Лабораторна робота №3

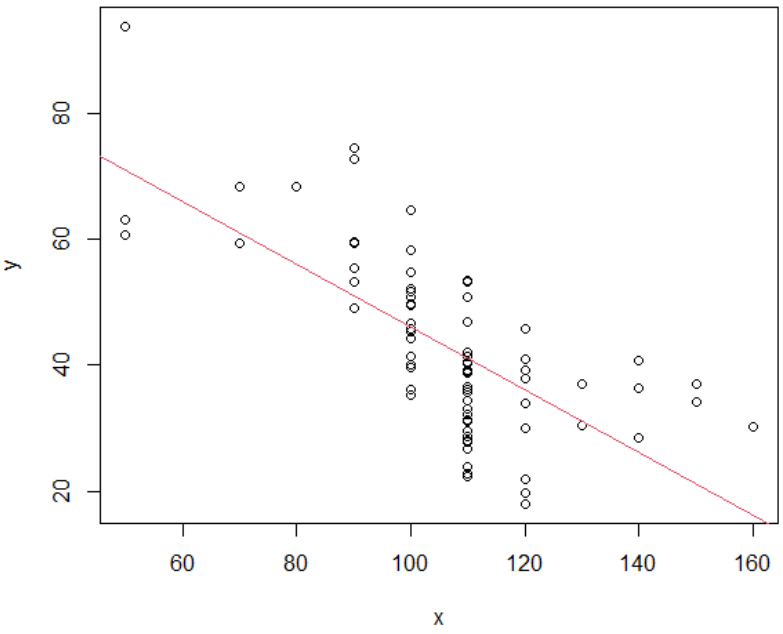
A. Для факторів $x1$, $x2$, $x3$ та y побудувати в R модель mi: lm(y~x_i) та виконати наступні завдання
a. - l. та зробити висновки щодо припущень 1, 2, 3, 4 для кожної пари відповідно.

Змінна X1

```
# Незалежна змінна x1 - калорій на порцію
x1 <- data$calories
mod1 <- lm(Y ~ x1)
summod1<-summary(mod1)
summary(mod1)
```

a) Побудувати діаграму розсіювання plot(*$$x_i$$, *$y) та накласти регресійну лінію.

```
> # діаграма розсіювання та регресійна лінія
> plot(x1, Y, xlab="x", ylab = "y")
> abline(mod1, col=2)
```



b) Перевірити значення *$r.squared та зробити висновки;

```
> summod1$r.squared
[1] 0.4752393
```

Залежна змінна пояснюється незалежною змінною x1 на 47.5%

c) Перевірити sum(*$residuals^2) та зробити висновки;

```
> sum(summod1$residuals^2)
[1] 7869.731
```

Сума квадратів відхилень достатньо велика

d) Обчислити var(*$$x_i$$);

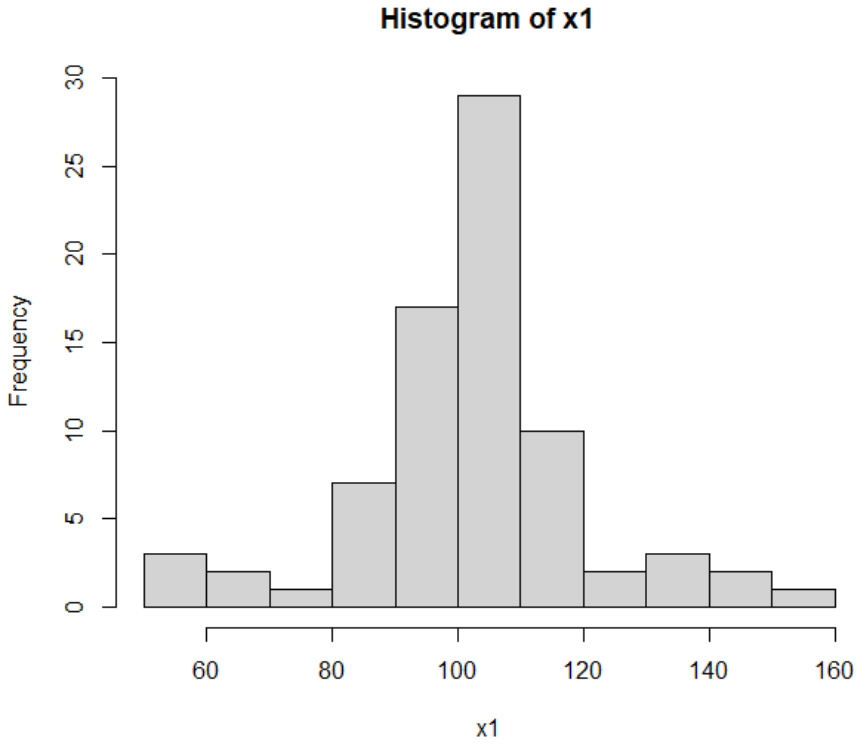
```
> var(x1)
[1] 379.6309
```

e) Обчислити var(*$y);

