Лабораторна робота № 7

Завдання:

(A) Описати модель зв'язку залежної змінної із категоріальною та незалежною змінною. Описати, який вплив категоріальної змінної на Y;
mod <- lm(Y ~ D + X, data)
summary(mod)

```
> #### (A) ####
> mean(data$fat)
[1] 1.012987
> data$D <- as.factor(data$fat > mean(data$fat))
> mod <- lm(rating ~ D + protein, data = data)
> summary(mod)

Call:
lm(formula = rating ~ D + protein, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.915  -6.997  -1.569   4.369  36.819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.371      3.229   8.478 1.61e-12 ***
DTRUE        -13.428      2.986  -4.497 2.50e-05 ***
protein        7.379      1.204   6.130 3.92e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.13 on 74 degrees of freedom
Multiple R-squared:  0.3886,    Adjusted R-squared:  0.3721
F-statistic: 23.51 on 2 and 74 DF,  p-value: 1.243e-08
```

$$Y = b_0 + b_1 D + b_2 X$$

$$\begin{cases} D = 1, & data\$fat > 1 \\ D = 0, & data\$fat \leq 1 \end{cases}$$

Отже, якщо x (жирність пластівців) починає перевищувати 1, то очікуване значення y (рейтингу) зменшується на 13.428. Тобто якщо x > 1, тоді D = 1 і $Y = b_0 + b_1 + b_2 X$ *, а якщо* $x \leq 1$*, то D = 0 і* $Y = b_0 + b_2 X$ *.*

(B) Для максимальної моделі визначити чи в оптимальній за критерієм Байєса моделі, залишається важливою категоріальна змінна. Описати її (тобто скільки додає до Y)
mod <- lm(Y ~ ., data)
modBIC <- MASS::stepAIC(mod, k = log(nrow(data)))

```
> modBIC <- MASS::stepAIC(modAll, k = log(nrow(data)))
Start:  AIC=5.45
rating ~ (calories + protein + fat + sodium + fiber + carbo +
    sugars + potass + vitamins + shelf + weight + cups + D) -
    fat

            Df Sum of Sq     RSS     AIC
- cups       1      0.13   39.82   1.355
- shelf      1      0.25   39.94   1.588
<none>                     39.69   5.447
- weight     1      5.02   44.72  10.280
- D          1     23.75   63.44  37.214
- vitamins   1     72.88  112.57  81.369
- sugars     1     91.00  130.69  92.862
- potass     1     97.07  136.76  96.359
- calories   1    273.73  313.42 160.214
- protein    1    445.30  485.00 193.831
- carbo      1    472.44  512.14 198.024
- fiber      1    669.04  708.73 223.040
- sodium     1   1150.59 1190.29 262.963
```

```
Step:  AIC=-2.64
rating ~ calories + protein + sodium + fiber + carbo + sugars +
    potass + vitamins + weight + D

           Df Sum of Sq     RSS     AIC
<none>                   40.00  -2.641
- weight    1      5.89   45.90   3.599
- D         1     26.74   66.74  32.427
- sugars    1     93.79  133.79  85.979
- vitamins  1     95.83  135.84  87.147
- potass    1    103.27  143.27  91.250
- calories  1    274.77  314.78 151.858
- protein   1    452.16  492.16 186.273
- carbo     1    489.45  529.45 191.897
- fiber     1    683.76  723.77 215.969
- sodium    1   1204.94 1244.94 257.732

Call:
lm(formula = rating ~ calories + protein + sodium + fiber + carbo +
    sugars + potass + vitamins + weight + D, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2093 -0.6002 -0.0496  0.5463  1.5912

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.502387   0.729465  73.345  < 2e-16 ***
calories    -0.283032   0.013293 -21.292  < 2e-16 ***
protein      3.507094   0.128404  27.313  < 2e-16 ***
sodium      -0.055059   0.001235 -44.587  < 2e-16 ***
fiber        3.649583   0.108659  33.587  < 2e-16 ***
carbo        1.255425   0.044179  28.417  < 2e-16 ***
sugars      -0.608438   0.048913 -12.439  < 2e-16 ***
potass      -0.046587   0.003569 -13.053  < 2e-16 ***
vitamins    -0.057661   0.004586 -12.574  < 2e-16 ***
weight       3.829385   1.227987   3.118  0.00269 **
DTRUE       -2.056097   0.309576  -6.642 6.99e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7785 on 66 degrees of freedom
Multiple R-squared:  0.9973,    Adjusted R-squared:  0.9969
F-statistic:  2468 on 10 and 66 DF,  p-value: < 2.2e-16
```

