3.270  0.00108 **
## Peak_position_times124  -5.742e+00  8.440e+00  -0.680  0.49631
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.285 on 11025 degrees of freedom
## Multiple R-squared:  0.7832, Adjusted R-squared:  0.7821
## F-statistic: 686.7 on 58 and 11025 DF,  p-value: < 2.2e-16
```