Виконали: Кузьменко Юрій, Болотов Єгор

Побудова та статистичний аналіз нелінійної 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

```
library(readr)
library(plotly)

y <- df$Top_ten_times
x1 <- df$Peak_position_times
x2 <- df$Peak_streams
x3 <- df$Days
x4 <- df$Total_streams
x5 <- df$Peak_position

x_binary <- ifelse(x5 <= 10, 'Yes', 'No')
x_binary
```

```
##     [1] "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"
"Yes"
##    [13] "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes"
"Yes"
##    [25] "Yes" "No"  "No"  "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "Yes" "No"
"Yes"
##    [37] "Yes" "Yes" "Yes" "Yes" "Yes" "No"  "Yes" "Yes" "Yes" "Yes" "Yes"
"Yes"
## [10933] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [10945] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [10957] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [10969] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [10981] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [10993] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [11005] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [11017] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [11029] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [11041] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [11053] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [11065] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
"No"
## [11077] "No"  "No"  "No"  "No"  "No"  "No"  "No"  "No"
```

```
typeof(x_binary)
```

```
## [1] "character"
```

Завдання 4: Нелінійні моделі за допомогою взаємодії між змінним.

(A) Побудуйте взаємодію між змінними xi для таких моделей. Визначте яка краща:

a) $y \sim x1 * x2$

```
mod_8 <- lm(y~x1*x2)
summary(mod_8)

##
## Call:
## lm(formula = y ~ x1 * x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -157.541   -1.503   -0.493    0.003  206.289
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.146e+00  1.375e-01  -8.329   <2e-16 ***
## x1           3.815e+00  6.260e-02  60.940   <2e-16 ***
## x2           4.996e-06  1.741e-07  28.689   <2e-16 ***
## x1:x2       -3.966e-07  2.153e-08 -18.419   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.63 on 11080 degrees of freedom
## Multiple R-squared:  0.5361, Adjusted R-squared:  0.536
## F-statistic:  4268 on 3 and 11080 DF,  p-value: < 2.2e-16
```

b) $y \sim x1 * x2 * x3$

```
mod_9 <- lm(y~x1*x2*x3)
summary(mod_9)

##
## Call:
## lm(formula = y ~ x1 * x2 * x3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -142.999   -0.295    0.133    0.192  133.960
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.065e-01  1.225e-01  -0.869   0.3847
## x1           1.784e+00  1.083e-01  16.480  < 2e-16 ***
## x2          -3.845e-07  1.677e-07  -2.292   0.0219 *
## x3           7.434e-03  1.084e-03   6.855 7.50e-12 ***
## x1:x2        2.358e-07  3.766e-08   6.262 3.95e-10 ***
## x1:x3        3.879e-03  2.171e-04  17.870  < 2e-16 ***
## x2:x3        3.313e-08  8.122e-10  40.789  < 2e-16 ***
## x1:x2:x3    -1.719e-09  7.246e-11 -23.719  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.709 on 11076 degrees of freedom
## Multiple R-squared:  0.6887, Adjusted R-squared:  0.6885
## F-statistic:  3501 on 7 and 11076 DF,  p-value: < 2.2e-16
```

