Лабораторна робота №4
з дисципліни ”регресійний аналіз”
Варіант №4

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

8 березня 2021 р.

1 Вступ.

У даній роботі проаналізовано дані обстеження домогосподарств України, проведеного Держкомстатом у 2001 році. Побудована регресійна модель залежності між віком голови домогосподарства та повними витрататами. Використано тест Чоу (Фішера про загальну лінійну гіпотезу для розшарованої вибірки) для перевірки, чи необхідно будувати різні регресійні моделі в залежності від наявності ванни.

2 Дескриптивна статистика вхідних даних. Виявлення залежності.

Після побудови гістограм абсолютних частот для дослідження форм розподілу маємо таку справу з початковими даними:

1. Гістограма за даними регресора показує, що розподіл може бути бімодальним, суттєве коливання між значеннями 45 та 65 не здаються випадковими. Взагалі кажучи, нагадує двокомпоненту суміш.

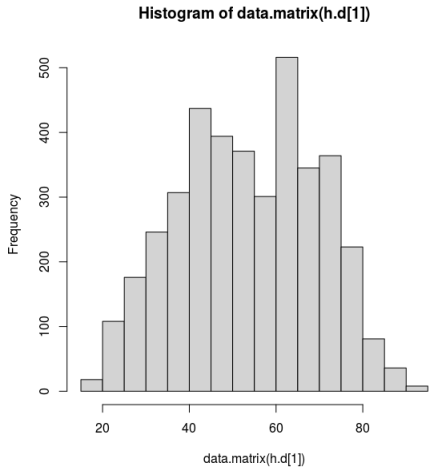

Рис. 1: Гістограма абсолютних частот регресора.

2. Гістограма значень відгуку, тобто повних витрат, має форму дзвона із зкошенням в ліву сторону. Це дещо нагадує за формою логнормальний розподіл. Справді, якщо прологарифмувати дані відгуку, спостерігаємо розподіл близький до нормального (ясно, що з результатів візуалізації та деякої інтуїції).

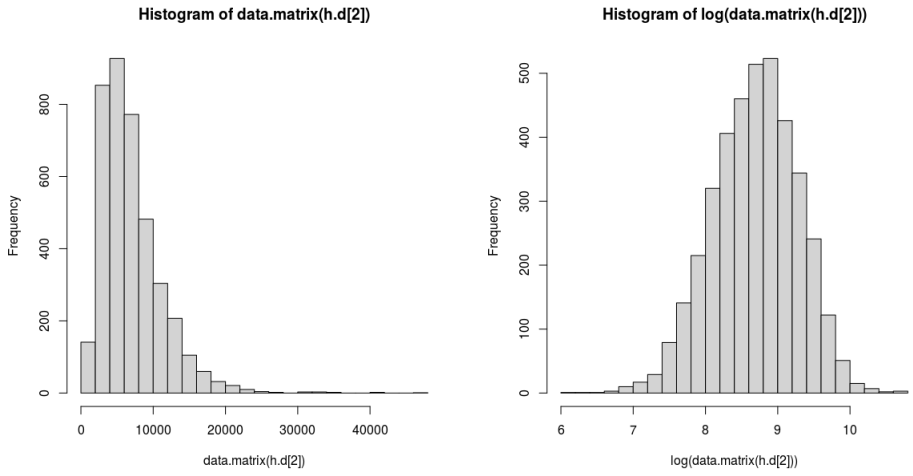

Рис. 2: Гістограми абсолютних частот відгуку. Зліва - до взяття логарифму, справа - після взяття.

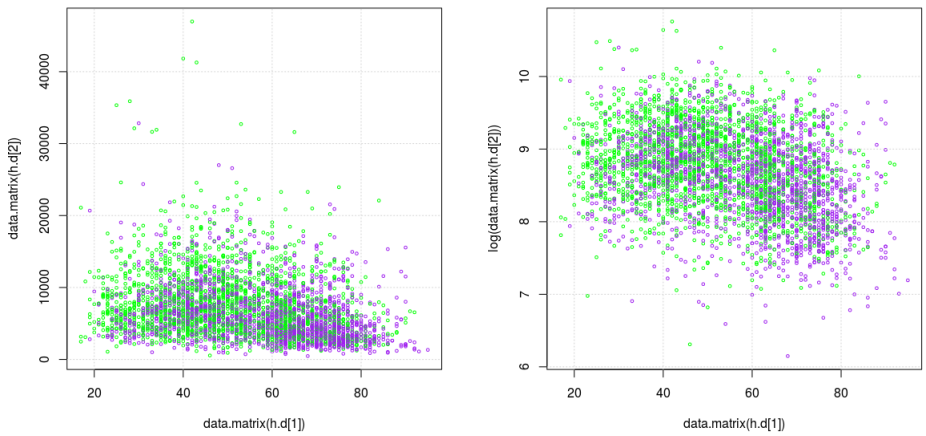

Рис. 3: Діаграма розсіювання спостережень. Зеленим кольором відмічено точки, для яких ванна є в наявності, а фіолетовим - ті спостереження, де ванни немає. Як і на попередньому рисунку, зліва зображені дані до логарифмування відгуку, а справа - після логарифмування.

