Лабораторна робота №1
з дисципліни "Чисельні методи у статистиці"

Горбунов, 5 курс, "Прикладна та теоретична статистика"

12 травня 2022 р.

Вступ.

У даній роботі наведено результати зі статистичної обробки даних, що подано у таблиці world95. Досліджено залежність тривалості життя населення від економічних та демографічних показників. На основі виявленої залежності, побудовано модель багатовимірної лінійної регресії. Перевірено адекватність цієї моделі на основі відповідних тестів для залишків. У рамках цієї роботи використано такі пакети в R, як "readxl", "corrplot", "car", "gvlma".

Хід роботи.

Постановка задачі.

Маємо справу з таблицею world95 – це статистичні дані станом на 1995 р. із ілюстративних прикладів програми SPSS. Пропущені дані замінено на "?" для сумісності з програмами типу PAST. Поставлено наступні задачі: проаналізувати, з якими чинниками пов'язано відмінність у тривалості життя в різних державах світу, і чи відрізняється цей зв'язок у різних регіонах світу. Використати ті або інші методики регресійного аналізу та програмування в системі R.

Короткий огляд таблиці.

Таблиця складається з $n = 109$ рядків (далі – спостережень) та 29 колонок. Таблиця містить деякі економічні та деміографічні показники для кожної з n країн. Деякі з колонок є перетвореннями (як лінійними, так і нелінійними) від інших колонок (наприклад, взяти змінні BIRTH_RT, DEATH_RT і B_TO_D – остання змінна є часткою першого на другого). Взагалі спочатку були сплутані значення у колонках RELIGION та RELIG_KOD (код був у колонці для назв переважаючих віросповідань, а відповідно назви у колонці з кодуваннями). А для конкретного виду клімату був відомий лише код, а не назва переважаючого клімату. Однак це незначні помилки у таблиці, які можна з легкістю виправити.

Охарактеризуємо лише основну частину колонок (не враховуємо ті колонки, що містять кодування факторних змінних типу регіону, релігії, клімату):

Назва колонки	Тлумачення
COUNTRY	Назва країни
POPULATN	Населення у тисячах
DENSITY	Кількість осіб / км^2
URBAN	Частка людей, що проживає у містах (%)
RELIGION	Переважаюче віросповідання
LIFEEXPF	Середня тривалість життя жінки
LIFEEXPM	Середня тривалість життя чоловіка
LITERACY	Частка людей, що читає (%)
POP_INCR	Приріст населення (% за рік))
BABYMORT	Смертність по новонародженим (кількість смертей на 1000 новонароджених)
GDP_CAP	Валовий внутрішній продукт на душу населення
REGION	Регіон або економічна група
CALORIES	Щоденна кількість вижитих калорій
AIDS	Кількість випадків виявлення СНІДу
BIRTH_RT	Народжуваність на 1000 осіб
DEATH_RT	Смертність на 1000 осіб
AIDS_RT	Кількість випадків виявлення СНІДу / 100000 осіб
LOG_GDP	Логарифм (з основою 10) від значень GDP_CAP
LG_AIDSR	Логарифм (з основою 10) від значень AIDS_RT
B_TO_D	Частка народжуваності до смертності
FERTILTY	Плодючість: середня кількість дітей
LOG_POP	Логарифм (з основою 10) від значень Population
CROPGROW	Зростання врожаю
LIT_MALE	Частка чоловіків, що читає (%)
LIT_FEMA	Частка жінок, що читає (%)
CLIMATE	Переважаючий клімат

Табл. 1: Тлумачення колонок у таблиці word95.

Початкова обробка даних.

Для так званої "безперебійної" роботи із даними, потрібно перекодувати пропуски з "?" на змінну NA (Not Available), з якою працювати набагато гнучкіше. Гнучкість пояснюється в тому сенсі, що тип даних відповідних колонок зберігається, порівняно з початковим кодуванням (яке збивало колонки у рядковий тип). Наприклад, це можна реалізувати наступним чином в R: спочатку перекодовуємо символи, а далі присвоюємо колонкам числовий тип.

```
# Зчитування даних
dat <- data.frame(read_xls("world1995.xls")[,-1])
# Колонки з числовими значеннями
cols.choose <- c("POPULATN", "DENSITY", "URBAN", "RELIG_KOD",
                 "LIFEEXPF", "LIFEEXPM", "LITERACY", "POP_INCR",
                 "BABYMORT", "GDP_CAP", "REGION_KOD", "CALORIES",
                 "AIDS", "BIRTH_RT", "DEATH_RT", "AIDS_RT", "LOG_GDP",
                 "LG_AIDSR", "B_TO_D", "FERTILTY", "LOG_POP",
                 "CROPGROW", "LIT_MALE", "LIT_FEMA", "CLIMATE_KOD")
...
```

```
...
for(col.c in cnames)
{
  # Перекодування пропусків
  dat[dat[, col.c] == "?", col.c] <- NA
  if(col.c%in%cols.choose)
    # Якщо колонка є числовою, то задаємо їй потрібний тип даних
    dat[, col.c] <- as.numeric(dat[, col.c])
}
```

Виявлення залежності.