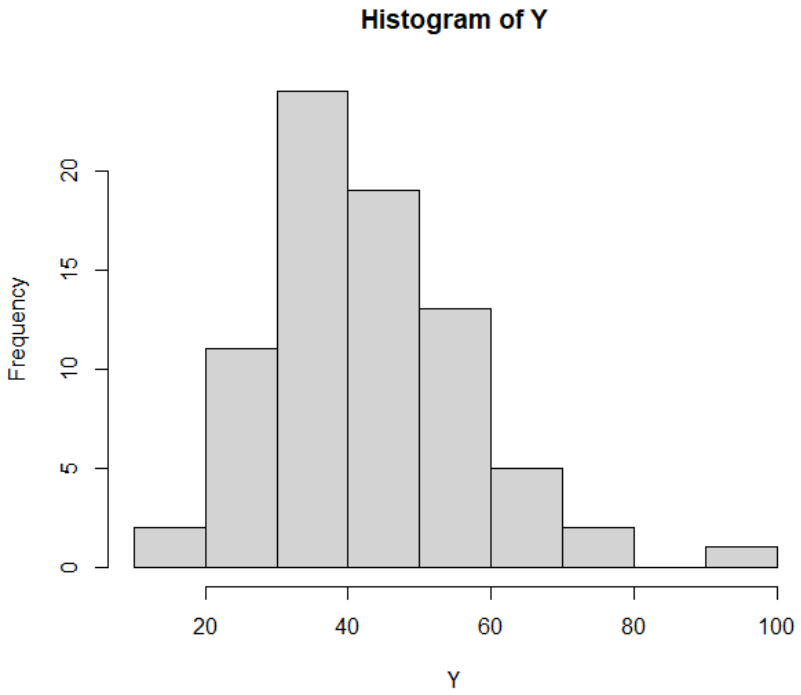
```
> var(Y)
[1] 197.3263
```

Варіація x більше, ніж варіація y, що добре

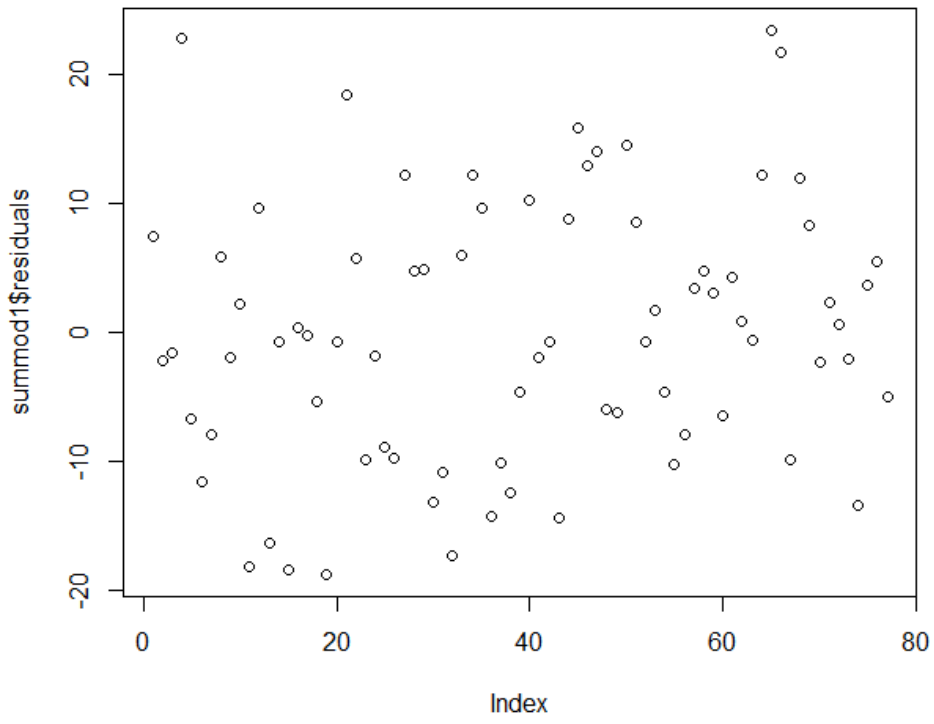
f) Побудувати hist(*$$x_i$$);



g) Побудувати hist(*$y);



h) Побудувати plot(*$residuals) зробити припущення чи відповідає $N(0; 1)$;



Схоже на нормальний розподіл, тобто на N(0, 1)

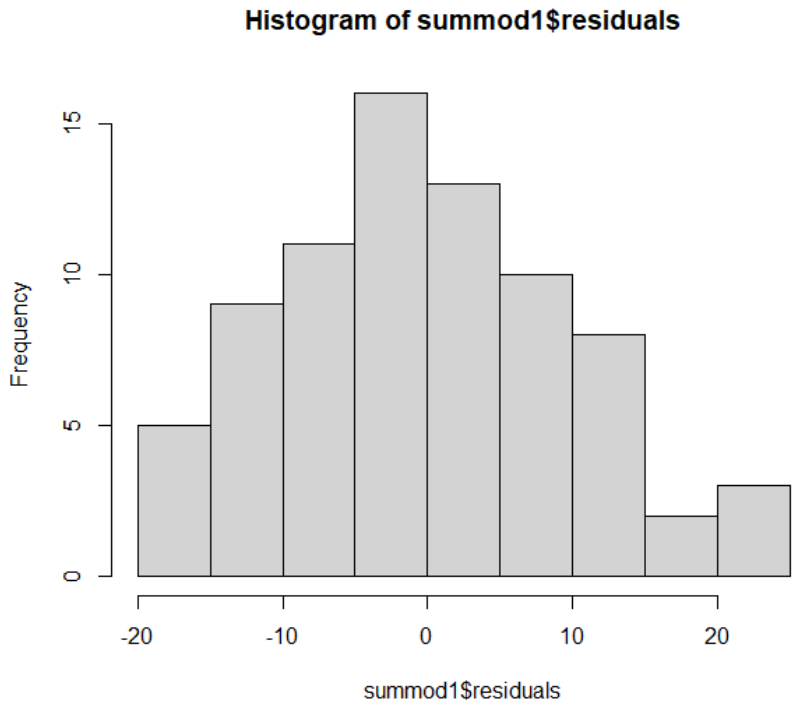
i) Перевірити mean(*$residuals);

```
> mean(summod1$residuals)
[1] 2.442693e-16
```

j) Обчислити var(*$residuals);

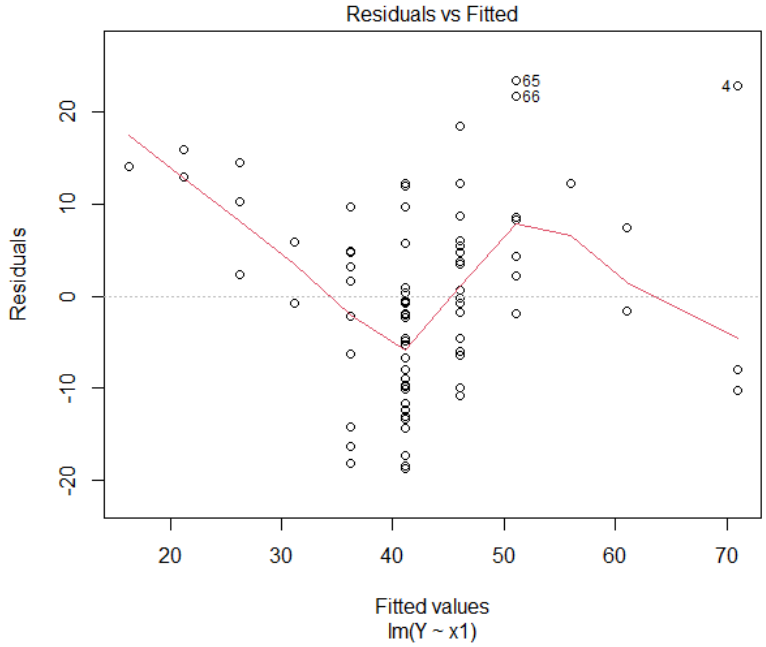
```
> var(summod1$residuals)
[1] 103.5491
```

k) Побудувати hist(*$residuals) та перевірити чи відповідає $N(0; 1)$;

