Course: Computer statistics

Lecture 9: Multiple regression.

Lecturer: Oleksandr Dykhovychnyi

Лекція 9. Багатовимірна регресія.

I. Модель. Оцінки
II. Властивості оцінок x2
III. Аналіз рівняння регресії
IV. Перевірка значущості рівняння множинної регресії
V. Довірчі інтервали для коефіцієнтів регресії. Прогноз
VI. Взаємодія предикторів
VII.Використання категоріальних предикторів

Множинний регресійний аналіз є логічним розвитком парного регресійного аналізу у випадку, коли залежна змінна пов'язана з більш ніж однією незалежною змінною. Модель парної регресії дає хороший результат в тому випадку, коли впливом інших факторів на залежну змінну можна нехтувати. Приміром, якщо коефіцієнт детермінації для побудованого рівняння регресії близький до одиниці. Однак, в практичних завданнях такі ситуації є скоріше винятком, ніж правилом. Але, якщо коефіцієнт детермінації суттєво відрізняється від одиниці, потрібно шукати зв'язок з іншими незалежними змінними, тобто розглядати множинну регресію. Тому моделі множинної лінійної регресії мають досить широке розповсюдження.

I. Модель. Оцінки

Розглянемо регресійні рівняння, в яких має місце лінійний зв'язок залежної змінної y від K незалежних змінних $x_1, x_2, ..., x_K$.

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_K x_K + \varepsilon,$$

де y - залежна випадкова змінна;

$x_1, x_2, ..., x_K$ - незалежні детерміновані змінні;

$a_0, a_1, ..., a_K$ - сталі;

ε - випадкова похибка.

В реальних статистичних дослідженнях це виглядає так. Маємо n+1 вибірку спостережень $Y = (y_1, y_2, ..., y_n)$ та $X^i = (x_{i1}, x_{i2}, ..., x_{iK}), i = \overline{1,n}$. За аналогією з одновимірним випадком лінійна залежність має вигляд :

$$y_i = a_0 + a_1 x_{i1} + ... + a_1 x_{iK} + \varepsilon_i, i = \overline{1,n}.$$

Знову, за відповідною аналогією з парною регресією, шукаємо оцінки параметрів $a^*_0, a_1^*, ..., a_K^*$, такі, що мінімізують суму квадратів похибок відхилення y_i від теоретичних $y_i^* = a_0 + a_1 x_{i1} + ... + a_1 x_{iK}$. Тобто,

$$Q_\varepsilon(a_0, a_1, ..., a_K) = \sum_{i=1}^{n} \left(y_i - y_i^* \right)^2 = \sum_{i=1}^{n} \varepsilon_i^2 \to \min .$$

Введемо наступні матриці:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \; X = \begin{pmatrix} 1\, x_{11} \dots x_{1K} \\ 1\, x_{21} \dots x_{2K} \\ \dots\dots\dots \\ 1\, x_{n1} \dots x_{nK} \end{pmatrix}, A = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \dots \\ a_K \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Тоді лінійна модель множинної регресії в матричної формі набуде вигляду:

$$Y = XA + \varepsilon \; .$$

Відповідна сума квадратів відхилень набуває вигляду:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \left(Y - XA \right)^T \left(Y - XA \right) \to \min .$$

На підставі геометричних міркувань складають відповідну систему **нормальних рівнянь**

$$X^T (Y - XA) = 0 .$$

Розв'язавши яку за умов невиродженої матриці $X^T X$, отримуємо оцінки коефіцієнтів регресії:

$$A^* = \left(X^T X\right)^{-1} X^T Y \ , \ A^* = \begin{pmatrix} a^*_0 \\ a^*_1 \\ a^*_2 \\ \dots \\ a^*_K \end{pmatrix}.$$

На підставі знайдених коефіцієнтів можна визначити розраховане значення залежної змінної y.

$$Y^{**} = XA^* = X\left(X^T X\right)^{-1} X^T Y.$$

II. Властивості оцінок

Розгляньмо властивості отриманих оцінок коефіцієнтів регресії. За припущення, невиродженої матриці $X^T X$ та умови, що $\varepsilon_i, i = \overline{1,n}$ - послідовність незалежних нормально розподілених величин з нульовим середнім і дисперсією σ^2 , оцінки $a^*_0, a_1^*, ..., a_K^*$ будуть **незсуненими, конзистентними** та **ефективними** серед усіх лінійних незсунених оцінок.

