Course: Computer statistics

Lecture 8: Regression analysis.

Lecturer: Oleksandr Dykhovychnyi

Лекція 8. Регресійний аналіз.

x2

I. Постановка задачі

II. Оцінка параметрів рівняння регресії

III.Оцінка якості вибіркового рівняння регресії

IV.Перевірка значущості рівняння регресії

V. Точковий й інтервальний прогнози

I. Постановка задачі

Регресійний аналіз досліджує і оцінює зв'язок між певною залежною змінною і незалежними змінними. Залежну змінну іноді ще називають **результативною ознакою**, а залежні змінні - **предикторами**, **регресорам**и або **факторами**.

Позначимо залежну змінну а y ,а незалежні $x_1, x_2, ..., x_K$.

При $K = 1$ є тільки одна незалежна змінна х і регресія називається **парною**. При $K > 1$ регресію називають **множинною.** Найпростіша парна регресійна модель має вигляд:

$$y = a_0 + a_1 x + \varepsilon$$

де y - залежна випадкова змінна;

x - незалежна детермінована змінна;

a_0, a_1 - сталі;

ε - випадкова похибка.

Вважатимемо, що між змінними x та y існує лінійна залежність $y = a_0 + a_1 x$.

Завдання регресійного аналізу полягає в отриманні оцінок коефіцієнтів a_0, a_1 за результатами спостережень за змінними x та y.

Похибка регресії вважається нормально розподіленою величиною з нульовим середнім і дисперсією σ^2 і, як правило, пов'язана з похибками вимірювань, або іншими випадковими факторами.

II. Оцінка параметрів рівняння регресії

Нехай є n спостережень за змінними x та y у вигляді вибірок $X = (x_1, x_2, ..., x_n)$ та $Y = (y_1, y_2, ..., y_n)$, тоді рівняння регресії можна переписати у вигляді:

x2

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

Будемо вважати, що $\varepsilon_i, i = \overline{1,n}$ - послідовність незалежних нормально розподілених випадкових величин з нульовим середнім і дисперсією σ^2 .

Основним методом отримання оцінок є метод найменших квадратів, який передбачає отримання вибіркових оцінок параметрів a^*_0, a_1^*, таких, що мінімізують суму квадратів похибок відхилення y_i від теоретичних $y_i^* = a_0 + a_1 x_i$:

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

Для мінімізації функції прирівнюємо до нуля її частинні похідні $\frac{\partial Q_\varepsilon(a_0, a_1)}{\partial a_0}$ та $\frac{\partial Q_\varepsilon(a_0, a_1)}{\partial a_1}$, одержуємо систему **нормальних рівнянь:**

$$\begin{cases} n a_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x^2{}_i = \sum_{i=1}^{n} x_i y_i \end{cases}.$$

Розв'язуючи яку, знаходимо оцінки (**вибіркові коефіцієнти регресії**):

$$a_0^* = \bar{y} - a_1^* \bar{x} ,$$

$$a_1^* = \frac{Cov_{xy}}{S_X^2},$$

де

$\bar{x}, \bar{y}$ — вибіркові середні значення відповідних змінних,

$S_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ — вибіркова дисперсія;

$Cov_{xy} = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n}$ — вибіркова коваріація;

n — обсяг вибірки.

III. Оцінка якості вибіркового рівняння регресії

Введемо наступні характеристики:

$$Q_t = \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 ,$$

$$Q_r = \sum_{i=1}^{n}\left(y^{**}_i - \bar{y}\right)^2 ,$$

$$Q_\varepsilon = \sum_{i=1}^{n}\left(y^{**}_i - y_i\right)^2 ,$$

$y_i^{**} = a_0^* + a^*_1 x_i, i = \overline{1,n}$ - розраховані значення y,

де Q_t – загальна сума квадратів відхилень значень залежної змінної від її вибіркового середнього значення;

Q_r – сума квадратів відхилень розрахункових значень залежної змінної від її вибіркового середнього значення;

Q_ε-сума квадратів відхилень y_i^{**} від лінії регресії, яку зазвичай називають сумою квадратів залишків або помилок.