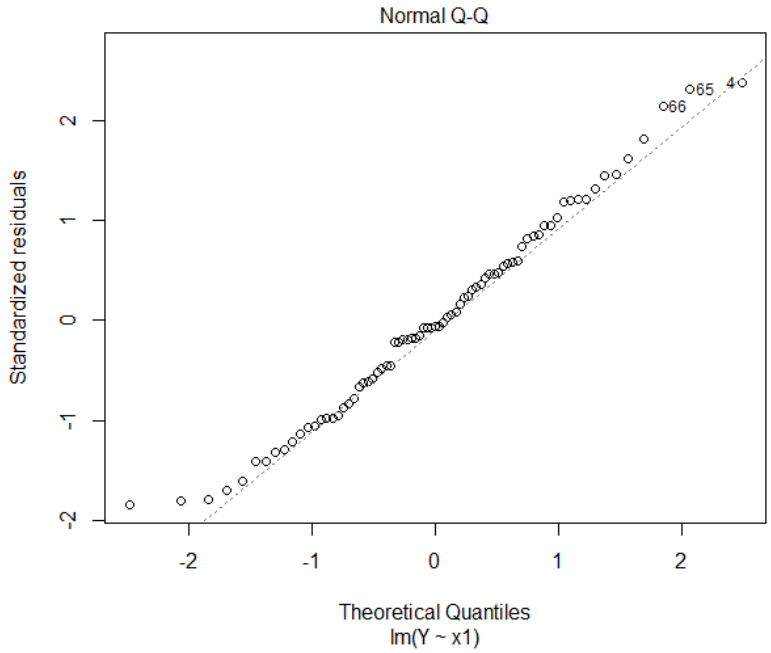


Схоже на нормальний розподіл, але хвости різні

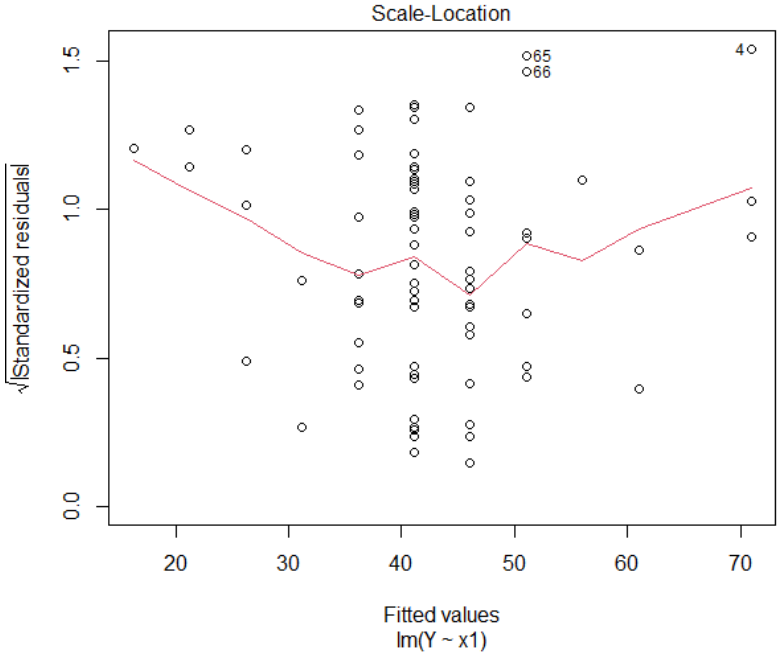
l) **Виконати перевірку 4-х припущень для МНК за допомогою plot(mod_AGST, 1), plot(mod_AGST, 2), plot(mod_AGST, 3), plot(mod_AGST$residuals, type = "o").**



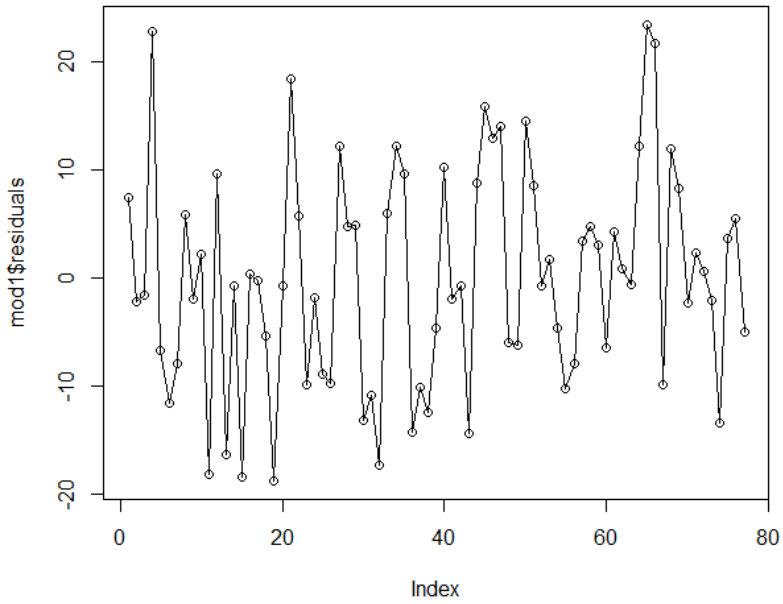
Оскільки червона лінія не близько до пунктирної, то лінійність не дотримується



Майже усі значення лежать на прямій, що означає що дані розподілені нормально



З графіку видно що припущення про гомоскедастичність не виконується, тобто ми маємо гетероскедастичність

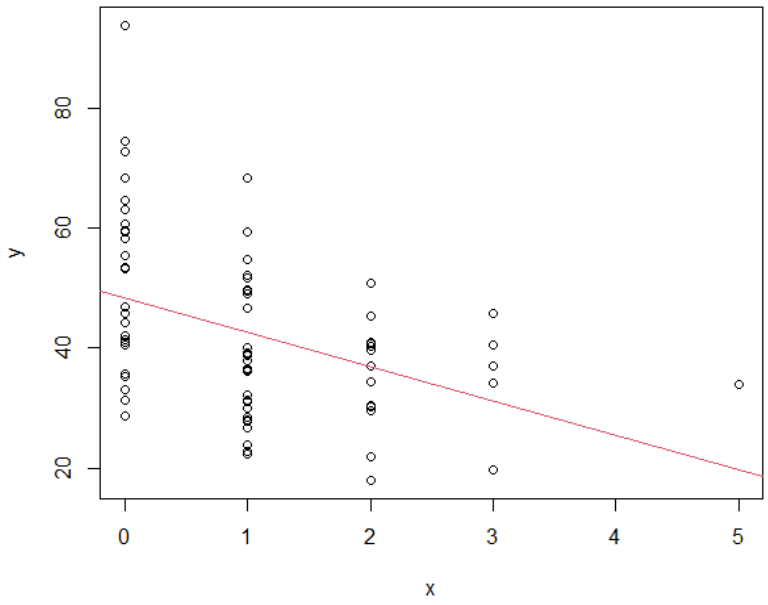


Графік майже повністю заповнює простір, це означає що припущення про незалежність виконується

Змінна X2

```
# Незалежна змінна x2 - грамів жиру
x2 <- data$fat
mod2 <- lm(Y ~ x2)
summod2<-summary(mod2)
summary(mod2)
```

a) Побудувати діаграму розсіювання plot(*$$x_i$, *$y) та накласти регресійну лінію;



b) Перевірити значення *$r.squared та зробити висновки;

```
> summod2$r.squared
[1] 0.1675131
```

Залежна змінна пояснюється незалежною змінною x1 на 16.7%

c) Перевірити sum(*$residuals^2) та зробити висновки;

```
> sum(summod2$residuals^2)
[1] 12484.64
```

Сума квадратів відхилень достатньо велика

d) Обчислити var(*$$x_i$);

