Лабораторна роботи №3 дисципліни "регресійний аналіз" Варіант №4

Горбунова Даніела Денисовича
4 курс бакалаврату
група "комп'ютерна статистика"

23 жовтня 2020 р.

1 Вступ.

У даній роботі наведені результати, отримані під час виконання третьої самостійної роботи. Досліджено залежність між заданими об'єктами, було підігнано регресійну модель.

2 Деякі відомості про дані.

Маємо вибірку $\Xi = (\xi_1, \ldots, \xi_n)$ з $n = 150$ спостережень. Кожне спостереження має дві характеристики $\xi_j = (X_j, Y_j)$. Надалі зауважимо, що в рамках цієї роботи $X = (X_1, \ldots, X_n)$ – регресор, а $Y = (Y_1, \ldots, Y_n)$ – відгук. Якщо зобразити точки на площині, можна побачити нелінійну закономірність між характеристиками, яку можна вгадати. Бачимо, що точки вишукуються по кривій, яка нагадує графік деякої степеневої функції.

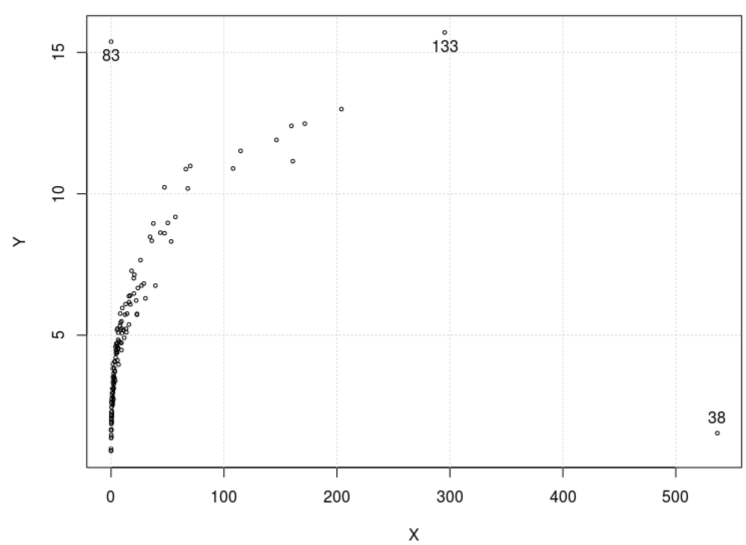

Рис. 1: Діаграма розсіювання спостережень вибірки Ξ.

З рисунку видно, що три спостереження сильно віддалені від загальної групи. За допомогою функції identify знайшли номери цих спостережень: 38, 83, 133. Чи справді вони є викидами? Якщо прологарифмувати X та Y (маємо на увазі, що береться натуральний логарифм від кожної точки), тоді вказана залежність стає лінійною, а загальну картину "псують" лише спостереження 38 та 83. Спостереження 133 тепер входить до сукупності, тому вважаємо що це не є викидом. Після побудови першої моделі регресії будемо робити висновки щодо сутності двох інших точок.

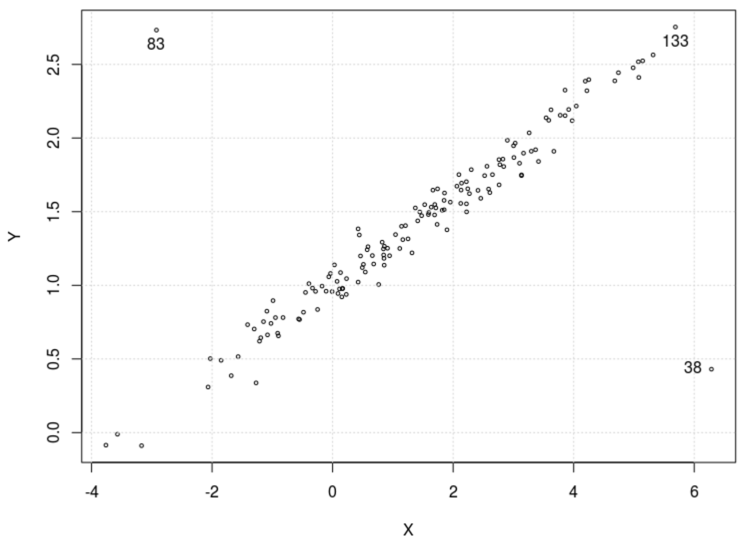

Рис. 2: Діаграма розсіювання спостережень вибірки $\ln \Xi$.

3 Підгонка моделі.