III. Аналіз рівняння регресії

Як і у випадку парної регресії, аналіз якості рівняння базується на дисперсійному аналізі. Аналогично справедливо співвідношення

$$Q_t = Q_r + Q_\varepsilon,$$

$Q_t = YY^T - n\overline{Y}^2$ – загальна сума квадратів відхилень значень залежної змінної від її вибіркового середнього значення;

$Q_r = A^{*T} X^T Y - n\overline{Y}^2$ – сума квадратів відхилень розрахункових значень залежної змінної від її вибіркового середнього значення;

$Q_\varepsilon = YY^T - A^{*T} X^T Y$ -сума квадратів залишків або помилок.

Тоді **коефіцієнт детермінації** R^2 , визначається так само, як і в випадку парної регресії

$$R^2 = \frac{Q_r}{Q_t} = \frac{A^{*T} X^T Y - n\overline{Y}^2}{YY^T - n\overline{Y}^2}$$

І має аналогічну інтерпретацію: чим ближче коефіцієнт детермінації R^2 до одиниці, тим вище якість отриманого рівняння регресії і тим краще і воно відповідає емпіричним даним.

Таким чином, R^2 характеризує тісноту зв'язку набору незалежних ознак $x_1, x_2, ..., x_K$ із залежною змінною у, тобто, оцінює ступінь тісноти їх зв'язку. При цьому можна показати, що коефіцієнт детермінації у разі лінійної моделі з точністю до знаку дорівнює вибірковому коефіцієнту кореляції між спостережуваними величинами Y і розрахунковими Y^{**} , тобто,

$$|R| = \rho_{YY^{**}}$$

Величину $|R| = \sqrt{R^2}$ у разі множинної регресійної моделі називають ще й **коефіцієнтом множинної кореляції**.

Недоліком коефіцієнта детермінації R^2, що обмежує його застосування, є те, що при додаванні нових незалежних змінних його значення завжди зростає, хоча це не означає поліпшення якості моделі як такої. Щоб уникнути цієї ситуації пропонується використовувати **скорегований коефіцієнт детермінації**.

$$R_1^2 = 1 - \frac{(n-K)(1-R^2)}{(n-k-1)} = \frac{(n-1)\varepsilon^T \varepsilon}{(n-k-1)Y^T Y} \quad .$$

На відміну від R^2 при впровадженні в модель нових незалежних змінних скоригований коефіцієнт R_1^2 може зменшуватися в тому випадку, коли ці змінні істотно не впливають на залежну змінну. Використання R_1^2 для порівняння регресій є більш коректним.

IV. Перевірка значущості рівняння множинної регресії

В цілому значущість рівняння регресії зазвичай розуміється як існування такої залежності, в якій на величину y впливає хоча б одна незалежна змінна x_i. І, навпаки, рівняння регресії вважається незначущим, якщо всі змінні x_i не зв'язані з y. В цьому випадку вся мінливість величини y пояснюється випадковою складовою ε, і, як було зазначено вище, коефіцієнт детермінації дорівнюватиме нулю: $R^2 = 0$.

Статистика критерію для перевірки значущості рівняння регресії може бути виражена через коефіцієнт множинної детермінації R^2. Якщо модель узгоджена з вибірковими даними, то статистика

$$\frac{Q_r(n-K-1)}{Q_\varepsilon K} = \frac{R^2(n-K-1)}{(1-R^2)K}$$

має розподіл Фішера $F(K, n-K-1)$.