Якщо побудувати повну модель без змінної(fat) від якої ми утворили категоріальну змінну(D), то в оптимальній за критерієм Байєса моделі ця категоріальна змінна залишиться. Її AIC = 32.427.

Отже, якщо x (жирність пластівців) починає перевищувати 1, то очікуване значення y (рейтингу) зменшується на 2.056.

(C) Побудувати нелінійні моделі

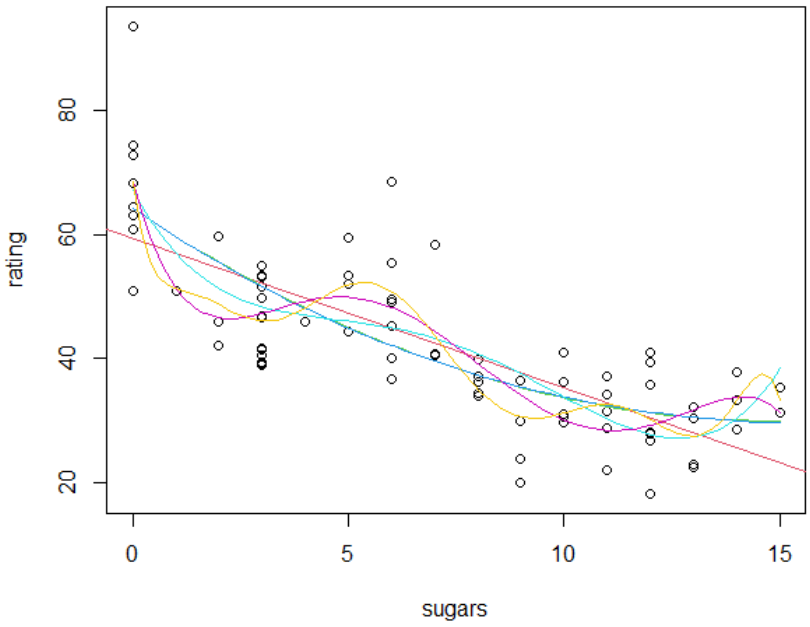
№	Модель	R^2	F	RSE
1	$y = b_0 + b_1 x$	0.5802	103.7	9.162
2	$y = b_0 b_1 x$	0.5931	109.3	0.2086
3	$y = b_0 e^{b_1 x}$	0.5931	109.3	0.2086
4	$y = e^{b_0 + b_1 x}$	0.5931	109.3	0.2086
5	$y = b_0 x^{b_1}$	0.5801	103.6	0.2119
6	$y = b_0 + b_1 \frac{1}{x}$	0.5299	84.54	9.695
7	$y = b_0 + b_1 x^2$	0.4619	64.37	10.37
8	$y = b_0 + b_1 x^3$	0.3714	44.31	11.21
9	$y = b_0 + b_1 \sqrt{x}$	0.6392	132.9	8.494

10	$y = b_0 + b_1 exp(x)$	0.05537	4.396	13.74
11	$y = b_0 + b_1 exp(-x)$	0.4382	58.5	10.6
12	$y = b_0 + b_1(X^3 - log(\|X\|) + 2^X$	0.1262	10.83	13.22

У моделі 9($y = b_0 + b_1\sqrt{x}$) найбільше значення R^2 і F, тому вона є найкращою. У моделях 3-5 значення R^2 і F не сильно менше ніж у моделі 9, а RSE набагато менше, тому вони також є непоганими.