c) $MASS::stepAIC(object = lm(y \sim ., data), scope = y \sim .\,\hat{}\,2, k = log(nobs(modBIC)), trace = 0)$.

```
modAll <- lm(y ~ x1 + x2 + x3 + x4 + x5);
modBIC <- MASS::stepAIC(modAll, k = log(nrow(df)))

## Start:  AIC=46736.17
## y ~ x1 + x2 + x3 + x4 + x5
##
##        Df Sum of Sq     RSS   AIC
## <none>               747702 46736
## - x5    1       639  748341 46736
## - x2    1      4731  752433 46797
## - x3    1     83558  831260 47901
## - x1    1    217243  964946 49554
## - x4    1    261191 1008893 50048

mod_10 <- MASS::stepAIC(object=lm(y ~ x1 + x2 + x3 + x4 + x5), scope = (y~ x1^2
+ x2^2 + x3^2 + x4^2 + x5^2),
k=log(nobs(modBIC)),trace=0)
summary(mod_10)

##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5)
##
## 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)  9.716e-01  2.360e-01   4.116 3.88e-05 ***
## x1           1.551e+00  2.733e-02  56.734  < 2e-16 ***
## x2          -1.479e-06  1.767e-07  -8.372  < 2e-16 ***
## x3          -6.711e-02  1.907e-03 -35.185  < 2e-16 ***
## x4           3.268e-07  5.253e-09  62.208  < 2e-16 ***
## x5          -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
```

(B) Визначте кращу модель з пункту (A)

За допомогою summary можна побачити, що mod_10 краща

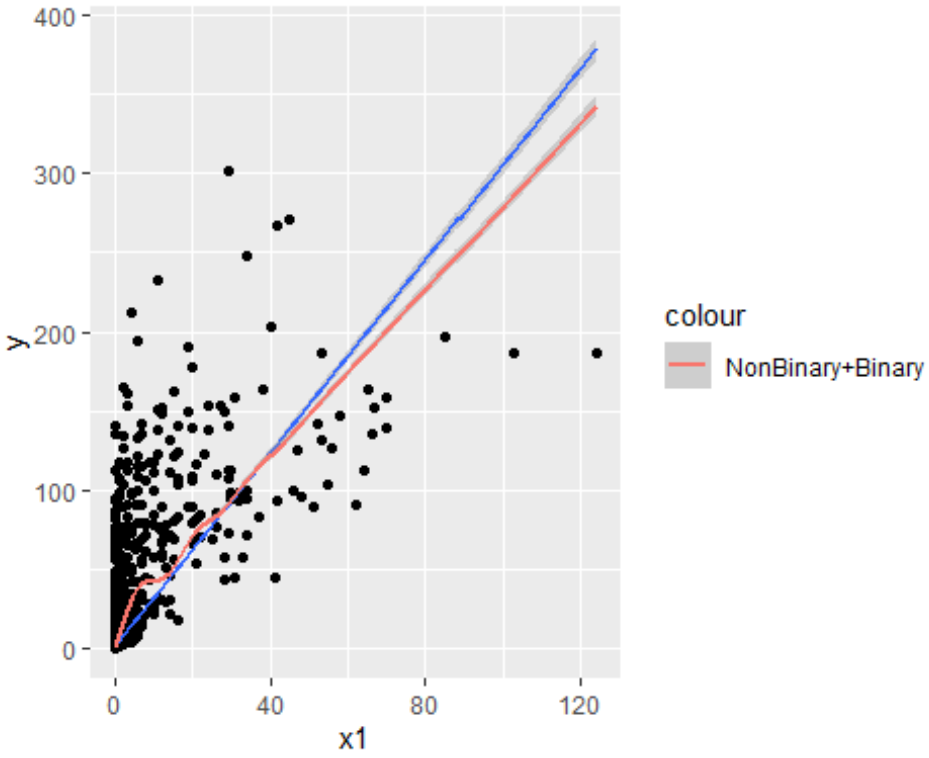
Завдання 5: Нелінійні моделі за допомогою взаємодії між неперервною та бінарною змінною.