V. Довірчі інтервали для коефіцієнтів регресії. Прогноз

У припущенні про нормальність і незалежність розподілу випадкових компонент $\varepsilon_i \sim N(0,\sigma^2)$ вибіркові оцінки коефіцієнтів рівняння регресії $a_i^*, i=\overline{0,K}$ матимуть нормальні розподіли, а їх нормовані відхилення від теоретичних значень - розподіл Стьюдента з n - К – 1 степенів свободи:

$$\frac{a^*_i - a_i}{\sigma_{a_i}} \sim t_{(n-K-1)}, i=\overline{1,n},$$

де $\sigma^2_{a_i} = S^2(X^TX)^{-1}$, $S^2 = \dfrac{\varepsilon^T\varepsilon}{n-K-1}$.

Це дозволяє будувати довірчі інтервали для коефіцієнтів регресії рівня довіри α:

$$a_i \in \left(a^*_i - t_{1-\frac{\alpha}{2},(n-K-1)}\sigma_{a_i}, a^*_i - t_{1-\frac{\alpha}{2},(n-K-1)}\sigma_{a_i} \right), i=\overline{1,n}.$$

На підставі цих довірчих інтервалів можна перевіряти гіпотези про рівність відповідних коефіцієнтів регресії нулеві $H_i = \{a_i = 0\}$ шляхом перевірки, чи накриває відповідний довірчий інтервал нульове значення, чи ні.

Перевірка гіпотези про значущість коефіцієнта a_i зводиться до

порівняння статистики $\left|\frac{a_i^*}{\sigma_{a_i}}\right|$ з відповідним квантілем розподілу Стьюдента з n-K-1 степенем свободи.

Маючи побудовану модель регресії, ми можемо побудувати прогноз. Нехай вектор $X_0 = (1, x_1^0, x_2^0, ..., x_K^0)$ представляє значення незалежних змінних, для яких потрібно визначити значення залежної змінної y_0^* .

$$y_0^* = a_0^* + a_1^* x_1^0 + ... + a_K^* x_K^0 .$$

Водночас отримуємо й довірчі інтеграли для прогнозу:

$$y_0 \in \left(y^*_0 - t_{1-\frac{\alpha}{2},(n-K-1)} S_{y_0}, y^*_0 + t_{1-\frac{\alpha}{2},(n-2)} S_{y_0} \right),$$

де $S_{Y_0}^2 = S^2(X_0^T (X^T X)^{-1} X_0)$

Множинний регресійний аналіз в **R** реалізує функція

lm(formula = …)

Приклади:

```
> #Зчитування дані з файла
> dat <- read.table(file.choose(), head=TRUE)
> attach(dat);
> dat
      y   x1    x2
1  12.2  4795  69
2   7.6   6962  82
3  10.4  6571  87
4   9.9   4249  92
5  15.7  9540  23
6  14.0  3488  31
7  12.7  4888  55
8  10.5  6237  81
9  15.1  2997  65
10 10.6 2990  98
11 15.2 1748  100
12 17.2 2128  69

> # візуальна перевірка залежностей
> pairs(dat,col=3)
```

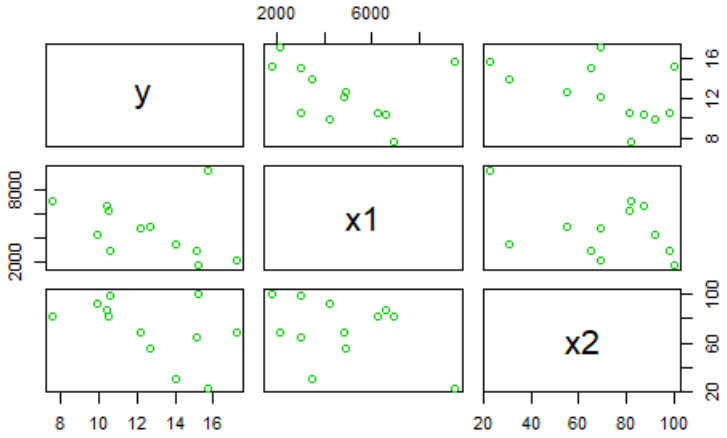

Рис.1. (Рисунок згенеровано **RStudio**).