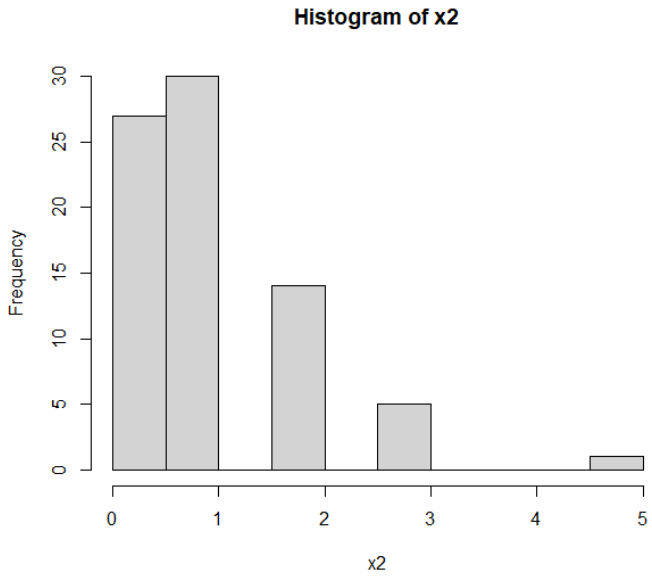
```
> var(x2)
[1] 1.012987
```

e) Обчислити var(*$y);

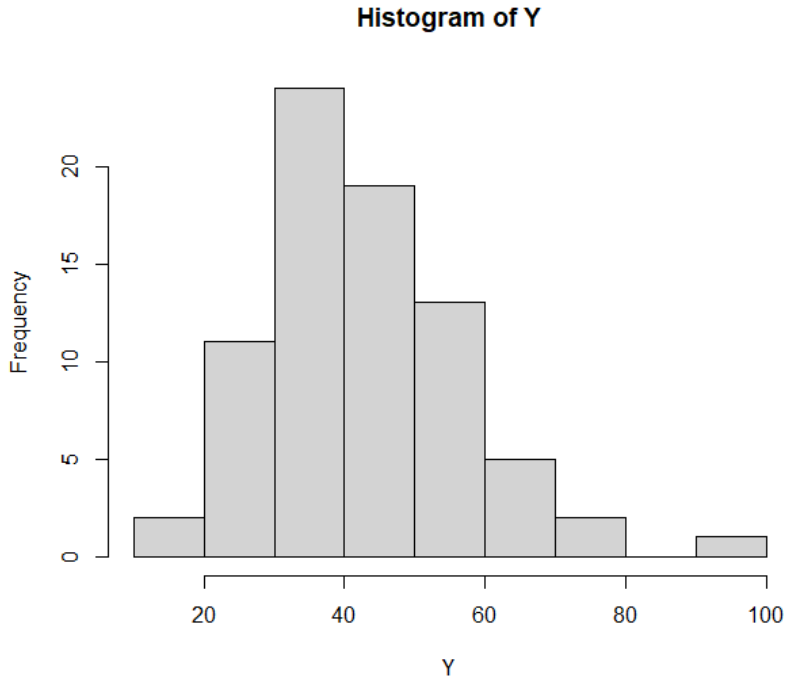
```
> var(Y)
[1] 197.3263
```

Варіація у більше, ніж варіація x, що погано

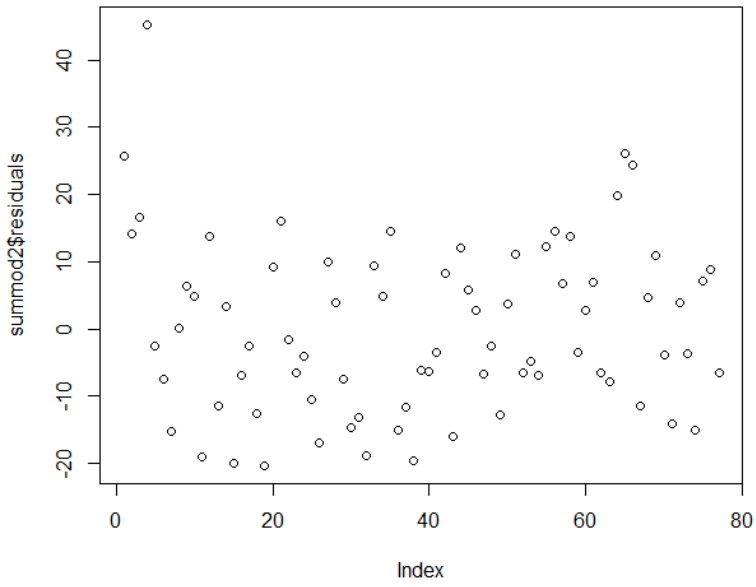
f) Побудувати hist(*$$x_i$$);



g) Побудувати hist(*$y);



h) Побудувати plot(*$residuals) зробити припущення чи відповідає $$N(0; 1)$$;



Схоже було б на нормальний розподіл, тобто на N(0, 1), але є викид

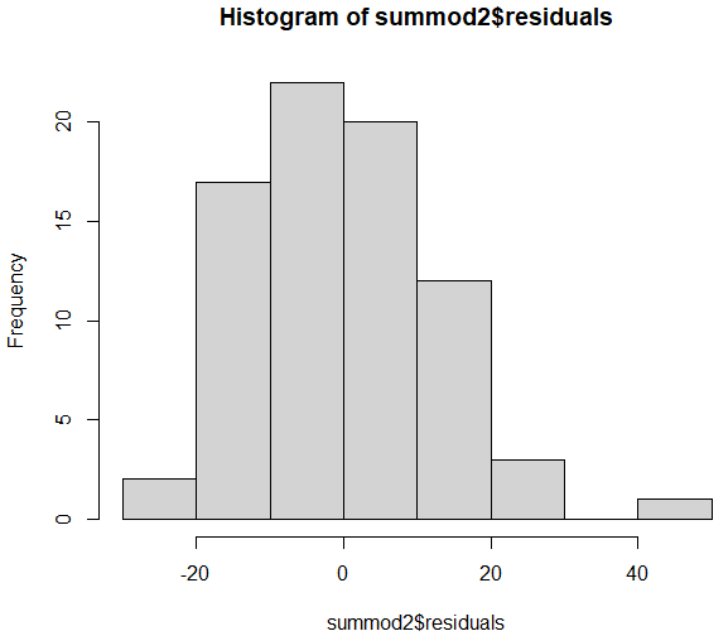
i) Перевірити mean(*$residuals);

```
> mean(summod2$residuals)
[1] -2.072882e-16
```

j) Обчислити var(*$residuals);

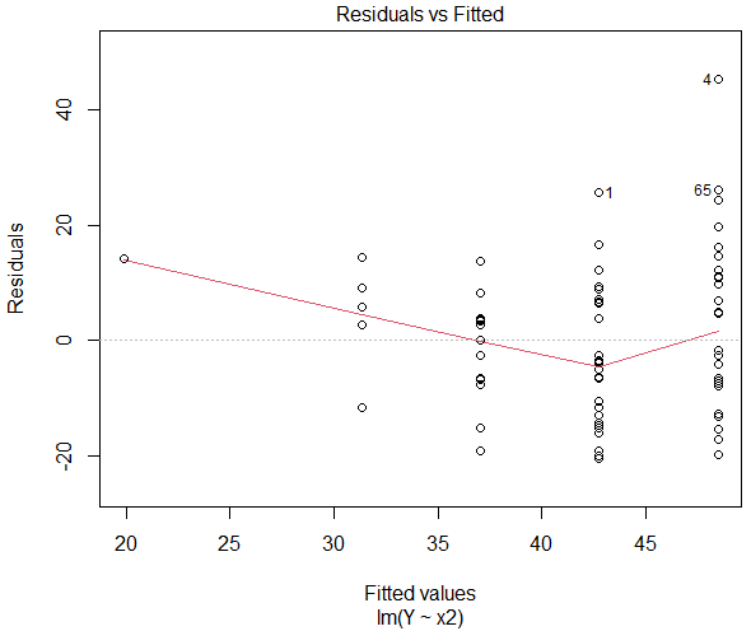
```
> var(summod2$residuals)
[1] 164.2716
```

k) Побудувати hist(*$residuals) та перевірити чи відповідає $N(0; 1)$;