(A) Побудуйте взаємодія між неперервною та бінарною змінною.

a) $Yi = \beta 0 + \beta 1Xi + \beta 2Di + ui$

```
mod_1 <- lm(y~x1 + x_binary)
pred_1 <- predict(mod_1)
ggplot(data = df, aes(x = x1, y = y)) +
  geom_point()+
  stat_smooth(method=lm)+
  geom_smooth(aes(color = "NonBinary+Binary", y=pred_1),)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```



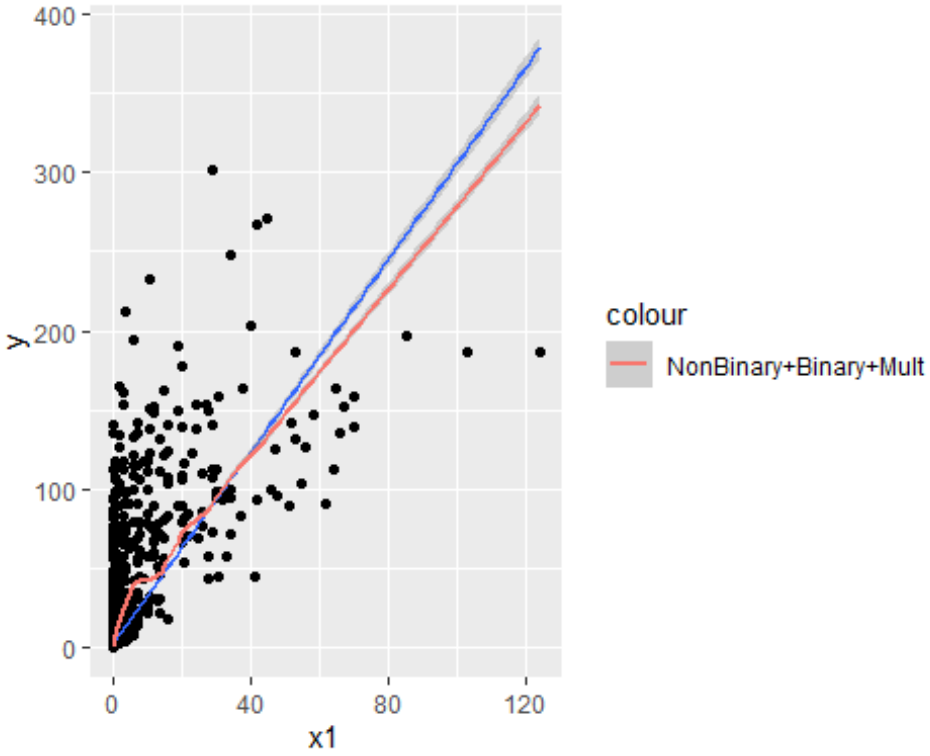

```
## `geom_smooth()` using formula = 'y ~ x1 + x_binary'
summary(mod_1)
```

```
##
## Call:
## lm(formula = y ~ x1 + x_binary)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -156.9    0.0    0.0    0.0  210.1
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.941e-12  1.018e-01    0.00        1
```

```
## x1          2.643e+00  2.796e-02   94.50   <2e-16 ***
## x_binaryYes 1.524e+01  3.209e-01   47.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 11081 degrees of freedom
## Multiple R-squared:  0.5804, Adjusted R-squared:  0.5804
## F-statistic:  7665 on 2 and 11081 DF,  p-value: < 2.2e-16
```

c) $Yi = \beta 0 + \beta 1Xi + \beta 2Di + \beta 3(Xi \times Di) + ui$

```
mod_2 <- lm(y~x1 + x_binary+ x1:x_binary)
pred_2 <- predict(mod_2)
ggplot(data = df, aes(x = x1, y = y)) +
  geom_point()+
  stat_smooth(method=lm)+
  geom_smooth(aes(color = "NonBinary+Binary+Mult", y=pred_2))

## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```

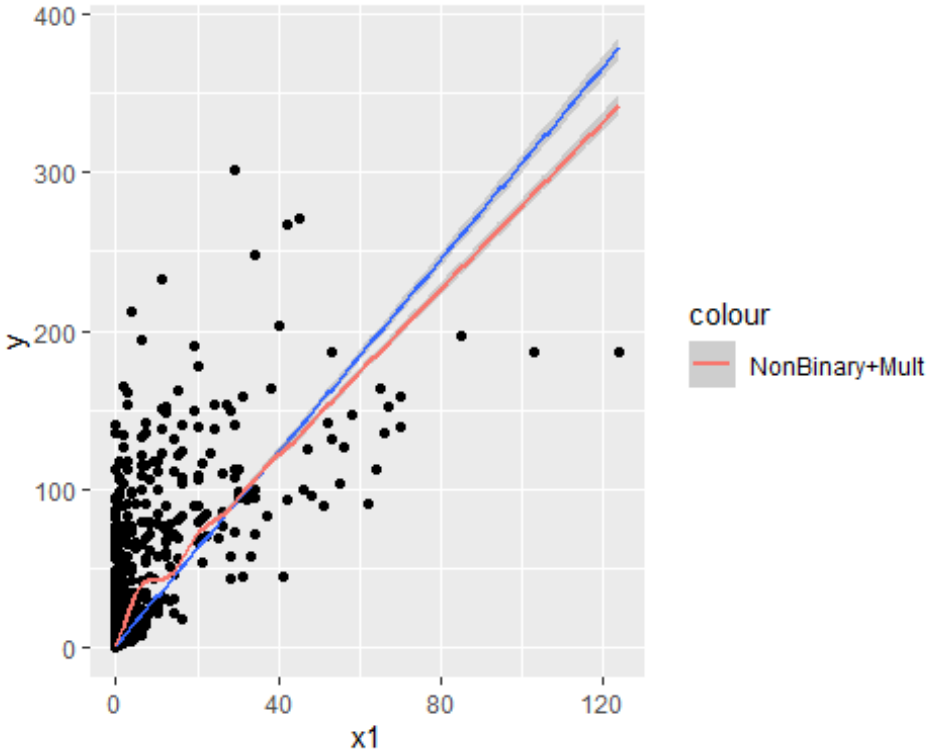



```
## `geom_smooth()` using formula = 'y ~ x1 + x_binary + x1*x_binary'
summary(mod_2)

##
## Call:
## lm(formula = y ~ x1 + x_binary + x1:x_binary)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -156.9    0.0     0.0     0.0   210.1
```

```
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.941e-12  1.018e-01    0.00        1
## x1             2.643e+00  2.796e-02   94.50   <2e-16 ***
## x_binaryYes    1.524e+01  3.209e-01   47.49   <2e-16 ***
## x1:x_binaryYes        NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.11 on 11081 degrees of freedom
## Multiple R-squared:  0.5804, Adjusted R-squared:  0.5804
## F-statistic:  7665 on 2 and 11081 DF,  p-value: < 2.2e-16
```

d) $Yi = \beta 0 + \beta 1 Xi + \beta 2 (Xi \times Di) + \mathcal{u}i$

```
mod_3 <- lm(y~x1 + x1:x_binary)
pred_3 <- predict(mod_3)
ggplot(data = df, aes(x = x1, y = y)) +
  geom_point()+
  stat_smooth(method=lm)+
  geom_smooth(aes(color = "NonBinary+Mult", y=pred_2))

## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using method = 'gam' and formula = 'y ~ s(x, bs = "cs")'
```




```
## `geom_smooth()` using formula = 'y ~ x1 + x_binary + x1*x_binary'
summary(mod_3)

##
## Call:
```