Діаграма розсіювання показує, що дані формують велику хмару точок на площині. Якщо розділити ці точки в залежності від значення факторної змінної про наявність ванни ('BATH'), то бачимо, що насправді ця хмарина складається з двох менших, які відрізняються незначним зсувом вздовж осі абсцис.

Регресійну модель можна застосувати, але буде слушною думкою одразу перевірити, чи варто будувати різні моделі, в залежності від значення факторної змінної. Для цього скористаємося тестом Чоу.

3 Тест Фішера у багатьох варіаціях.

3.1 Тест Чоу.

Спочатку перевіримо гіпотезу про те, чи потрібно використовувати різні регресійні моделі. Загальний вигляд лінійної моделі матиме вигляд:

$$\ln Y_j \approx (b_0^{no}\delta^{no} + b_0^{yes}\delta^{yes}) + (b_1^{no}\delta^{no} + b_1^{yes}\delta^{yes})X_j,\ j \in \overline{1,N},\ N = 3931$$

$$\delta^{no} = \begin{cases} 1, & BATH_j = 1 \\ 0, & BATH_j = 2 \end{cases},\ \delta^{yes} = 1 - \delta^{no}$$

Формулюємо статистичні гіпотези:

$$H_0 : b_j^{no} = b_j^{yes},\ j = 1, 2$$

$$H_1 : \text{Хоча б один з коефіцієнтів відрізняється}$$

За виконання H_0, (обмежена) модель набуває вигляду:

$$\ln Y_j \approx b_0 + b_1 X_j,\ j \in \overline{1,N}$$

Кількість регресорів в загальній моделі, що розглядається, дорівнює $d = 4$ (а в обмеженій - $d_r = 2$). Статистика тесту Чоу має вигляд:

$$F_{emp} = \frac{\frac{1}{d_r}(||U_0||^2 - ||U_1||^2)}{\frac{1}{N-d}||U_0||^2} = \frac{\frac{1}{2}(||U_0||^2 - ||U_1||^2)}{\frac{1}{N-4}||U_0||^2},$$

де U_j - залишки моделі регресії за виконанням H_j, $j = 0, 1$. F_{emp} будемо порівнювати з $F_{theor} = Q^{F(d_r, N-d)}(1-\alpha)$, $\alpha := 0.05$

Сума квадратів залишків необмеженої регресійної моделі становить $||U_0||^2 = 1212.732$, а обмеженої: $||U_1||^2 = 1230.488$. Числа, звісно, великі, але більше цікавить значення статистики тесту: $F_{emp} = 28.747202 > 2.998019 = F_{theor}$. Оскільки значення статистики перевищує критичний рівень тесту, то приймається альтернативна гіпотеза про параметри регресійної моделі. Далі перевіримо, чи є значущою відмінність між коефіцієнтами зсуву в загальній моделі. Тільки перш ніж переходити до цього, додамо в кінці отримані оцінки коефіцієнтів регресії для використаних моделей:

$$Unrestricted: \hat{b}_0^{no} = 9.29565029,\ \hat{b}_1^{no} = -0.01008343,\ \hat{b}_0^{yes} = 9.26064065,\ \hat{b}_1^{yes} = -0.01191635$$
$$Restricted: \hat{b}_0 = 9.32903538,\ \hat{b}_1 = -0.01187054$$