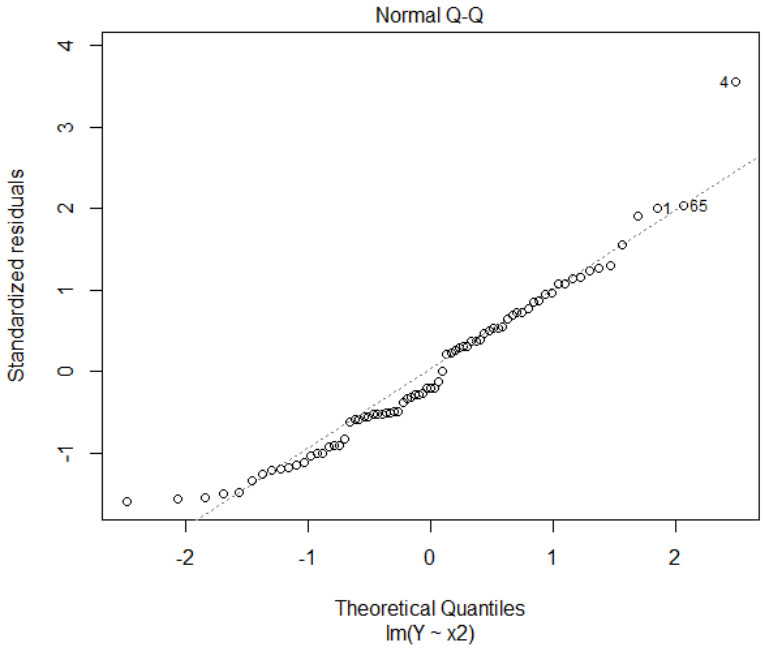


Схоже на нормальний розподіл, але на правому хвості є викид

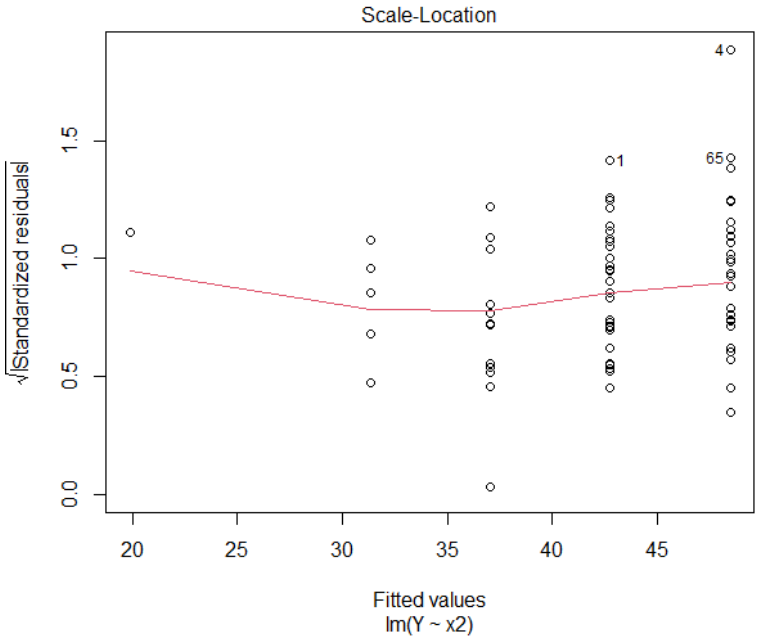
l) Виконати перевірку 4-х припущень для МНК за допомогою plot(mod_AGST, 1), plot(mod_AGST, 2), plot(mod_AGST, 3), plot(mod_AGST$residuals, type = "o");



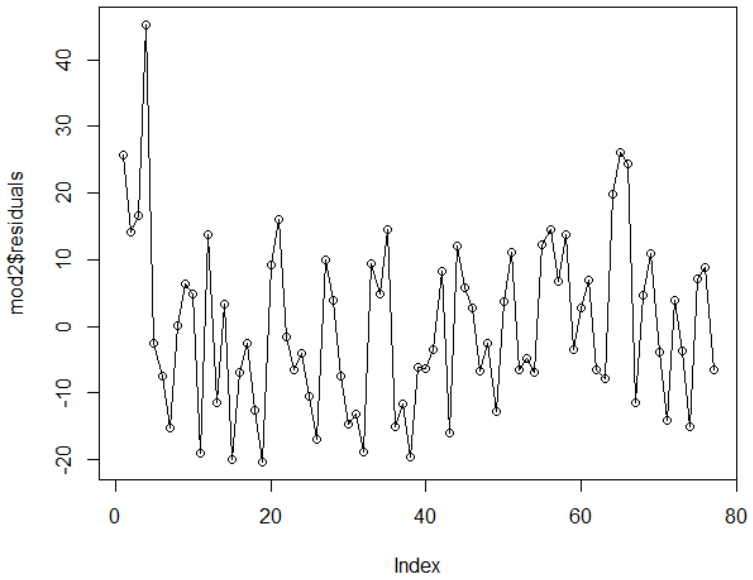
Оскільки червона лінія не близько до пунктирної, то лінійність не дотримується



Майже усі значення лежать на прямій, окрім одного викиду, що означає що данні розподілені нормально



З графіку схоже, що припущення про гомоскедастичність виконується

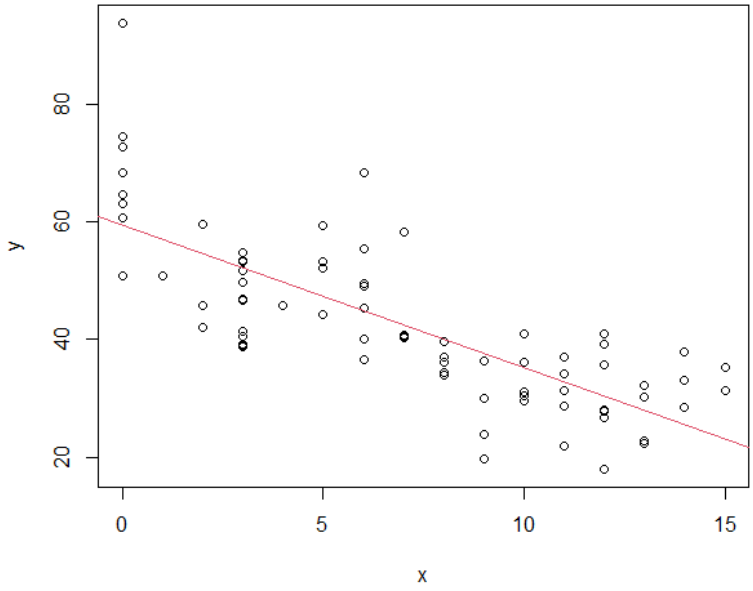


Спочатку є викид, якщо його не було то виконувалось б припущення про незалежність похибок