x2

```
> # перевірка корельованості
> cor(x1,x2) #між x1,x2
[1] -0.3920887
> cor(y,x1)   #між x1,y
[1] -0.3384028
> cor(y,x2) #між y,x2
[1] -0.4810236
> cor(dat,method="pearson") #кореляційна матриця
           y          x1          x2
y  1.0000000 -0.3384028 -0.4810236
x1 -0.3384028  1.0000000 -0.3920887
x2 -0.4810236 -0.3920887  1.0000000

> cor.test(y,x1,method="pearson") # перевірка коефіцієнта корелляції
                                  # на нуль

        Pearson's product-moment correlation

data:  y and x1
t = -1.1372, df = 10, p-value = 0.282
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7639397  0.2922583
sample estimates:
      cor
-0.3384028
> # побудова регресійної моделі
> fit <- lm(y ~ x1 + x2) # модель задається формулою
> # резюме
```

> summary(fit)

Call:
lm(formula = y ~ x1 + x2)

Residuals:

Min	1Q	Median	3Q	Max
-2.9491	-1.1543	-0.2731	1.0857	2.8351

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	22.3218043	2.9415849	7.588	3.37e-05	***
x1	-0.0007869	0.0003039	-2.590	0.0292	*
x2	-0.0847747	0.0281108	-3.016	0.0146	*

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 2.115 on 9 degrees of freedom
Multiple R-squared: 0.5596, Adjusted R-squared: 0.4617
F-statistic: 5.717 on 2 and 9 DF, p-value: 0.02497

Модель має вигляд: $y = 22.32 - 0.008x_1 - 0.08x_2$. Всі коефіцієнти можна вважа ти значущими. Діаграми розсіювання зображені на рисунку 1.

Перевіримо залишки на нормальність (рис.2)

Приклади:
> # перевірка залишків моделі на нормальність
> qqnorm(fit[[2]], pch=4); qqline(as.vector(fit[[2]]))

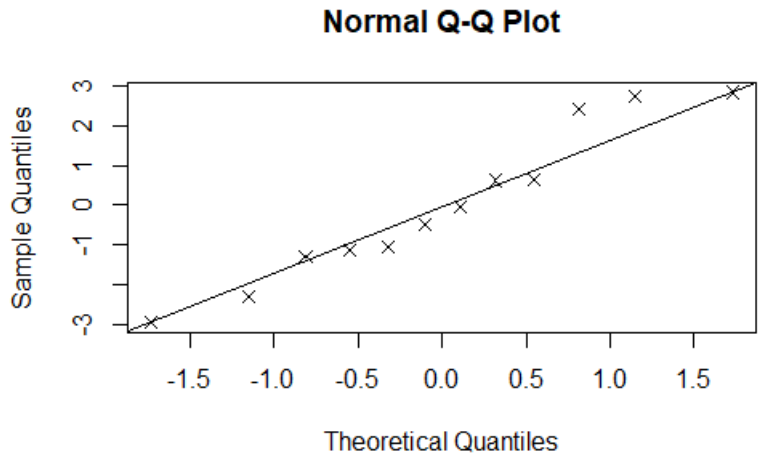

Рис. 2 (Рисунок згенеровано **RStudio**).

Зобразимо графік залишків (рис. 3).