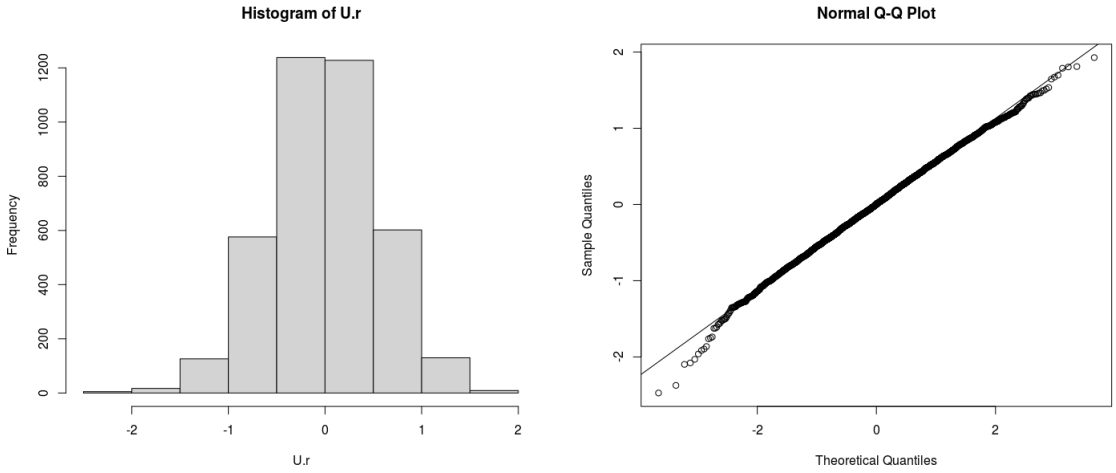

Рис. 4: Гістограма абсолютних частот і QQ-діаграма залишків обмеженої моделі.

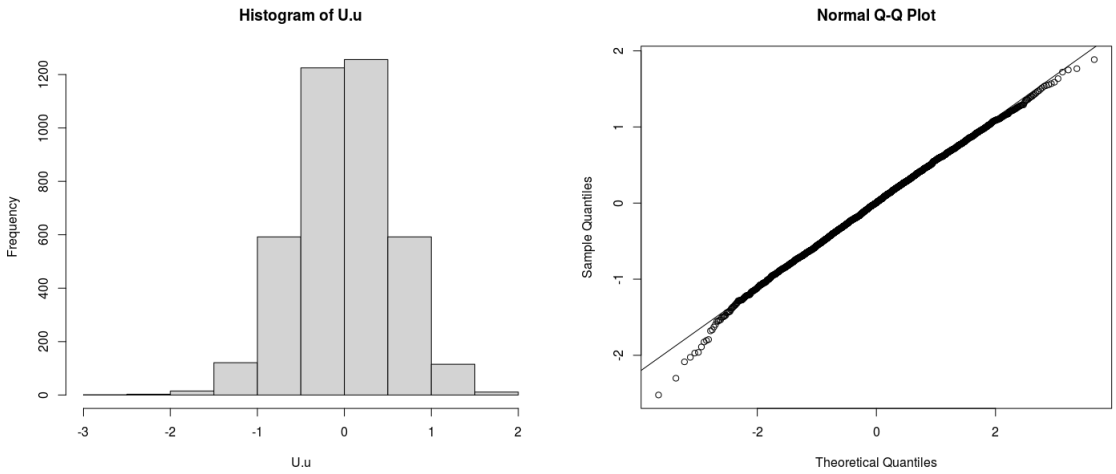

Рис. 5: Гістограма абсолютних частот і QQ-діаграма залишків необмеженої моделі.

3.2 Перевірка гіпотези про рівність коефіцієнтів зсуву.

Знову, формулюємо такі статистичні гіпотези:

$$H_0 : b_0^{no} \neq b_0^{yes},\ b_1^{no} = b_1^{yes}$$
$$H_1 : b_j^{no} \neq b_j^{yes},\ j = 1, 2$$

Обмежена модель, у даному випадку, така:

$$\ln Y_j \approx (b_0^{no}\delta^{no} + b_0^{yes}\delta^{yes}) + b_1 X_j,\ j \in \overline{1, N}$$

Кількість регресорів в такій дорівнює $d_r = 3$. Для необмеженої все так само, як і в попередньому випадку. Запишемо статистику тесту Фішера:

$$F_{emp} = \frac{\frac{1}{3}(||U_0||^2 - ||U_1||^2)}{\frac{1}{N-4}||U_0||^2}$$

Тоді $F_{emp} = 0.8525359 < 2.6071707 = F_{theor}$. Звідси приймаємо гіпотезу про те, що коефіцієнти зсуву відрізняються, а різниця між коефіцієнтами нахилу несуттєва. Отримані оцінки коефіцієнтів для моделі за H_0:

$$b_0^{no} = 9.3391173,\ b_0^{yes} = 9.2036853,\ b_1 = -0.0109355$$

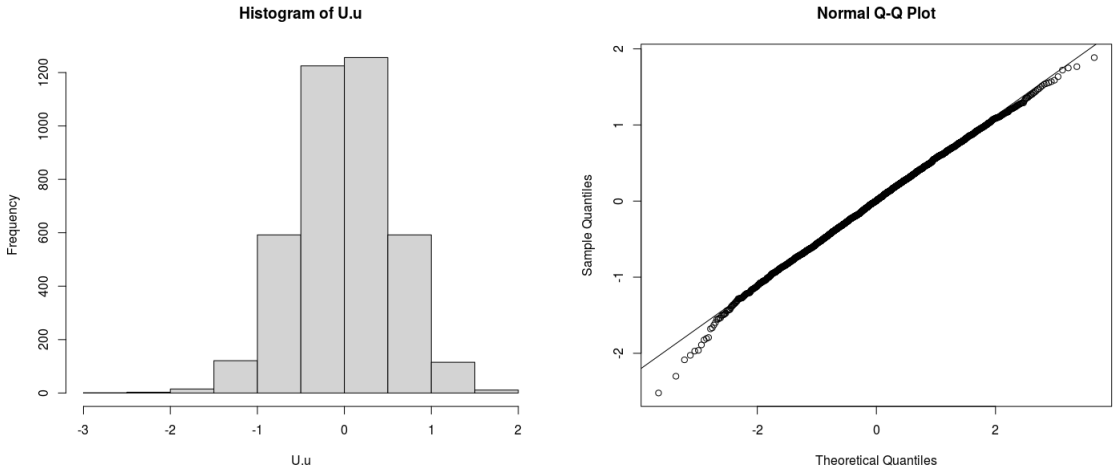

Рис. 6: Гістограма абсолютних частот і QQ-діаграма залишків іншої обмеженої моделі.

4 Висновки.

Дослідили, що в залежності від значення факторної змінної, слід будувати відповідну регресійну модель. Розподіл похибок адекватний.

5 Додаток.

5.1 Програмна реалізація.

```
alpha <- 0.05

h.table <- read.table('house01.txt', header=T)

# Побудуйте регресійну модель залежності між віком голови домогосподарства
# (AGE_HEAD) та повними витратами (TOTALEXP). Чи потрібно враховувати при
# побудові моделі наявність ванни або душу (BATH)?

h.d <- h.table[c("AGE_HEAD", "TOTALEXP", "BATH")]

plot(data.matrix(h.d[1]),data.matrix(h.d[2]),
     col = c('green', 'purple')[data.matrix(h.d[3])], cex = 0.5);grid()

plot(data.matrix(h.d[1]), log(data.matrix(h.d[2])),
     col = c('green', 'purple')[data.matrix(h.d[3])], cex = 0.5);grid()

plot(data.matrix(h.d[1])[h.d$BATH == 1],
     log(data.matrix(h.d[2]))[h.d$BATH == 1],
     col = 'green', cex = 0.5);grid()

plot(data.matrix(h.d[1])[h.d$BATH == 2],
     log(data.matrix(h.d[2]))[h.d$BATH == 2],
     col = 'purple', cex = 0.5);grid()


X.u <- as.matrix(
  data.frame(
    b.0.0 = (h.d[3] == 1) * 1,
    b.1.0 = (h.d[3] == 1) * h.d$AGE_HEAD,
    b.0.1 = (h.d[3] == 2) * 1,
    b.1.1 = (h.d[3] == 2) * h.d$AGE_HEAD
  )
)

n <- nrow(X.u)

Y <- log(data.matrix(h.d$TOTALEXP))

A.u <- t(X.u)%*%X.u
print(A.u)
print(det(A.u))

A.u.inv <- solve(A.u)
b.u <- A.u.inv%*%t(X.u)%*%Y
print(b.u)
```

```
X.r <- matrix(cbind(1 + numeric(n), h.d$AGE_HEAD), nrow = n, ncol = 2)

A.r <- t(X.r)%*%X.r
print(A.r)
print(det(A.r))

A.r.inv <- solve(A.r)
b.r <- A.r.inv%*%t(X.r)%*%Y
print(b.r)

d <- ncol(X.u)
d.r <- ncol(X.r)

U.u <- Y - X.u%*%b.u
U.r <- Y - X.r%*%b.r

sq.U.u <- sum(U.u^2)
sq.U.r <- sum(U.r^2)
F.chou.stat <- ((1/d.r) * (sq.U.r - sq.U.u))/((1/(n - d)) * sq.U.u)
F.chou.theo <- qf(1 - alpha, d.r, n - d)
print(c(F.chou.stat, F.chou.theo))

X.r.new <- as.matrix(
  data.frame(
    b.0.0 = (h.d[3] == 1) * 1,
    b.0.1 = (h.d[3] == 2) * 1,
    angle = h.d$AGE_HEAD
  )
)

n <- nrow(X.r.new)

A.r.new <- t(X.r.new)%*%X.r.new
print(A.r.new)
print(det(A.r.new))

A.r.new.inv <- solve(A.r.new)
b.r.new <- A.r.new.inv%*%t(X.r.new)%*%Y
print(b.r.new)

U.r.new <- Y - X.r.new%*%b.r.new
sq.U.r.new <- sum(U.r.new^2)
d.new <- 3

F.fis.stat <- ((1/d.new) * (sq.U.r.new - sq.U.u))/((1/(n - d)) * sq.U.u)
F.fis.theo <- qf(1 - alpha, d.new, n - d)
print(c(F.fis.stat, F.fis.theo))
```