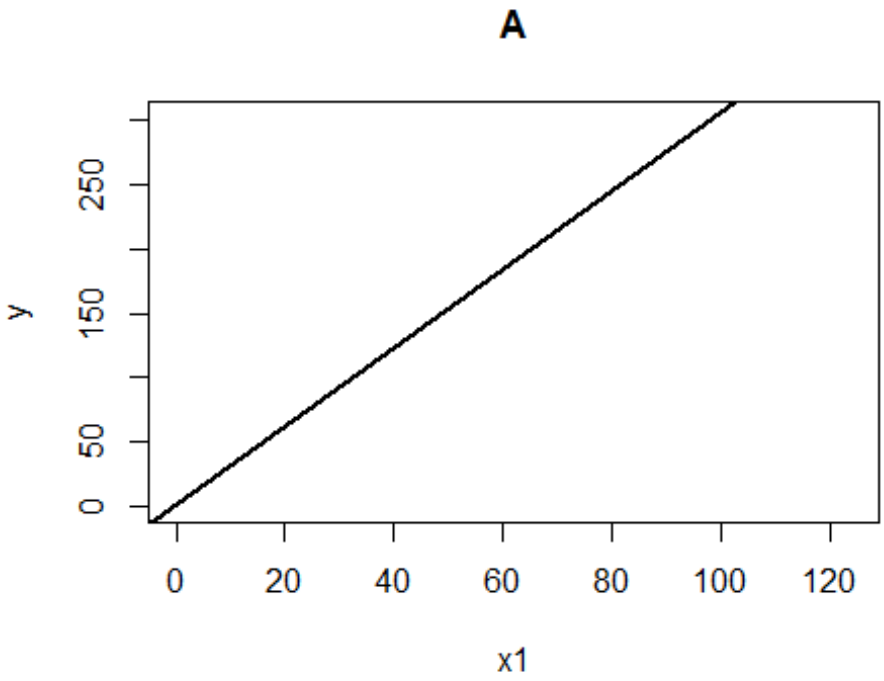
```
## lm(formula = y ~ x1 + x1:x_binary)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -193.265   -1.535   -1.535   -1.535  212.125
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.53490    0.10593   14.49   <2e-16 ***
## x1              3.04621    0.02923  104.23   <2e-16 ***
## x1:x_binaryYes       NA         NA      NA       NA
## ---
## 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
```

(B) Визначте кращу модель з пукту (A) Найкращі моделі 1 та 2

Завдання 6: Візуальне представлення нелінійної моделі де присутня взаємодія між неперервною та бінарною змінною.

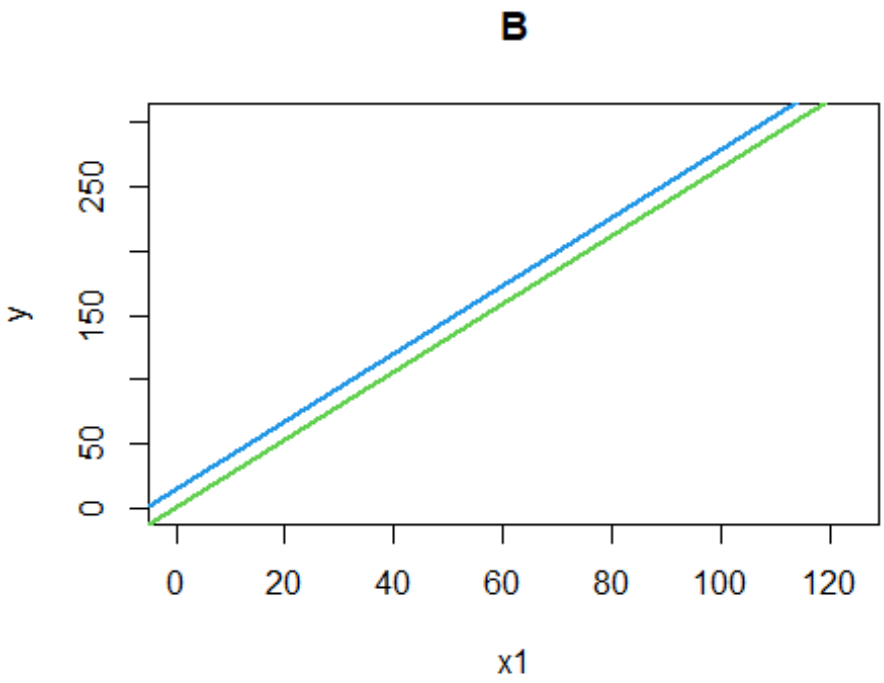
(A) Побудуйте розсіювання з налаштуванням груп. Варто використати такі налаштування col <- data$D + 3cex < -0.5 + 0.25 * dataD plot(y ~ x, data = data, col = col, pch = 16, cex = cex, main = "1")

```
col <- as.integer(x_binary) +3

## Warning: в результате преобразования созданы NA

cex <- 0.5 + 0.25*as.integer(x_binary)

## Warning: в результате преобразования созданы NA

mod_lin <- lm(y~x1)
plot(y ~ x1,  col = col, pch = 16, cex = cex, main = "A")
abline(coef=mod_lin$coefficients, lwd=2)
```