З минулого розділу можна висунути припущення про те, що нелінійна залежність між X та Y описується степеневою функцією:

$$\forall j \in \overline{1,n}:\ Y_j \approx C X_j^{\alpha} \tag{1}$$

Лінеаризація дає можливість перетворити (1) на модель простої лінійної регресії:

$$\forall j \in \overline{1,n}:\ \ln Y_j \approx \ln C + \alpha \ln X_j \tag{2}$$

Перша модель базується на формулі (2), враховуючи всі спостереження Ξ:

```
# Зчитування даних
c.table <- read.table('c4.txt', header=T, sep='\t', fileEncoding = "UTF16LE")
# Логарифмування спостережень
ln.c.table <- log(c.table)
# Перша спроба підігнати модель
lm.log.1 <- lm(Y ~ X, data=ln.c.table)
```

Підставимо оцінки в (2):

$$\forall j \in \overline{1,n}:\ \ln Y_j \approx 1.0563073 + 0.2556595 \times \ln X_j$$

Маємо наступні відомості про модель:

```
Call:
lm(formula = Y ~ X, data = ln.c.table)

Residuals:
     Min       1Q   Median       3Q      Max
-2.23284 -0.09118  0.00611  0.07934  2.42391

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.05631    0.02954   35.75   <2e-16 ***
X            0.25566    0.01235   20.70   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2961 on 148 degrees of freedom
Multiple R-squared:  0.7433,Adjusted R-squared:  0.7416
F-statistic: 428.6 on 1 and 148 DF,  p-value: < 2.2e-16
```

У звіті бачимо, що модель за висунутою нами формулою залежності має місце. Дисперсії отриманих оцінок для коефіцієнтів $\widehat{\ln C}, \hat{\alpha}$ низькі, а самі значення є значущими за тестом Стьюдента. З іншого боку, коефіцієнт детермінації моделі кульгає. Числових значень недостатньо, застосуємо графічне представлення результатів, щоб зрозуміти у чому проблема.

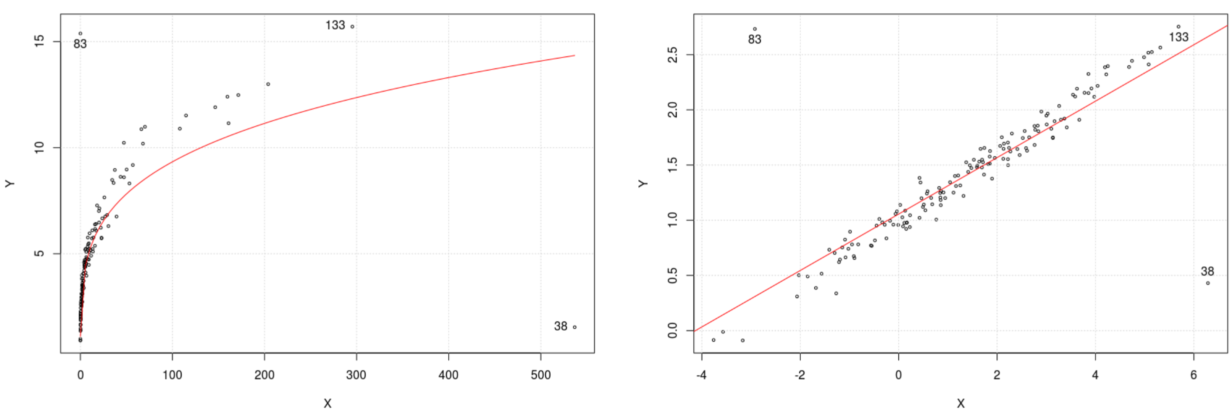

Рис. 3: Діаграма розсіювання спостережень Ξ (справа $\ln \Xi$) та підігнана крива за оцінками першої моделі.

З рисунків видно, що оцінка кута нахилу (показника степеня) не відповідає дійсності та на його основі маємо некоректні криві для опису залежності між об'єктами: графіки не охоплюють значну частину точок на площині.