(D) Побудуйте поліноми $Y = \beta_0 + \beta_1 X + \ldots + \beta_k X^k + \varepsilon$ до 5-го ступеня та 10-й. Визначте оптимальний поліном за допомогою BIC(*). Побудуйте розсіювання x та y, накладіть пряму лінію та поліноміальні моделі;

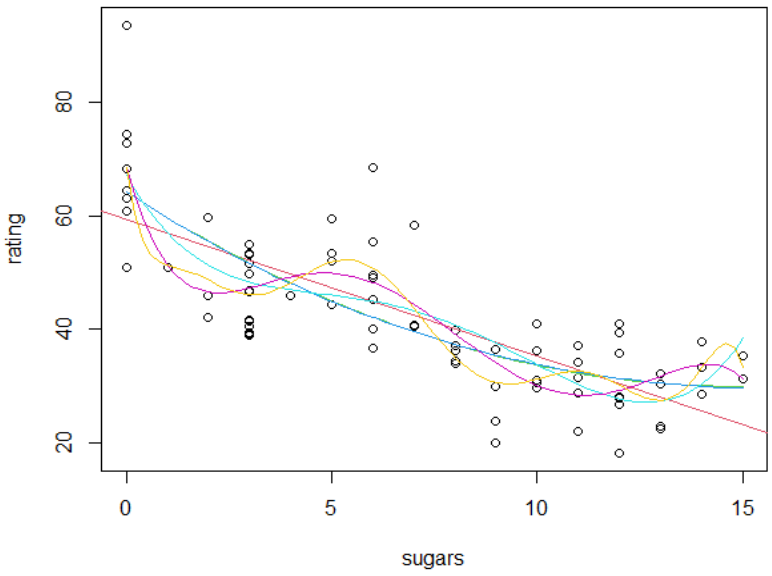
```
> mod1 <- lm(Y ~ X)
> plot(X,Y, xlab = "sugars", ylab = "rating")
> abline(coef = mod1$coefficients, col = 2)
> mod2 <- lm(Y ~ poly(X, degree = 2, raw = TRUE))
> d <- seq(0, 15, length.out = 77)
> lines(d, predict(mod2, new = data.frame(X = d)), col = 3)
> mod3 <- lm(Y ~ poly(X, degree = 3, raw = TRUE))
> lines(d, predict(mod3, new = data.frame(X = d)), col = 4)
> mod4 <- lm(Y ~ poly(X, degree = 4, raw = TRUE))
> lines(d, predict(mod4, new = data.frame(X = d)), col = 5)
> mod5 <- lm(Y ~ poly(X, degree = 5, raw = TRUE))
> lines(d, predict(mod5, new = data.frame(X = d)), col = 6)
> mod10 <- lm(Y ~ poly(X, degree = 10, raw = TRUE))
> lines(d, predict(mod10, new = data.frame(X = d)), col = 7)
```




```
> BIC(mod1, mod2, mod3, mod4, mod5, mod10)
       df      BIC
mod1    3 570.6347
mod2    4 567.7247
mod3    5 572.0508
mod4    6 565.2177
mod5    7 558.3124
mod10  12 572.1336
```