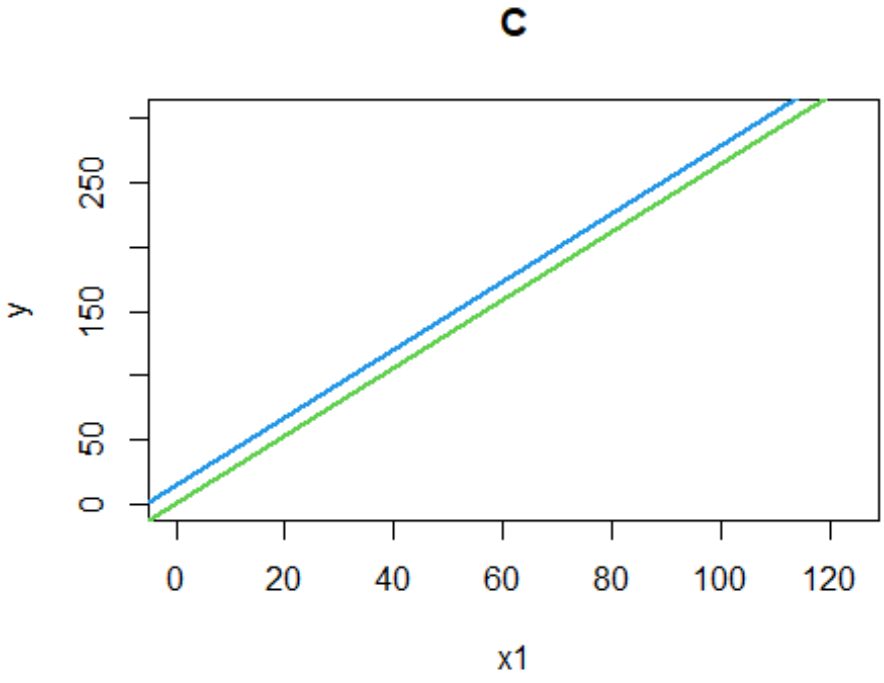
(B) Побудуйте розсіювання з бінарною змінною.

```
plot(y ~ x1,  col = col, pch = 16, cex = cex, main = "B")
abline(a=mod_1$coefficients[1],b=mod_1$coefficients[2],col=3, lwd=2)
abline(a=mod_1$coefficients[1] +
mod_1$coefficients[3],b=mod_1$coefficients[2],col=4, lwd=2)
```