Величину $\sqrt{\frac{Q_\varepsilon}{n-2}}$ називають **середньоквадратичною похибкою** або помилкою рівняння регресії**.**

Між наведеними вище сумами квадратів існує зв'язок: $Q_t = Q_r + Q_\varepsilon$, який і дозволяє характеризувати якість побудованого рівняння регресії. Рівняння регресії вважається тим краще, чим більше сума квадратів Q_r, у порівнянні з сумою квадратів залишків Q_ε.

Для строгої формалізувати цього твердження використовується **коефіцієнт детермінації**:

$$R^2 = \frac{Q_r}{Q_t} = \frac{\sum_{i=1}^{n}\left(y^{**}{}_i - \overline{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \overline{y}\right)^2}.$$

Очевидно, що $R^2 \in [0,1]$. Причому, чим ближче коефіцієнт детермінації R^2 до одиниці, тим вище якість отриманого рівняння регресії. Максимальна значення коефіцієнта детермінації $R^2 = 1$ досягається в тому випадку, коли всі залишки $\varepsilon_i = 0, i = \overline{1,n}$, а рівняння прямої регресії проходить точно через всі точки $(x_i, y_i), i = \overline{1,n}$.

IV. Перевірка значущості рівняння регресії

Розгляньмо властивості отриманих оцінок коефіцієнтів регресії. За припущення, що $\varepsilon_i, i=\overline{1,n}$ - послідовність незалежних нормально розподілених величин з нульовим середнім і дисперсією σ^2 , оцінки a^*_0, a^*_1 будуть **незсуненими, конзистентними** та **ефективними**.

Причому вони мають нормальний розподіл, а саме:

$$a^*_0 \sim N(a_0, \sigma^2_{a_0}),$$

$$a^*_1 \sim N(a_1, \sigma^2_{a_1}),$$

де

$$\sigma^2_{a_0} = \frac{\sigma\sqrt{\sum_{i=1}^{n} x_i^2}^{x2}}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}, \quad \sigma^2_{a_1} = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}.$$

На підставі цього маємо:

$$\frac{a^*_0 - a_0}{\sigma_{a_0}} \sim t_{(n-2)}, \frac{a^*_1 - a_1}{\sigma_{a_1}} \sim t_{(n-2)},$$

тобто, мають розподіл Стьюдента з $n-2$ степенями свободи. Враховуючи це, можна отримати довірчі інтервали і перевірити гіпотезу про рівність коефіцієнтів нулеві. Отже, отримуємо довірчі інтервали для рівня довіри α :

$$a_0 \in \left(a^*_0 - t_{1-\frac{\alpha}{2},(n-2)}\sigma_{a_0}, a^*_0 - t_{1-\frac{\alpha}{2},(n-2)}\sigma_{a_0} \right),$$

$$a_1 \in \left(a^*_1 - t_{1-\frac{\alpha}{2},(n-2)}\sigma_{a_1}, a^*_1 - t_{1-\frac{\alpha}{2},(n-2)}\sigma_{a_1} \right).$$

Перевірка значущості моделі полягає у перевірці гіпотез $H = \{a_0 = 0\}$ або $H = \{a_1 = 0\}$ і здійснюється шляхом перевірки, чи накриває відповідний довірчий інтервал нульове значення, чи ні?

Перевірка згоди моделі лінійної регресії з даними спостережень, показує наскільки саме лінійна модель відповідає даним. Її реалізують за

наступним критерієм. Якщо модель узгоджена з вибірковими даними, то статистика

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

має розподіл Фішера $F(1,n-2)$.

V. Точковий й інтервальний прогнози

Рівняння регресії, одержане в результаті аналізу емпіричних даних, може бути використано для прогнозування значень залежної змінної y_0 у при заданих значеннях незалежної змінної x_0 шляхом підстановки цих значень у рівняння: $y^*_0 = a^*_0 + a^*_1 x_0$. При цьому $y_0 \sim N(y_0, k_y \sigma_y^2)$. Де

$$k_y = \sqrt{\frac{1}{n} + \frac{\left(x-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}},$$

а в якості оцінки σ_y^2 використовується статистика:

$$S_y^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - y_i^{**}\right)^2 .$$

Водночас отримуємо й довірчі інтеграли для прогнозу:

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

Регресійний аналіз в **R** реалізує функція

lm(formula = …)

Розглянемо приклад, у якому досліджується залежність шкідливих викидів підприємства від інвестицій. Спочатку відобразимо їх на рисунку 1.