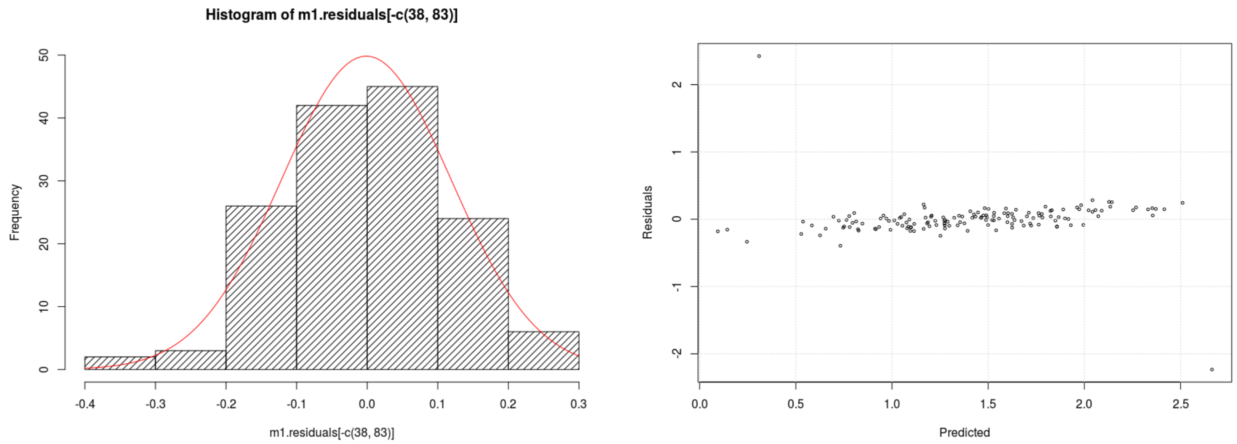

Рис. 4: Гістограма абсолютних частот залишків з нормованим графіком щільності нормального розподілу (зліва). Справа - діаграма розсіювання "прогноз-залишок".

Гістограма залишків показує, що їх розподіл близький до гауссового. Хоча видно, що хвости не є подібними між собою - на правий хвіст припадає більша кількість елементів. QQ-діаграма також натякає на гауссовість.

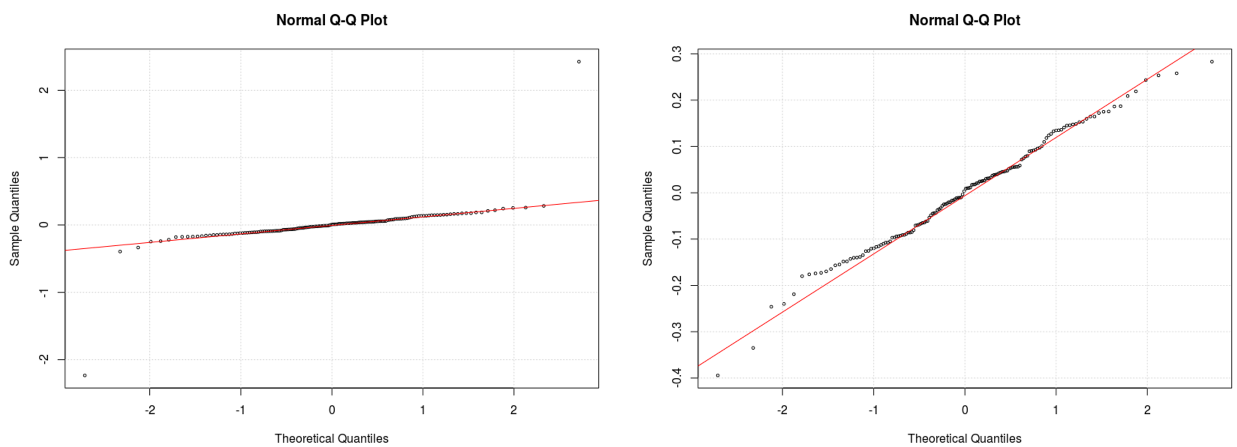

Рис. 5: QQ-діаграми. Зліва з урахуванням спостережень 38, 83; справа без них.

Така регресійна модель є недоречною, тому спробуємо ще раз. У другій моделі все залишається так само, але не будемо враховувати спостереження 38 та 83, оскільки є викидами.

```
# Масив з індексів тих спостережень, які бажано усунути
out <- c(38, 83)
# Друга спроба підігнати модель
lm.log.2 <- lm(Y ~ X, data=ln.c.table[-out,])
```

Підставимо нові оцінки в (2):

$$\forall j \in \overline{1,n}: \ \ln Y_j \approx 0.9999856 + 0.2958150 \times \ln X_j$$

Отримані відомості:

```
Call:
lm(formula = Y ~ X, data = ln.c.table[-c(38, 83), ])

Residuals:
      Min        1Q    Median        3Q       Max
-0.286961 -0.065794  0.003973  0.061415  0.258190

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.999986   0.009422  106.13   <2e-16 ***
X           0.295815   0.004026   73.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09291 on 146 degrees of freedom
Multiple R-squared:  0.9737,Adjusted R-squared:  0.9735
F-statistic:  5398 on 1 and 146 DF,  p-value: < 2.2e-16
```