(E) Побудуйте ортогональні поліноми Лежанра до 5-го ступеня та 10-й. Визначте оптимальний поліном за допомогою BIC(*). Побудуйте розсіювання x та y, накладіть пряму лінію та поліноміальні моделі;

```
> mod1 <- lm(Y ~ X)
> plot(X,Y, xlab = "sugars", ylab = "rating")
> abline(coef = mod1$coefficients, col = 2)
> mod2 <- lm(Y ~ poly(X, degree = 2))
> d <- seq(0, 15, length.out = 77)
> lines(d, predict(mod2, new = data.frame(X = d)), col = 3)
> mod3 <- lm(Y ~ poly(X, degree = 3))
> lines(d, predict(mod3, new = data.frame(X = d)), col = 4)
> mod4 <- lm(Y ~ poly(X, degree = 4))
> lines(d, predict(mod4, new = data.frame(X = d)), col = 5)
> mod5 <- lm(Y ~ poly(X, degree = 5))
> lines(d, predict(mod5, new = data.frame(X = d)), col = 6)
> mod10 <- lm(Y ~ poly(X, degree = 10))
> lines(d, predict(mod10, new = data.frame(X = d)), col = 7)
```




```
> BIC(mod1, mod2, mod3, mod4, mod5, mod10)
       df      BIC
mod1    3 570.6347
mod2    4 567.7247
mod3    5 572.0508
mod4    6 565.2177
mod5    7 558.3124
mod10  12 572.1336
```

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

a. $y \sim x_1 * x_2$

```
> mod1 <- lm(Y~X1*X2)
> summary(mod1)

Call:
lm(formula = Y ~ X1 * X2)

Residuals:
    Min      1Q  Median      3Q     Max
-14.404  -4.540  -1.270   5.024  16.640

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.894170   2.339078  20.476  < 2e-16 ***
X1          -1.850082   0.276373  -6.694 3.82e-09 ***
X2           0.124035   0.019531   6.351 1.63e-08 ***
X1:X2       -0.006413   0.002248  -2.853  0.00563 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.956 on 73 degrees of freedom
Multiple R-squared:  0.7645,    Adjusted R-squared:  0.7548
F-statistic: 78.99 on 3 and 73 DF,  p-value: < 2.2e-16
```

b. $y \sim x_1 * x_2 * x_3$

```
> mod2 <- lm(Y~X1*X2*X3)
> summary(mod2)

Call:
lm(formula = Y ~ X1 * X2 * X3)

Residuals:
    Min      1Q  Median      3Q     Max
-10.450  -4.464  -0.948   4.019  17.410

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 77.9091898  8.5008369   9.165  1.5e-13 ***
X1           1.0948067  1.9848848   0.552  0.58302
X2           0.0854559  0.0526070   1.624  0.10885
X3          -0.3078528  0.0912911  -3.372  0.00123 **
X1:X2       -0.0297203  0.0116856  -2.543  0.01323 *
X1:X3       -0.0227218  0.0180753  -1.257  0.21297
X2:X3        0.0002881  0.0007146   0.403  0.68807
X1:X2:X3     0.0002158  0.0001049   2.057  0.04346 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.894 on 69 degrees of freedom
Multiple R-squared:  0.8402,    Adjusted R-squared:  0.824
F-statistic: 51.82 on 7 and 69 DF,  p-value: < 2.2e-16
```