Змінна X3

```
# Незалежна змінна x3 - грам цукрів
x3 <- data$sugars
mod3 <- lm(Y ~ x3)
summod3<-summary(mod3)
summary(mod3)
```

а) Побудувати діаграму розсіювання plot(*$$x_i$$, *$y) та накласти регресійну лінію;



b) Перевірити значення *$r.squared та зробити висновки;

```
> summod3$r.squared
[1] 0.5802363
```

Залежна змінна пояснюється незалежною змінною x1 на 58%

c) Перевірити sum(*$residuals^2) та зробити висновки;

```
> sum(summod3$residuals^2)
[1] 6295.113
```

Сума квадратів відхилень достатньо велика

d) Обчислити var(*$$x_i$$);

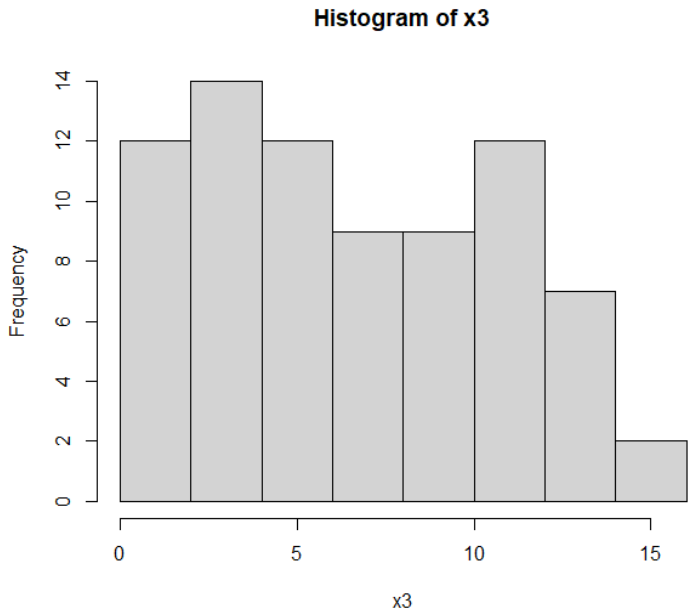
```
> var(x3)
[1] 19.56152
```

e) Обчислити var(*$y);

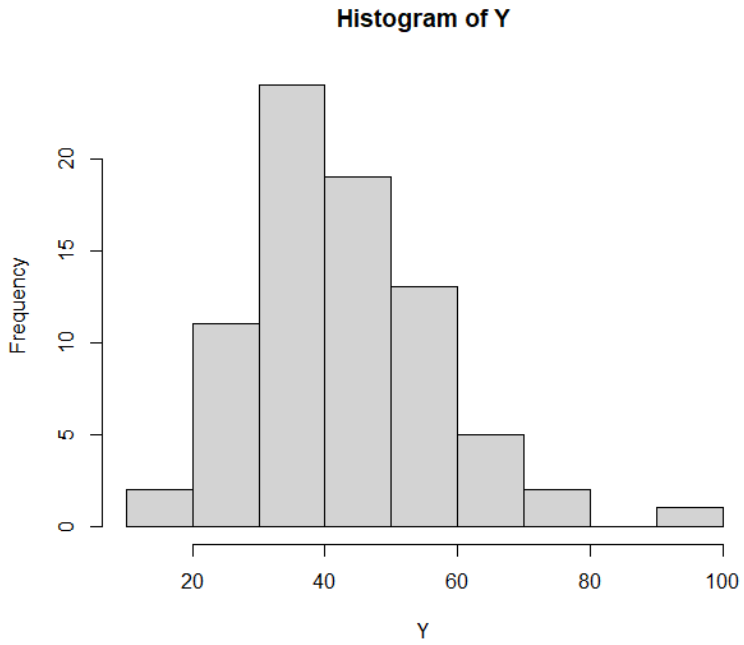
```
> var(Y)
[1] 197.3263
```

Варіація у більше, ніж варіація x, що погано

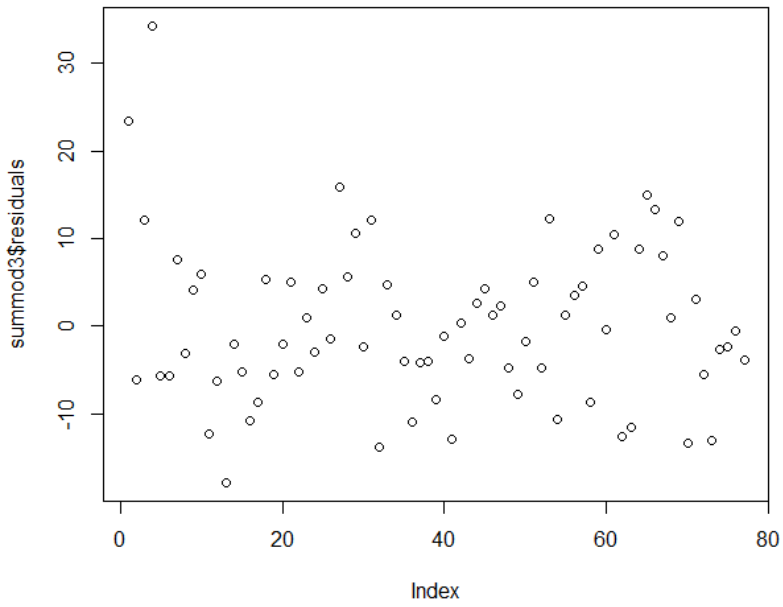
f) Побудувати hist(*$$x_i$$);



g) Побудувати hist(*$y);



h) Побудувати plot(*$residuals) зробити припущення чи відповідає $N(0; 1)$;



Схоже було б на нормальний розподіл, тобто на N(0, 1), але є декілька викидів

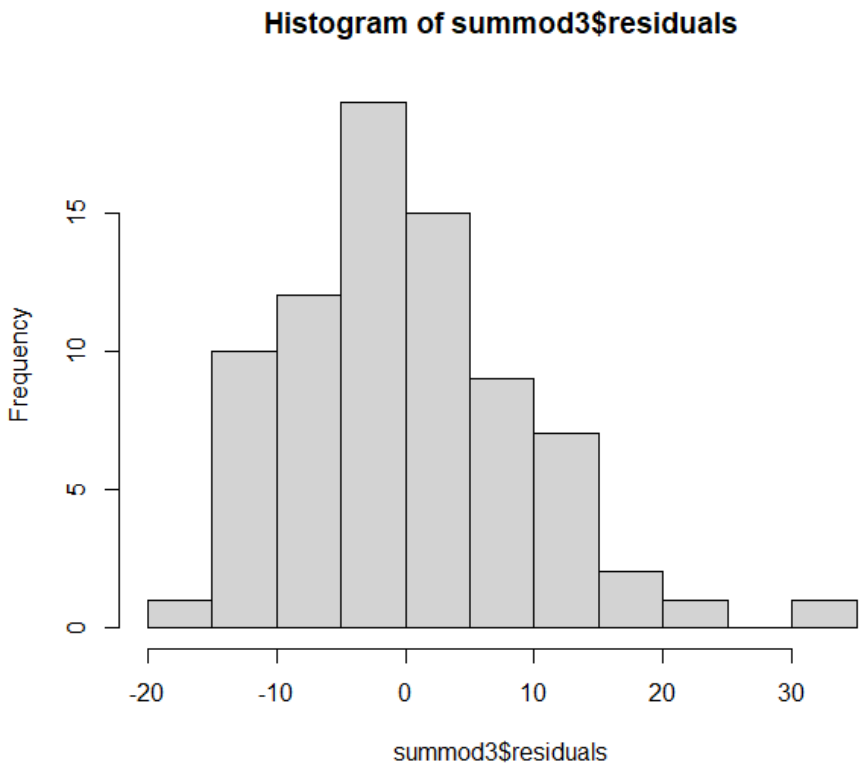
i) Перевірити mean(*$residuals);

```
> mean(summod3$residuals)
[1] -2.565588e-16
```

