Лабораторна робота № 4

Для вхідних даних користуємось кейсом який ви обрали.

Роботу виконуємо в R. Опишіть ваші дії, припущення та висновки до кожного пункту.

Завдання: Для множинної лінійної регресійної моделі представити прогноз та довірчі інтервали.

(A) Використовуючи комп'ютерне програмне забезпечення RStudio побудуйте множинну лінійну регресійну модель для вашого кейсу за 4-ма факторами.

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4;$$

```
> # Залежна змінна rating
> Y <- data$rating
> x1 <- data$calories
> x2 <- data$sodium
> x3 <- data$protein
> x4 <- data$cups
> mod <- lm(Y ~ x1+x2+x3+x4)
```

```
> sumMod <- summary(mod)
> sumMod

Call:
lm(formula = Y ~ x1 + x2 + x3 + x4)

Residuals:
    Min      1Q  Median      3Q     Max 
-14.782  -5.951  -0.516   4.028  16.299 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 82.21044    5.96959  13.772  < 2e-16 ***
x1          -0.46359    0.04560 -10.166 1.47e-15 ***
x2          -0.03033    0.01063  -2.853  0.00564 ** 
x3           6.03616    0.79740   7.570 9.72e-11 ***
x4          -0.62914    3.77569  -0.167  0.86813    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.367 on 72 degrees of freedom
Multiple R-squared:  0.7394,	Adjusted R-squared:  0.725 
F-statistic: 51.08 on 4 and 72 DF,  p-value: < 2.2e-16
```

(B) Зробити аналіз за допомогою критерію Фішера (F-тест);

$$H_0: \beta_j = 0, \;\; j = 1, \dots, p \qquad F < F_{\alpha k_1 k_2}$$

$$k_1 = p, \quad k_2 = n - p - 1;$$

```
F-statistic: 51.08 on 4 and 72 DF,  p-value: < 2.2e-16
```

a. **Вказати k_1**

$k_1 = 4$

b. **k_2**

$k_2 = 72$

c. **F**

$F = 51.08$

d. **$F_{\alpha k_1 k_2}$**

$F_{\alpha k_1 k_2} \approx 2.75$(отримано з таблиці)

e. **Висновки**

Оскільки $F > F_{\alpha k_1 k_2}$ і p-value < $\alpha(0.05)$, то ми відхиляємо нульову гіпотезу і приймаємо альтернативну.

(C) Перевірка t-статистики для 5-ти коефіцієнтів регресії:

$$H_0\colon \beta_j = 0 \text{ vs } H_1\colon \beta_j \neq 0$$

$$t_j = \frac{\hat{\beta}_j - 0}{\widehat{\mathrm{SE}}(\hat{\beta}_j)}$$

$$-t_{\text{кр}} > t_j > t_{\text{кр}}$$

a. Вказати $\boldsymbol{k}$

Df = 72

b. $\boldsymbol{t_j}$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 82.21044    5.96959  13.772  < 2e-16 ***
x1          -0.46359    0.04560 -10.166 1.47e-15 ***
x2          -0.03033    0.01063  -2.853  0.00564 ** 
x3           6.03616    0.79740   7.570 9.72e-11 ***
x4          -0.62914    3.77569  -0.167  0.86813    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c. $\boldsymbol{t}_{\text{кр}}$

```
> # Знаходимо t_k для моделі
> alpha <- 0.05
> n <- length(data$calories)
> p <- 4
> t_k <- qt(p = 1 - alpha / 2, df = n - p - 1)
> t_k
[1] 1.993464
```

d. Висновки

- t value для b_0 = 13.772, воно більше за $t_{\text{кр}}$ і p-value менше за 0.05, тому ми відхиляємо нульову гіпотезу і приймаємо альтернативну.
- t value для b_1 = -10.166, воно менше за -$t_{\text{кр}}$ і p-value менше за 0.05, тому ми відхиляємо нульову гіпотезу і приймаємо альтернативну.
- t value для b_2 = -2.853, воно менше за -$t_{\text{кр}}$ і p-value менше за 0.05, тому ми відхиляємо нульову гіпотезу і приймаємо альтернативну.
- t value для b_3 = 7.570, воно більше за $t_{\text{кр}}$ і p-value менше за 0.05, тому ми відхиляємо нульову гіпотезу і приймаємо альтернативну.
- t value для b_4 = -0.167, воно знаходиться в межах від -$t_{\text{кр}}$ до $t_{\text{кр}}$ і p-value більше за 0.05, тому ми приймаємо нульову гіпотезу.