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

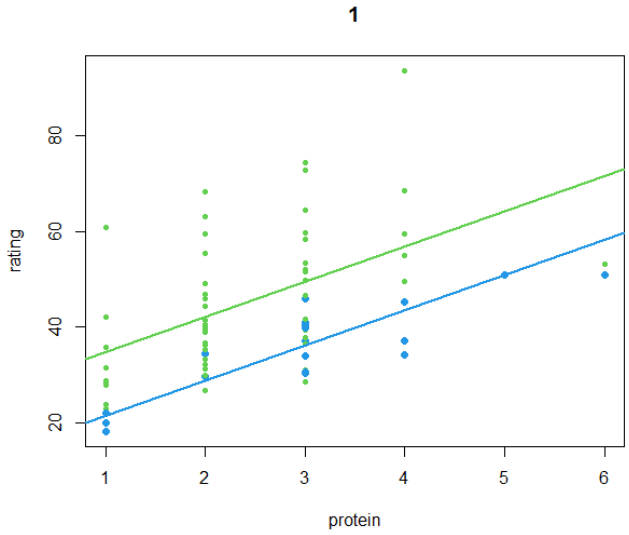
```
> modIntBIC <- MASS::stepAIC(object = lm(rating ~ .,data=data),
+                            scope = rating ~ .^2, k = log(nobs(modBIC)), trace = 0)
> summary(modIntBIC)

Call:
lm(formula = rating ~ calories + protein + fat + sodium + fiber +
    carbo + sugars + potass + vitamins + shelf + weight + cups +
    protein:carbo + carbo:shelf + sugars:shelf + sodium:carbo +
    protein:sugars + shelf:weight + fiber:carbo + fiber:shelf +
    protein:weight + protein:shelf + potass:weight + vitamins:cups +
    protein:fiber + vitamins:shelf, data = data)

Residuals:
       Min         1Q     Median         3Q        Max
-0.0089063 -0.0033416 -0.0001595  0.0030167  0.0128037

Coefficients:
                  Estimate Std. Error   t value Pr(>|t|)
(Intercept)     5.486e+01  5.419e-02  1012.367  < 2e-16 ***
calories       -2.230e-01  1.705e-04 -1307.930  < 2e-16 ***
protein         3.307e+00  1.577e-02   209.633  < 2e-16 ***
fat            -1.688e+00  1.955e-03  -863.116  < 2e-16 ***
sodium         -5.420e-02  4.270e-05 -1269.382  < 2e-16 ***
fiber           3.459e+00  5.092e-03   679.390  < 2e-16 ***
carbo           1.102e+00  2.681e-03   410.845  < 2e-16 ***
sugars         -7.186e-01  2.311e-03  -310.884  < 2e-16 ***
potass         -3.463e-02  1.501e-04  -230.648  < 2e-16 ***
vitamins       -4.989e-02  4.572e-04  -109.112  < 2e-16 ***
shelf           1.769e-03  1.596e-02     0.111  0.91215
weight         -1.557e-01  7.422e-02    -2.098  0.04095 *
cups            2.459e-02  1.029e-02     2.389  0.02071 *
protein:carbo   3.621e-03  5.054e-04     7.164 3.33e-09 ***
carbo:shelf    -5.416e-03  5.238e-04   -10.340 5.16e-14 ***
sugars:shelf   -4.901e-03  6.297e-04    -7.783 3.60e-10 ***
sodium:carbo   -1.956e-05  2.792e-06    -7.006 5.88e-09 ***
protein:sugars  2.415e-03  4.300e-04     5.617 8.56e-07 ***
shelf:weight    1.399e-01  2.097e-02     6.669 1.98e-08 ***
fiber:carbo    -5.595e-04  1.639e-04    -3.413  0.00128 **
fiber:shelf    -5.627e-03  1.291e-03    -4.359 6.51e-05 ***
protein:weight -9.183e-02  1.636e-02    -5.614 8.65e-07 ***
protein:shelf  -4.223e-03  1.565e-03    -2.699  0.00946 **
potass:weight   4.677e-04  1.349e-04     3.467  0.00109 **
vitamins:cups  -7.074e-04  3.692e-04    -1.916  0.06106 .
protein:fiber   2.469e-03  1.029e-03     2.400  0.02015 *
vitamins:shelf -2.393e-04  1.363e-04    -1.755  0.08532 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.005886 on 50 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:      1
F-statistic: 1.665e+07 on 26 and 50 DF,  p-value: < 2.2e-16
```

Між моделями a і b, краще модель b, оскільки у неї більше R^2 *і менше RSE. Проте модель c краще ніж a і b, у неї дуже гарні значення* R^2*, RSE і F.*