Приклади:

```
> windows();
> par(mfrow=c(3,1))
> plot(x1, fit[[2]], pch=4); abline(h=0, lty=1)
> plot(x2, fit[[2]], pch=4); abline(h=0, lty=1)
```

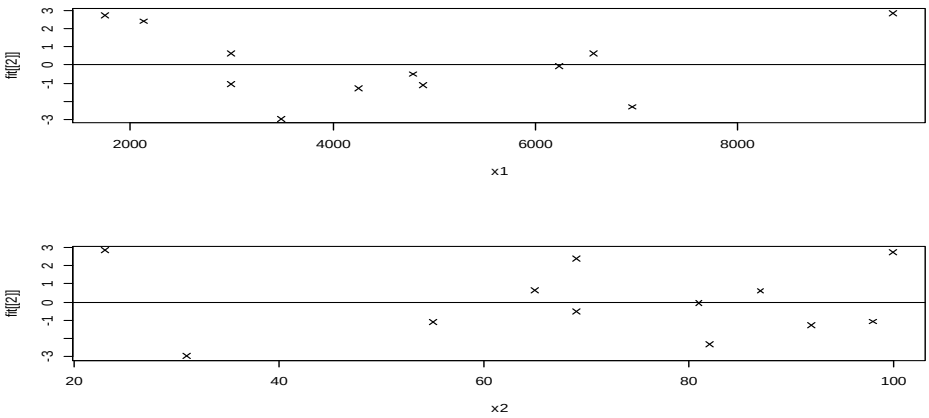

Рис. 3 (Рисунок згенеровано **RStudio**).

Зробимо прогноз за кожною змінною і побудуємо графіки (рис. 4,5).

Приклади:

```
> # розраховуємо прогноз
> x1i <- seq(0.9*min(x1), 1.1*max(x1), len=100)
> x2i <- seq(0.9*min(x2), 1.1*max(x2), len=100)
> pre <- predict(fit,data.frame(x1=x1i,x2=x2i),interval="confidence")

> windows();# зображуємо прогноз
> plot(x1, y, pch=3)
            > matplot(x1i, pre, type="l", lty=c(1,2,2), add=TRUE)
```

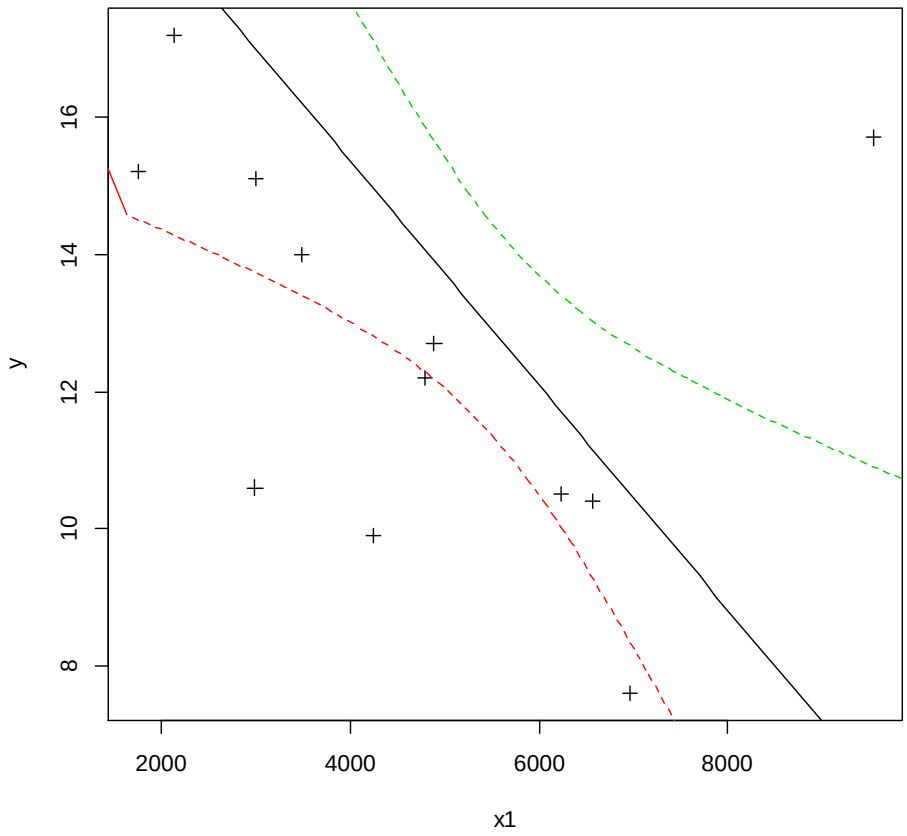

Рис. 4 (прогноз по змінній x_1)
(Рисунок згенеровано **RStudio**).