(D) Знайти та записати довірчі інтервали для 5-ти коефіцієнтів множинної регресії:

$$\left(\hat{\beta}_j \pm \widehat{\mathrm{SE}}(\hat{\beta}_j) t_{n-p-1;\alpha/2}\right)$$

Вкажіть α – рівень значущості $0.1, 0.05, 0.01$;

$df = n - p - 1$ – ступені вільності;

$df = 77 - 4 - 1 = 72$

```
> #### (D) ####
>
> confint(mod, level = 0.90)
                    5 %        95 %
(Intercept) 72.26334670 92.15752674
x1          -0.53957999 -0.38760118
x2          -0.04804191 -0.01261784
x3           4.70745344  7.36485760
x4          -6.92055498  5.66226668
>
> confint(mod, level = 0.95)
                 2.5 %      97.5 %
(Intercept) 70.3102767 94.110596690
x1          -0.5545002 -0.372680975
x2          -0.0515196 -0.009140152
x3           4.4465683  7.625742750
x4          -8.1558475  6.897559153
>
> confint(mod, level = 0.99)
                   0.5 %      99.5 %
(Intercept)  66.41578586 98.005087575
x1           -0.58425161 -0.342929556
x2           -0.05845423 -0.002205524
x3            3.92635402  8.145957016
x4          -10.61906454  9.360776241
```

(E) Знайти та записати довірчі інтервали для регресійних значень $\widehat{Y}$:

$$\hat{y}_i - \Delta\hat{y}_i < y_i < \hat{y}_i + \Delta\hat{y}_i$$

```
> predict(mod, interval = "confidence", level = 0.90)
       fit      lwr      upr
1 69.75322 65.43130 74.07513
2 43.60394 39.97826 47.22962
3 65.81033 60.61962 71.00104
4 78.61477 73.57970 83.64985
5 36.74995 34.91968 38.58022
```

```
> predict(mod, interval = "confidence", level = 0.95)
       fit      lwr      upr
1 69.75322 64.58272 74.92372
2 43.60394 39.26638 47.94151
3 65.81033 59.60045 72.02022
4 78.61477 72.59108 84.63847
5 36.74995 34.56031 38.93959
```

```
> predict(mod, interval = "confidence", level = 0.99)
       fit      lwr      upr
1 69.75322 62.89060 76.61584
2 43.60394 37.84685 49.36103
3 65.81033 57.56818 74.05249
4 78.61477 70.61974 86.60981
5 36.74995 33.84372 39.65618
```

На скрінах показані лише перші 5 значень, в R можна побачити усі 77

(F) Зробіть прогноз для середнього $\hat{y}_i$ та для $\hat{y}_p$ ($p = max + 10\%$) на наступний період. Опишіть для яких змінних і які значення беруться для прогнозу;

$$\hat{y}_p = a + bx_p$$

$$\hat{y}_p - \Delta\hat{y}_p < y_p < \hat{y}_p + \Delta\hat{y}_p;$$

a. Вказати середнє значення $\hat{y}_i$

```
> mean(x1)
[1] 106.8831
> mean(x2)
[1] 159.6753
> mean(x3)
[1] 2.545455
> mean(x4)
[1] 0.821039
> dataNewMean <- data.frame(x1 = 106.8831,x2=159.6753 ,x3=2.545455, x4 = 0.821039)
```

Знаходимо середнє для іксів і створюємо нову data

b. $fit \quad lwr \quad upr$ для a.

```
> predict(mod, newdata = dataNewMean, interval = "confidence")
       fit      lwr      upr
1 42.66572 40.99209 44.33934
```

c. $\hat{y}_p$ ($p = max + 10\%$)

```
> max(x1)*1.1
[1] 176
> max(x2)*1.1
[1] 352
> max(x3)*1.1
[1] 6.6
> max(x4)*1.1
[1] 1.65
> dataNewMax <- data.frame(x1 = 176,x2=352 ,x3=6.6, x4= 1.65)
```

Знаходимо максимальне + 10% для іксів і створюємо нову data

d. $fit \quad lwr \quad upr$ для c.

```
> predict(mod, newdata = dataNewMax, interval = "prediction")
       fit      lwr      upr
1 28.74292 10.09019 47.39564
```