Приклади:

 # формуємо дані

```
 # інвестиції
> x<-c(2.36, 2.67, 2.98, 3.30, 3.61, 3.93, 4.24, 4.56 , 4.87, 5.18, 5.50)
 #викиди
> y<-c(1.12, 0.46, 0.19, -0.27, -0.85, -0.79, -1.17, -1.88, -1.62, -1.25 , -1.04)
> x
  [1] 2.36 2.67 2.98 3.30 3.61 3.93 4.24 4.56 4.87 5.18 5.50
> y
  [1]  1.12  0.46  0.19 -0.27 -0.85 -0.79 -1.17 -1.88 -1.62 -1.25 -1.04
> plot(x,y,col="red",xlab="інвестиції", ylab="викиди") #графік
```

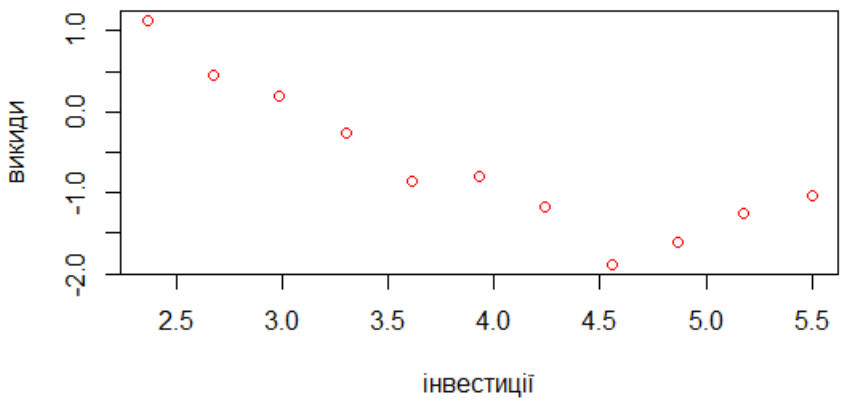

Рис.1 (Рисунок згенеровано **RStudio**).

```
> cor(x,y) # перевіряємо наявність лінійної залежності
[1] -0.8697513
```

Як бачимо, кореляція достатньо сильна , але від'ємна.

Приклади:

```
> fit <- lm(y ~ x) # будуємо модель
> summary(fit)
```

Отримуємо резюме:

```
Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
```

-0.7473 -0.2662 -0.1076 0.2487 0.8165

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	**2.3786**	**0.5900**	**4.032**	**0.002965**	******
x	**-0.7700**	**0.1456**	**- 5.287**	**0.000502**	*******

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

Residual standard error: 0.4799 on 9 degrees of freedom
Multiple R-squared: 0.7565, Adjusted R-squared: 0.7294
F-statistic: 27.96 on 1 and 9 DF, p-value: 0.0005022

Residuals - містить дані про похибки ε_i .
Estimate - коефіцієнти регресії a^*_0, a_1^*, отже , модель має вигляд:

$$y = \mathbf{2.3786} - \mathbf{0.7700}x \ .$$

Multiple R-squared – коефіцієнт детермінації **= 0.7565.**
Гіпотезу про рівність рівність нулеві коєфіцієнтів відхиляємо, оскільки для гіпотези $a_0 = 0$ **P-value=0.002965** і для гіпотези $a_1 = 0$ **P-value=0.000 502.**
Модель є значущою, оскільки для критерія Фішера **P-value=0.0005022.**

Побудуємо графік функції регресії разом з прогнозом і відповідними довірчими інтервалами (рисунок 2).