j) Обчислити var(*$residuals);

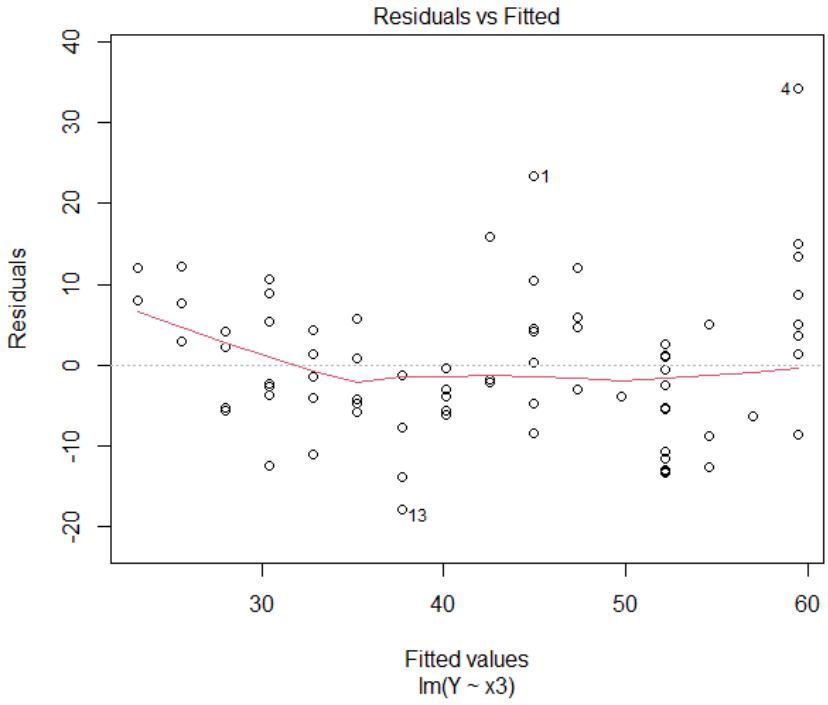
```
> var(summod3$residuals)
[1] 82.83043
```

k) Побудувати hist(*$residuals) та перевірити чи відповідає $N(0; 1)$;

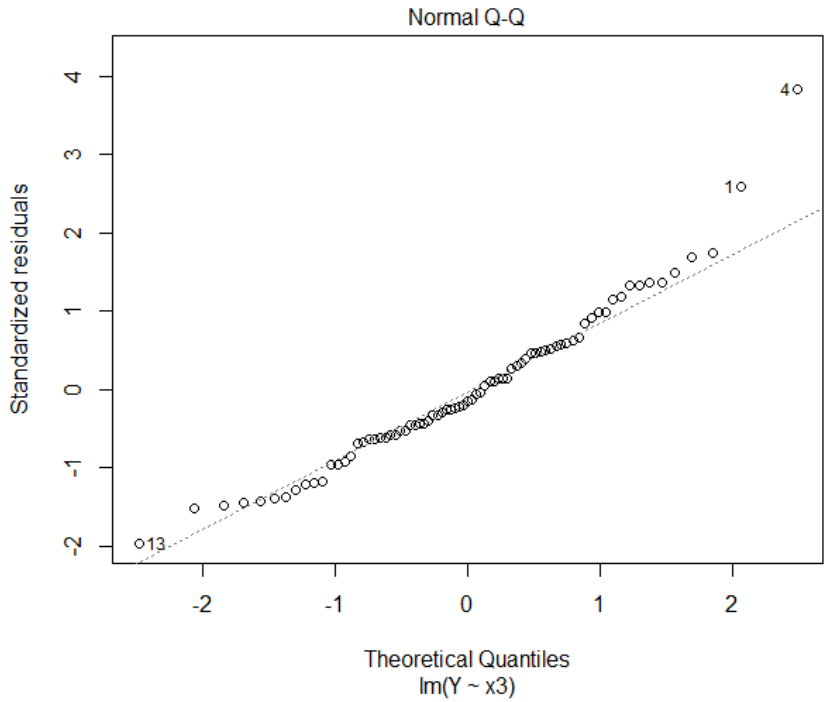


Схоже на нормальний розподіл, але на правому хвості є декілька викидів

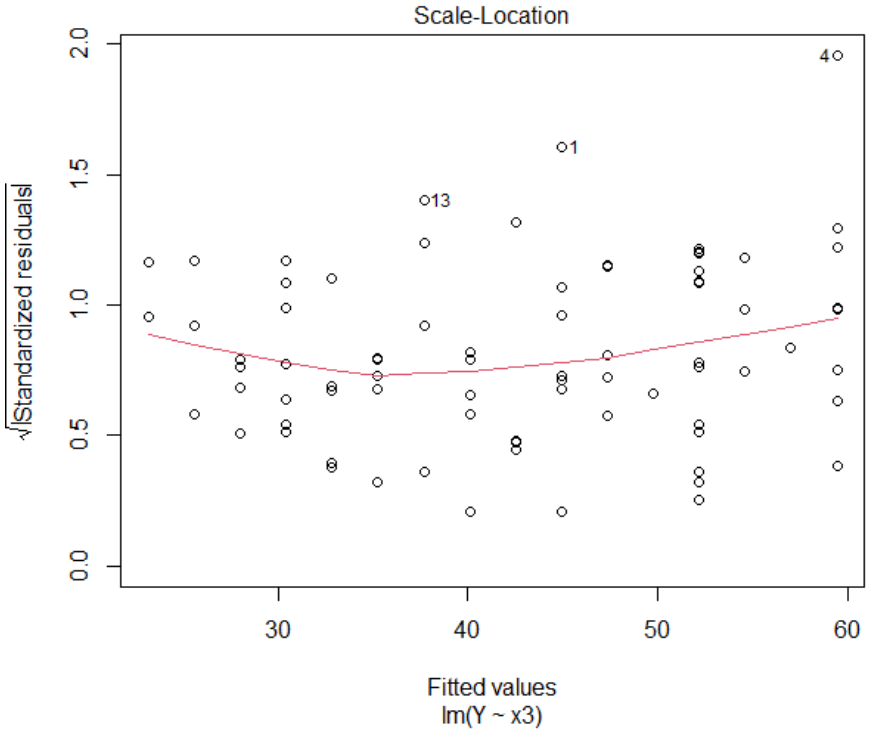
l) Виконати перевірку 4-х припущень для МНК за допомогою plot(mod_AGST, 1), plot(mod_AGST, 2), plot(mod_AGST, 3), plot(mod_AGST$residuals, type = "o").