Приклади:

```
> windows(); plot(x2, y, pch=3)
> matplot(x2i, pre, type="l", lty=c(1,2,2), add=TRUE)
```

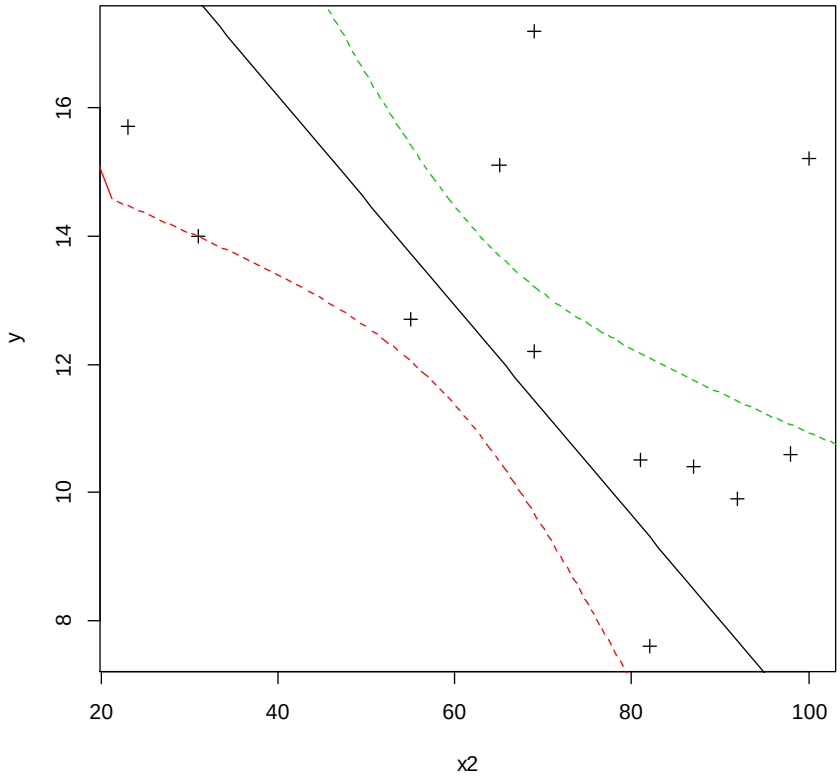

Рис. 5 (прогноз по змінній x_2)
(Рисунок згенеровано **RStudio**).

VI. Взаємодія предикторів

У розглянутих регресійних моделях всі змінні вважались незалежним и. Але у ряді випадків виникають ефекти, пов'язані із взаємодією змінних

Розглянемо вбудований набір **swiss,** який містить дані про народжуваність у кантонах Швейцарії.

Приклади:

```
> swiss
             Fertility Agriculture Examination Education Catholic
Courtelary        80.2        17.0          15        12     9.96
Delemont          83.1        45.1           6         9    84.84
Franches-Mnt      92.5        39.7           5         5    93.40
Moutier           85.8        36.5          12         7    33.77
.....
```

Народжуваність (Fertility) залежить від пятьох змінних змінних: Agricul ture(рівень сільського господарства), Examination(фізичної підготовки), Education(освіти), Catholic (Релігії), Infant.Mortality (дитячої смертності).