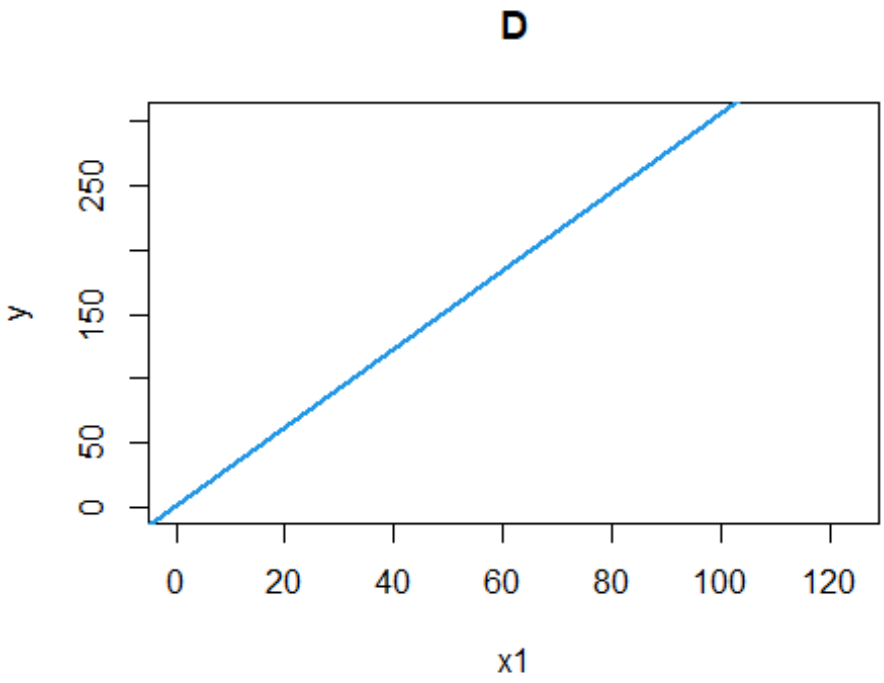
(C) Предиктор, фіктивна змінна та їх взаємодія. $Y = \beta 0 + \beta 1X + \beta 2D + \beta 3(X \cdot D) + \varepsilon$

```
plot(y ~ x1,  col = col, pch = 16, cex = cex, main = "C")
abline(a=mod_2$coefficients[1],b=mod_2$coefficients[2],col=3, lwd=2)
abline(a=mod_2$coefficients[1] + mod_2$coefficients[3],
       b=mod_2$coefficients[2] +0,col=4, lwd=2)
```



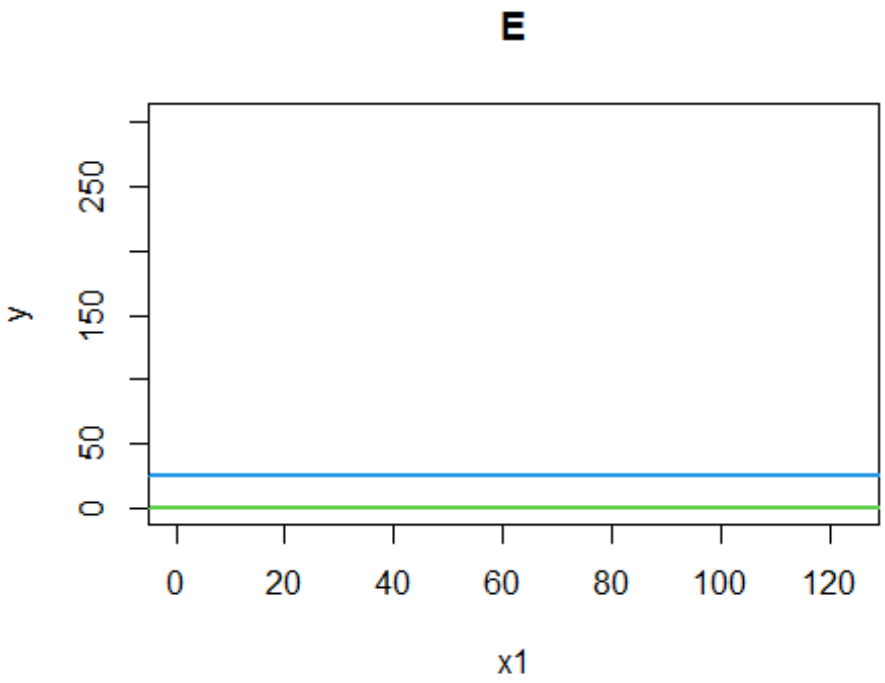
(D) Фіктивна змінна, присутня лише у взаємодії $Y = \beta 0 + \beta 1 X + \beta 2(X \cdot D) + \varepsilon$

```
plot(y ~ x1,  col = col, pch = 16, cex = cex, main = "D")
abline(a=mod_3$coefficients[1],b=mod_3$coefficients[2],col=3, lwd=2)
abline(a=mod_3$coefficients[1],
       b=mod_3$coefficients[2] +0,col=4, lwd=2)
```