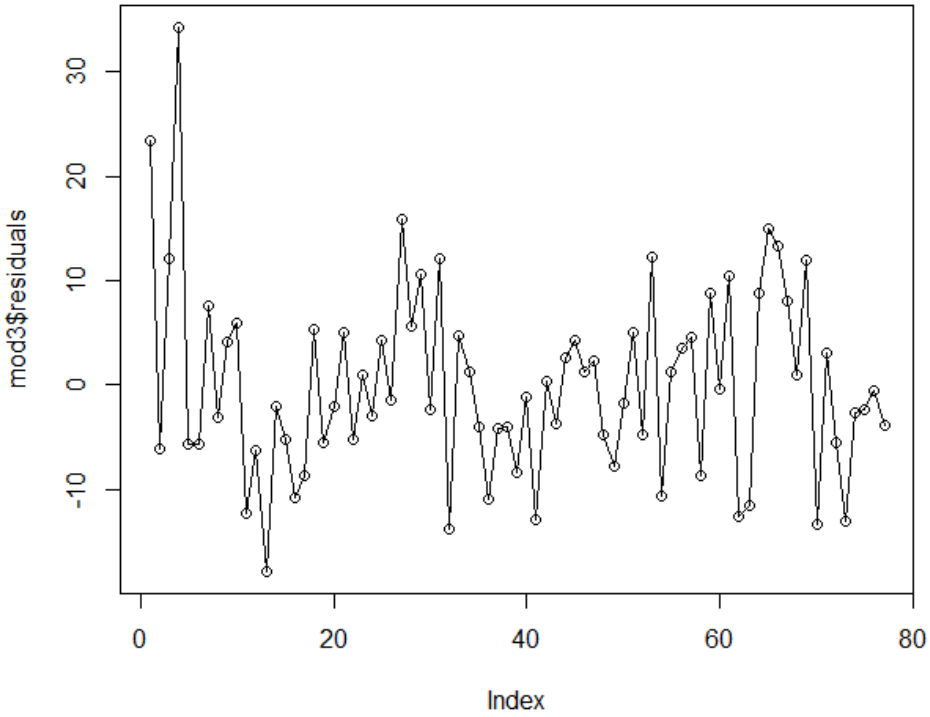
Майже вся червона лінія лежать на пунктирній лінії, тобто можна зробити висновок що лінійність дотримується



Майже усі значення лежать на прямій, окрім декількох викидів в кінці, що означає що данні розподілені нормально



З графіку схоже, що припущення про гомоскедастичність виконується



Спочатку є декілька викидів, якщо їх не було то виконувалось б припущення про незалежність похибок

B. Побудувати лінійну модель (m1) за не менше ніж 5-ма параметрами;

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5;$$

Визначити з summary() чому дорівнює RSE та порахувати вручну, а також перевірити чи вони співпадають.

```
> x4 <- data$protein
> x5 <- data$weight
> m1 <- lm(Y ~ x1+x2+x3+x4+x5)
> sumM1 <- summary(m1)
> sumM1
> sum((Y - m1$coefficients[1] - m1$coefficients[2] * x1 - m1$coefficien
ts[3] * x2- m1$coefficients[4] * x3- m1$coefficients[5] * x4- m1$coeffi
cients[6] * x5)^2)
[1] 2441.586
> sum(m1$residuals^2)
[1] 2441.586
> sqrt(sum(m1$residuals^2)/m1$df.residual)
[1] 5.864174
> sumM1$sigma
[1] 5.864174
```

RSE = 5.864174

Значення яке ми отримали з summary збіглось з значенням яке ми порахували вручну

C. Створити модель (m2) в якої на 1-н параметр менше; Визначити ступені вільності для (m1) та (m2).

```
> m2 <- lm(Y ~ x1+x2+x3+x4)
> summary(m2)

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

Residuals:
     Min       1Q   Median       3Q      Max
-13.8007  -3.8030  -0.7069   3.3371  16.5003

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 70.82795    4.68446  15.120  < 2e-16 ***
x1          -0.28956    0.04957  -5.842 1.38e-07 ***
x2          -2.75911    0.84251  -3.275  0.00163 **
x3          -1.10876    0.21432  -5.173 1.99e-06 ***
x4           5.21368    0.73465   7.097 7.32e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.209 on 72 degrees of freedom
Multiple R-squared:  0.8149,    Adjusted R-squared:  0.8046
F-statistic: 79.26 on 4 and 72 DF,  p-value: < 2.2e-16


> sumM1

Call:
lm(formula = Y ~ x1 + x2 + x3 + x4 + x5)

Residuals:
     Min       1Q   Median       3Q      Max
-12.2182  -3.7736  -0.6054   2.6367  15.5582

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.51511    5.00849  12.681  < 2e-16 ***
x1          -0.40559    0.05982  -6.780 2.97e-09 ***
x2          -1.88527    0.84374  -2.234  0.02860 *
x3          -1.29847    0.21139  -6.143 4.20e-08 ***
x4           4.17997    0.76912   5.435 7.32e-07 ***
x5          22.12101    7.09966   3.116  0.00265 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.864 on 71 degrees of freedom
Multiple R-squared:  0.8372,    Adjusted R-squared:  0.8257
F-statistic: 73.02 on 5 and 71 DF,  p-value: < 2.2e-16
```

Для m1 ступені вільності дорівнюють 71, а для m2 - 72

D. Порівняти моделі (m1) та (m2) за допомогою функцій summary() та car::compareCoefs(m1, m2) на предмет: $R2$, RSE, $SE(\beta i$). Зробити висновок, яка модель краща

```
> car::compareCoefs(m1, m2)
Calls:
1: lm(formula = Y ~ x1 + x2 + x3 + x4 + x5)
2: lm(formula = Y ~ x1 + x2 + x3 + x4)

            Model 1 Model 2
(Intercept)   63.52   70.83
SE             5.01    4.68

x1          -0.4056 -0.2896
SE           0.0598  0.0496

x2           -1.885  -2.759
SE            0.844   0.843

x3           -1.298  -1.109
SE            0.211   0.214

x4            4.180   5.214
SE            0.769   0.735

x5             22.1
SE              7.1
```

Summary:
R^2: m1 — 0.8372, m2 — 0.8149. R^2 в моделі m1 вищий, тобто кращий
Adjusted R^2: m1 — 0.8257, m2 — 0.8046. Adjusted R^2 в моделі m1 вищий, тобто кращий
RSE: m1 — 5.864, m2 — 6.209. RSE в моделі m1 менший, ніж в m2, тобто кращий

Для моделі m2 усі SE трішки менше, окрім x3

Можна зробити висновок, що модель m1 є краща, бо вона є кращою майже у всіх параметрах які ми перевіряли