Приклади:

```
> fit <- lm(Fertility ~ Examination + Catholic, data = swiss)
> summary(fit)
```

Збудуємо лінійну залежність між народжуваністю та фізичною підготовк ою та належністю до католіцизму:

```
Call:
lm(formula = Fertility ~ Examination + Catholic, data = swiss)

Residuals:
    Min      1Q  Median      3Q     Max
-26.2643  -5.6510  -0.0017   7.7268  17.7103

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 83.03566    4.97730  16.683  < 2e-16 ***
Examination -0.88619    0.21736  -4.077 0.000188 ***
Catholic     0.04179    0.04158   1.005 0.320322
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.641 on 44 degrees of freedom
Multiple R-squared:  0.4302,    Adjusted R-squared:  0.4043
F-statistic: 16.61 on 2 and 44 DF,  p-value: 4.218e-06
```

Як бачимо, фізична підготовка є значущою ї від’ємним коефіцієнтом, а вп лив католіцизму є незначущим. Модель має вигляд:

$$F = 83.03566 - 0.88619 * E + 0.04179 * C$$

Також можна визначити вплив взаємодії двох факторів E і C. Для цього в формулі задаємо **Fertility ~ Examination*Catholic** і викликаємо:

Приклади:

```
> fit2 <- lm(Fertility ~ Examination*Catholic, data = swiss)
> summary(fit2)

Call:
lm(formula = Fertility ~ Examination * Catholic, data = swiss)
```

Residuals:

Min	1Q	Median	3Q	Max
-25.5446	-5.3640	0.5461	7.5383	18.5540

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	80.957567	6.471732	12.509	6.37e-16	***
Examination	-0.765480	0.323031	-2.370	0.0224	*
Catholic	0.083823	0.092648	0.905	0.3706	
Examination:Catholic	-0.003337	0.006559	-0.509	0.6135	

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.723 on 43 degrees of freedom
Multiple R-squared: 0.4337, Adjusted R-squared: 0.3941
F-statistic: 10.98 on 3 and 43 DF, p-value: 1.77e-05

Ця модель математично має вигляд:

$$F = 80.957567 - 0.765480 * E + 0.083823 \ *C - 0.003337 * E * C$$

Взагалі кажучи, така модель не є лінійною, але її також умовно вважаю ть такою. Зауважимо, що цей самий результат можна було отримати кома ндою:

Приклади:

```
fit <- lm(Fertility ~ Examination + Catholic+Examination*Catholic, data = swiss)
```

VI. Використання категоріальних предикторів

У розглянутих регресійних моделях всі незалежні змінні були числов ими(кількісними). Однак на практиці зустрічаються змінні іншого типу - **категоріальні**. Приміром: чоловік/жінка; дитина/дорослий/пенсіонер; кат олик/некатолик.

Якщо ми хочемо ввести у модель такий предиктор, то слід ввести нову зм інну такого виду:

$$x_i = \begin{cases} 1, \textit{якщо } i - \textit{та людина католик}; \\ 0, \textit{якщо } i - \textit{та людина не католик}. \end{cases}$$

Тоді, впровадження такої змінної приводить до моделі залежності

народжуваності від релігійності:

$$y_i = a_0 + a_1 x_{i1} + \varepsilon_i = \begin{cases} a_0 + a_1 + \varepsilon_i, \text{ якщо } i - \text{та людина католик}; \\ a_0, \text{ якщо } i - \text{та людина не католик}. \end{cases}$$

Подивимось, як це працює на практиці. Збудуємо гістограму релігійності (рис.6).