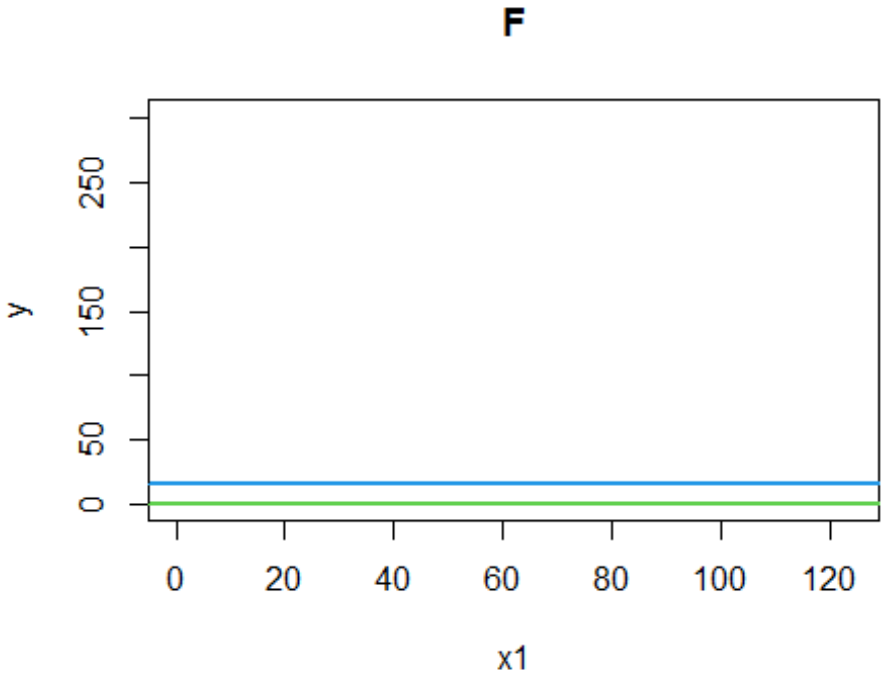
(E) Фіктивна змінна та відсутній предиктор$Y = \beta 0 + \beta 1 D + \varepsilon$

```
mod_4 <- lm(y~x_binary)
plot(y ~ x1,  col = col, pch = 16, cex = cex, main = "E")
abline(a=mod_4$coefficients[1],b=0,col=3, lwd=2)
abline(a=mod_4$coefficients[1]+mod_4$coefficients[2], b=0,col=4, lwd=2)
```



(F) Фіктивна змінна та взаємодія з предиктором.$Y = \beta 0 + \beta 1 D + \beta 2 (X \cdot D) + \varepsilon$

```
mod_5 <- lm(y~x_binary+ x1:x_binary)
plot(y ~ x1,  col = col, pch = 16, cex = cex, main = "F")
abline(a=mod_5$coefficients[1],b=0,col=3, lwd=2)
abline(a=mod_5$coefficients[1]+mod_5$coefficients[2], b=0,col=4, lwd=2)
```



(G) Взаємодія фіктивної змінної з предиктором.$Y = \beta 0 + \beta 1(X \cdot D) + \varepsilon$

```
mod_6 <- lm(y ~ x1:x_binary)
plot(y ~ x1,  col = col, pch = 16, cex = cex, main = "G")
abline(a=mod_6$coefficients[1],b=0,col=3, lwd=2)
abline(a=mod_6$coefficients[1], b=0, col=4, lwd=2)
```