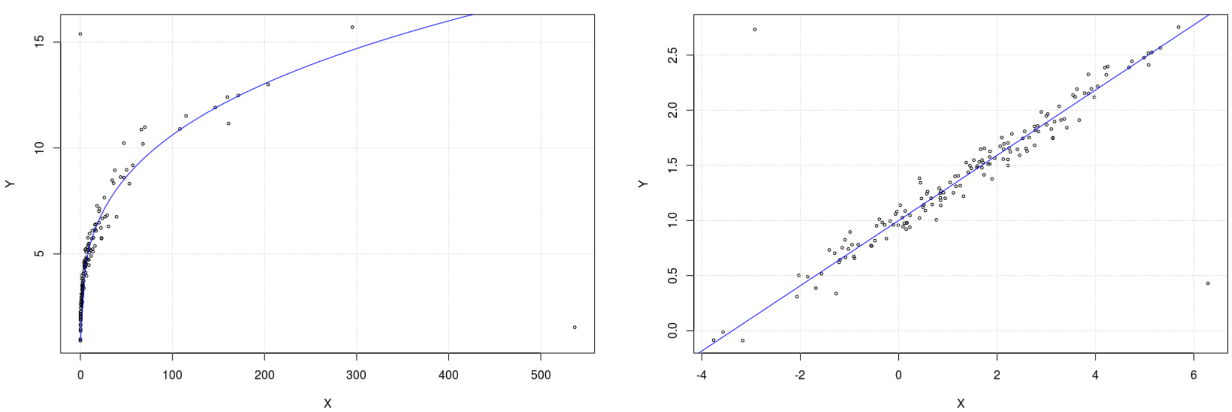

Рис. 6: Діаграма розсіювання спостережень Ξ (справа $\ln \Xi$) та підігнана крива за оцінками другої моделі.

Маємо кращі результати: розташування підігнаних кривих досить хороше, як видно з графіків. У даній моделі дисперсії оцінок, порівняно з першими, суттєво нижчі; коефіцієнт детермінації досягає ≈ 0.97.

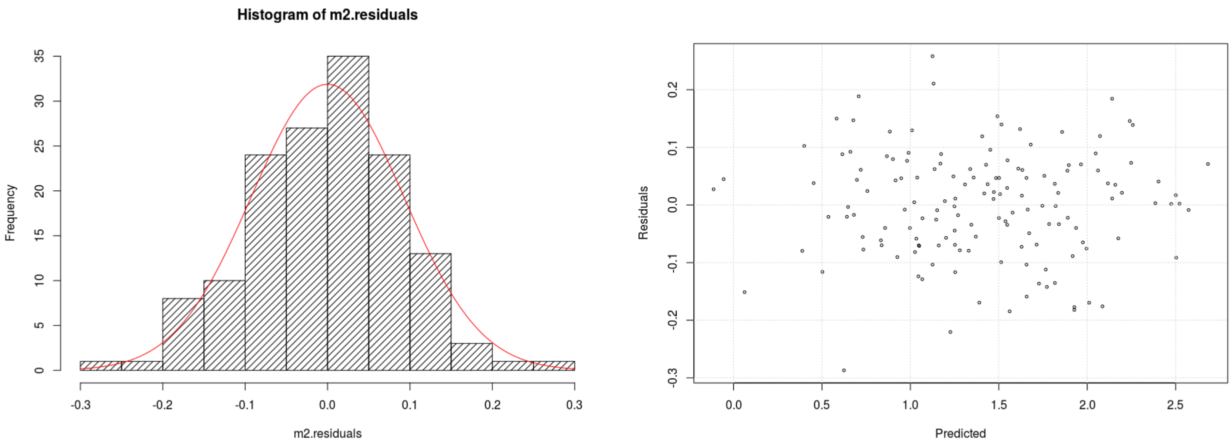

Рис. 7: Гістограма абсолютних частот залишків з нормованим графіком щільності нормального розподілу (зліва). Справа - діаграма розсіювання "прогноз-залишок".

Гістограма та QQ-діаграма залишків свідчать про те, що розподіл може бути гауссовим. Інших цікавих явищ не спостерігається.

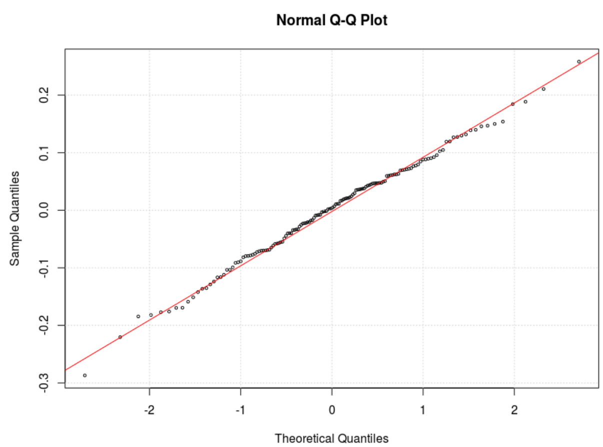

Рис. 8: QQ-діаграма залишків другої моделі.

4 Висновки.

Визначити формулу залежності між об'єктами вибірки було неважко: достатньо було спробувати взяти логарифм та зобразити перетворені точки на площині. Але підігнати регресійну модель вдалося з другої спроби, заважали викиди. В цілому, вдалося побудувати таку модель, яка гарно описує знайдену нелінійну залежність.