Приклади:

```
# будуємо прогноз разом з довірчими інтервалами
> xin <- seq(0.9*min(x), 1.1*max(x), length=100)
> pre <- predict(fit, data.frame(x=xin), interval="confidence")
> plot(x,y, pch=3); points(mean(x), mean(y))
> matplot(xin, pre, type="l", lty=c(1,2,2), add=TRUE)
```

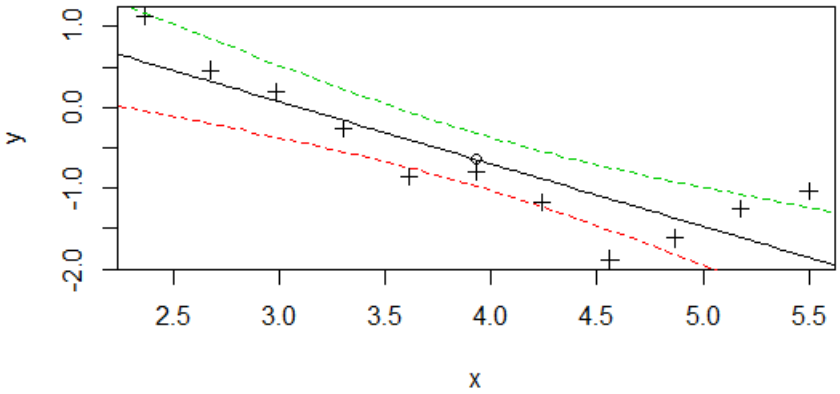

Рис. 2 (Рисунок згенеровано **RStudio**).

Перевіримо похибки. Побудуємо графік(рис.3). На ньому видно, що за лишки коливаються навколо нульового рівня.

Приклади:

```
> plot(x, fit[[2]], pch=4)
> abline(h=0)
```

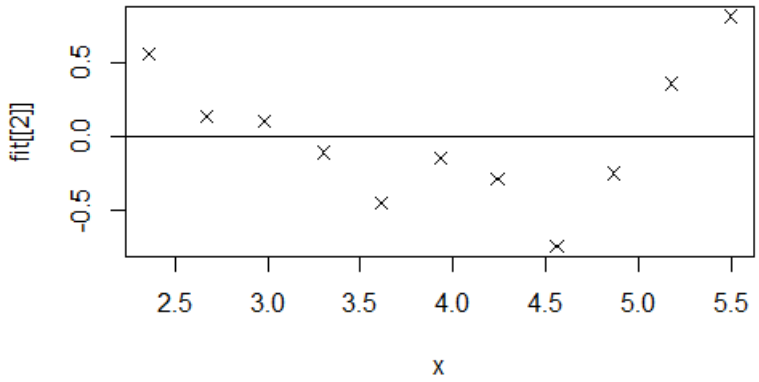

Рис.3 (Рисунок згенеровано **RStudio**).

Перевіримо залишки на нормальність. Як видно з рисунку 4 вони «схожі» на нормальні.

Приклади:

```
> qqnorm(fit[[2]], pch=4)
> qqline(as.vector(fit[[2]]))
```

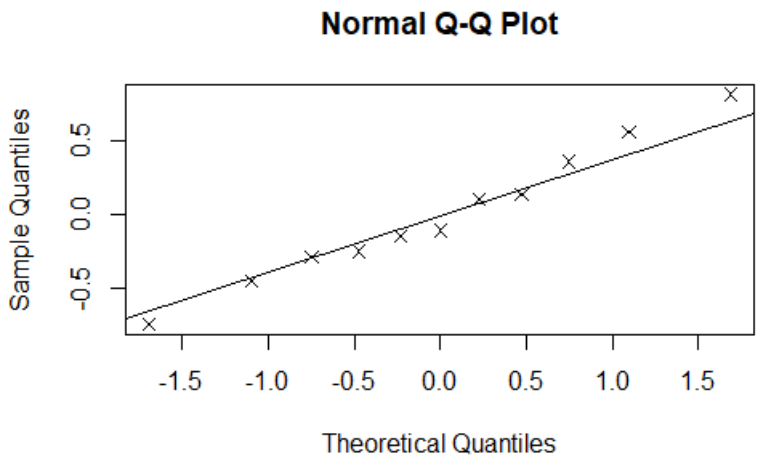

Рис.4 (Рисунок згенеровано **RStudio**).