(G) Побудуйте взаємодія між неперервною та бінарною змінною. Визначте яка краща:

a. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$

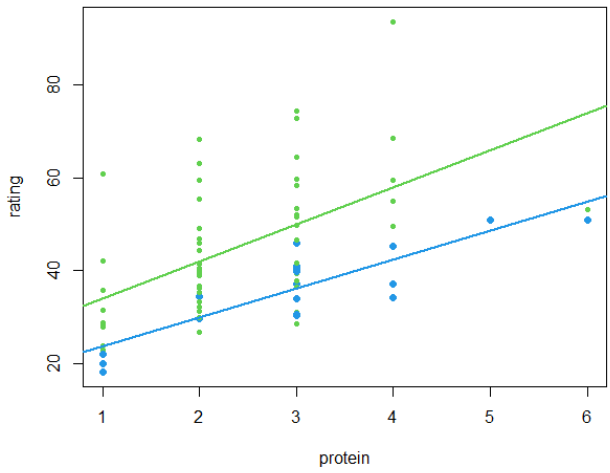
```
> mod1<- lm(rating ~ protein+D, data = data)
> summary(mod1)

Call:
lm(formula = rating ~ protein + D, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-20.915  -6.997  -1.569   4.369  36.819

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   27.371      3.229   8.478 1.61e-12 ***
protein        7.379      1.204   6.130 3.92e-08 ***
DTRUE        -13.428      2.986  -4.497 2.50e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.13 on 74 degrees of freedom
Multiple R-squared:  0.3886,    Adjusted R-squared:  0.3721
F-statistic: 23.51 on 2 and 74 DF,  p-value: 1.243e-08
```



b. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$

```
> mod2 <- lm(rating ~ protein * D, data = data)
> summary(mod2)

Call:
lm(formula = rating ~ protein * D, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-21.293  -6.144  -1.785   4.369  35.825

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    25.903      3.871   6.691 3.88e-09 ***
protein         7.994      1.499   5.332 1.04e-06 ***
DTRUE          -8.542      7.659  -1.115    0.268
protein:DTRUE  -1.755      2.531  -0.693    0.490
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.17 on 73 degrees of freedom
Multiple R-squared:  0.3926,    Adjusted R-squared:  0.3676
F-statistic: 15.73 on 3 and 73 DF,  p-value: 5.51e-08
```

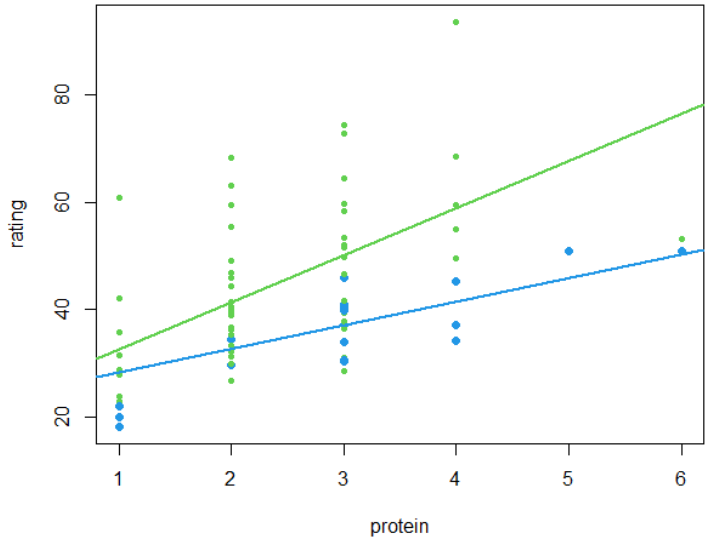
2



c. $Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$

```
> mod3 <- lm(rating ~ protein + protein:D, data = data)
> summary(mod3)

Call:
lm(formula = rating ~ protein + protein:D, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-23.241  -7.271  -1.165   3.918  34.883

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    23.720      3.346   7.089 6.69e-10 ***
protein         8.775      1.328   6.608 5.24e-09 ***
protein:DTRUE  -4.353      0.992  -4.388 3.74e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11.19 on 74 degrees of freedom
Multiple R-squared:  0.3822,    Adjusted R-squared:  0.3655
F-statistic: 22.89 on 2 and 74 DF,  p-value: 1.823e-08
```

3



Модель а трошки краще ніж інші, оскільки вона має трошки менший RSE і трошки більший F і adjusted R^2. Проте моделі не сильно відрізняються від один одного.