Приклади:

```
> hist(swiss$Catholic, col = 'red')
```

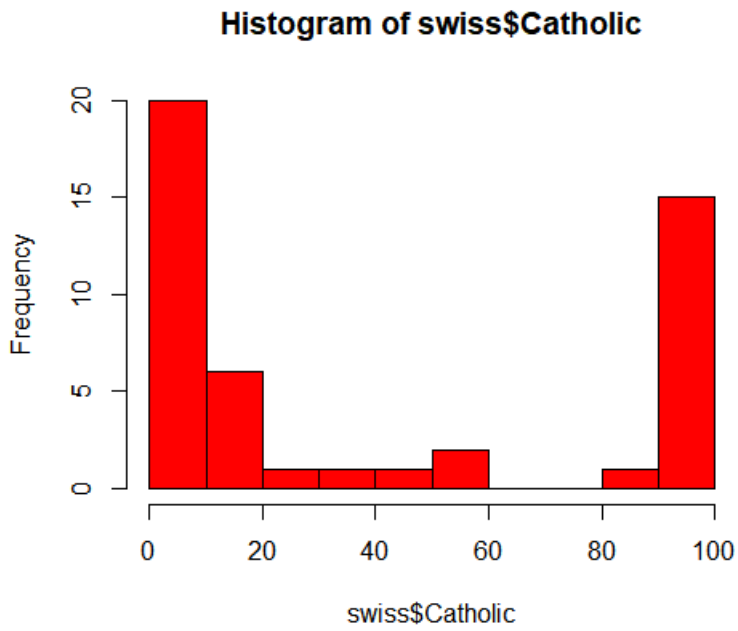

Рис. 6 (Рисунок згенеровано **RStudio**).

Як бачимо, вона має два явних піка. Тобто, є кантони, де живуть католики, а є такі, де не живуть. Створимо категоризовану змінну та внесемо її у фрейм як фактор. Цей фактор набуває двох значень:1 і 2.

Приклади:

```
> swiss$religious <- ifelse(swiss$Catholic > 60, 1,)
> swiss$religious <- as.factor(swiss$religious)
```

Тоді нова регресійна модель залежності народжуваності від фізичної підг отовки та релігійності має вигляд:

Приклади:

```
> fit3 <- lm(Fertility ~ Examination + religious, data = swiss)
> summary(fit3)
```

```
Call:
lm(formula = Fertility ~ Examination + religious, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-22.9026  -4.8974   0.1926   7.1239  15.4542

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  78.5753     4.7701  16.472   <2e-16 ***
Examination  -0.6858     0.2222  -3.086   0.0035 **
religious     8.4469     3.7016   2.282   0.0274 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.221 on 44 degrees of freedom
Multiple R-squared:  0.4788,   Adjusted R-squared:  0.4552
F-statistic: 20.21 on 2 and 44 DF,  p-value: 5.934e-07
```

Цю модель можна розшифрувати наступним чином:
Середнє значення народжуваності (Intercept) для кантонів з R=0 дорівнює 78.5753
Народжуваність в цих кантонах залежить від фізичної підготовки (Examination) лінійно з коефіцієнтом -0.6858
а стрибок народжуваності для кантонів з R=1 дорівнює 8.4469

$$F = 78.5753 - 0.6858 * E + 8.4469 * R$$

Якщо врахувати взаємодію народжуваності та релігійності, то отримаємо

Приклади:

```
> fit4 <- lm(Fertility ~ Examination*religious, data = swiss)
> summary(fit4)
```

```
Call:
lm(formula = Fertility ~ Examination * religious, data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-23.6289  -4.2417   0.0795   6.4508  14.0243

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          82.1160     5.0736  16.185  < 2e-16 ***
Examination          -0.8617     0.2389  -3.607 0.000801 ***
religious            -2.9615     7.4096  -0.400 0.691366
Examination:religious  1.0096     0.5723   1.764 0.084839 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.007 on 43 degrees of freedom
Multiple R-squared:  0.514,    Adjusted R-squared:  0.4801
F-statistic: 15.16 on 3 and 43 DF,  p-value: 7.128e-07
```

Цю модель можна розшифрувати наступним чином:

Середнє значення народжуваності(Intercept) для кантонів з R=0 - 82.1160
Народжуваність в цих кантонах залежить від фізичної підготовки (Examination) лінійно з коефіцієнтом -0.8617
а стрибок народжуваності для кантонів з R=1 дорівнює -2.9615(але він незначущий)
Фактор впливу фізичної підготовки в кантонах з R=1 високий й додатний 1.0096.

$$F = 82.1160 - 0.8617 * E - 2.9615 * R + 1.0096 * E * R$$

x2