Інтуїтивно можна вважати, що у добре розвинених країнах рівень життя є кращим, тому й тривалість життя може бути непоганою. Природньо також допустити безпосередній вплив народжуваності й смертності на птривалість, однак це краще обгрунтовується до відслідковування залежності популяції від цих факторів. Додатково можна припустити, що освічені люди довше живуть (бо в певному наближенні знають що роблять). Окрім певних міркувань, потрібно ще чимось підкріпити. Обчислимо матриці кореляцій (по числовим змінним) за Пірсоном та Спірменом, зобразимо їх. Відштовхуючись від них, зробимо конкретні висновки про можливі форми й чинники залежності.

```
I <- cols.choose[-c(4,11,25)] # -c(4,11,25) - вилучення факторних колонок
# Підрахунок матриць кореляцій
c.p <- cor(dat[,I], method = "pearson", use = "complete.obs")
c.s <- cor(dat[,I], method = "spearman", use = "complete.obs")
# Візуалізація кореляційних матриць
corrplot(c.p, method = "color", type = 'upper', addCoef.col = 'black')
corrplot(c.s, method = "color", type = 'upper', addCoef.col = 'black')
```

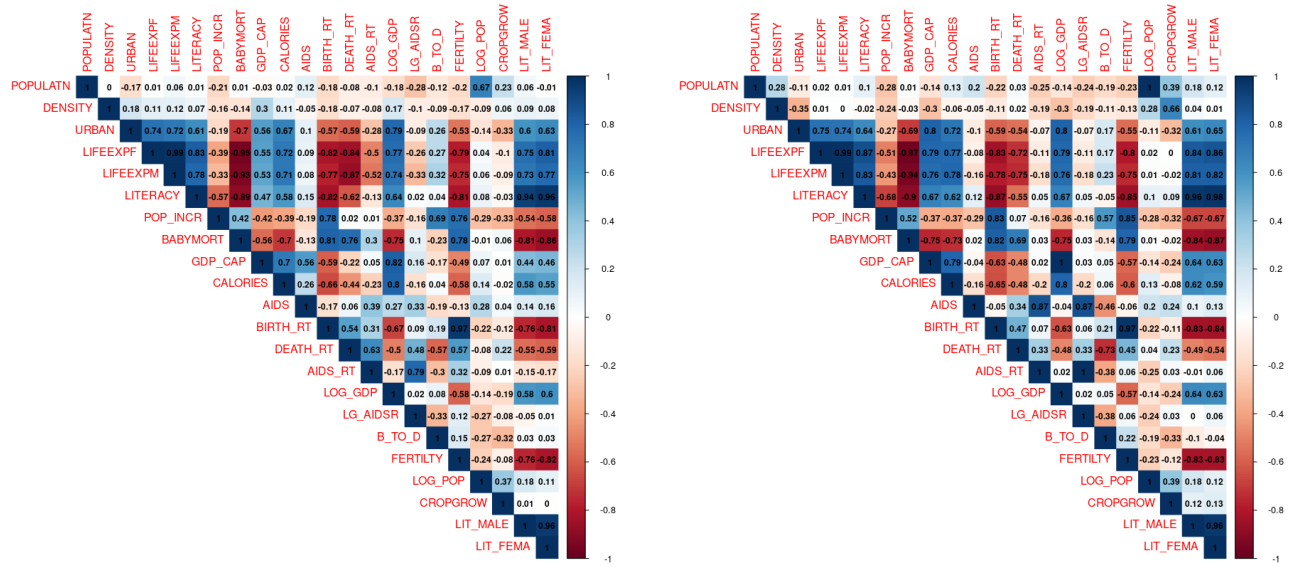

Рис. 1: Візуалізація матриць кореляцій (зліва – за Пірсоном, справа – за Спірменом).

Як видно з візуалізацій, наші попередні припущення про залежність тривалості життя від конкретних чинників виправдовується. Сильно виражену лінійну залежність маємо від таких змінних, як LITERACY, BIRTH_RT, DEATH_RT, CALORIES, AIDS (та деякі інші). Цікаво зауважити, що кореляція тривалості життя від народжуваності або плодючості є від'ємною. Не обговорювалося, але цілком очевидною є виявлення позитивної кореляції між тривалістю життя та щоденною кількістю спожитих калорій: грубо кажучи, коли людина накопичує більше "енергії", то має більше сил на виживання.

Лінійна залежність. Відбір змінних.

З попередніх результатів виявилося, що тривалість життя корелює з багатьма змінними у таблиці. Втім, треба відібрати для прогнозної моделі лише ті, що мають реальний зв'язок з досліджуваною змінною (інакше кажучи, потрібно вибрати змінні адекватно, від цього реалізація моделі стає не тільки простою, а й легшою для інтерпретації). Серед кандидатів на побудову моделі можна висунути досить тісну групу змінних, вплив яких обговорювався раніше: кількість вживаних калорій, рівень освіченості населення, ВВП країни та деякі базові демографічні показники. Варто ще брати до уваги те, що ми обмежені в обсязі даних, тому сильну складну модель робити не варто (особливо якщо згадати, що в даних наявні пропуски). Виходячи з емпіричного правила, що на оцінення одного параметра припадає приблизно 10 спостережень, то для лінійної моделі доцільно було б взяти не більше 4-5 змінних (не забуваємо, що доведеться ще оцінювати коефіцієнт зсуву). Тому у рамках цієї роботи спробуємо будувати фундамент навколо грамотності (LITERACY / LIT_MALE / LIT_FEMALE), ВВП країни (LOG_GDP), народжуваності (BIRTH_RT) та смертності (DEATH_RT). Цікаво перебирати інші змінні для розгляду та відбору, можливо, кращих моделей, однак залишимо це на вільну годину. Подивимося на діаграми розсіювання.

```
# Побудова діаграм розсіювання для тривалості життя чоловіків від змінних вище
plot(LIFEEXPM ~ LIT_MALE, data = dat)
# ... і так далі. Аналогічний код для діаграм, де фігурує тривалість життя жінок
```

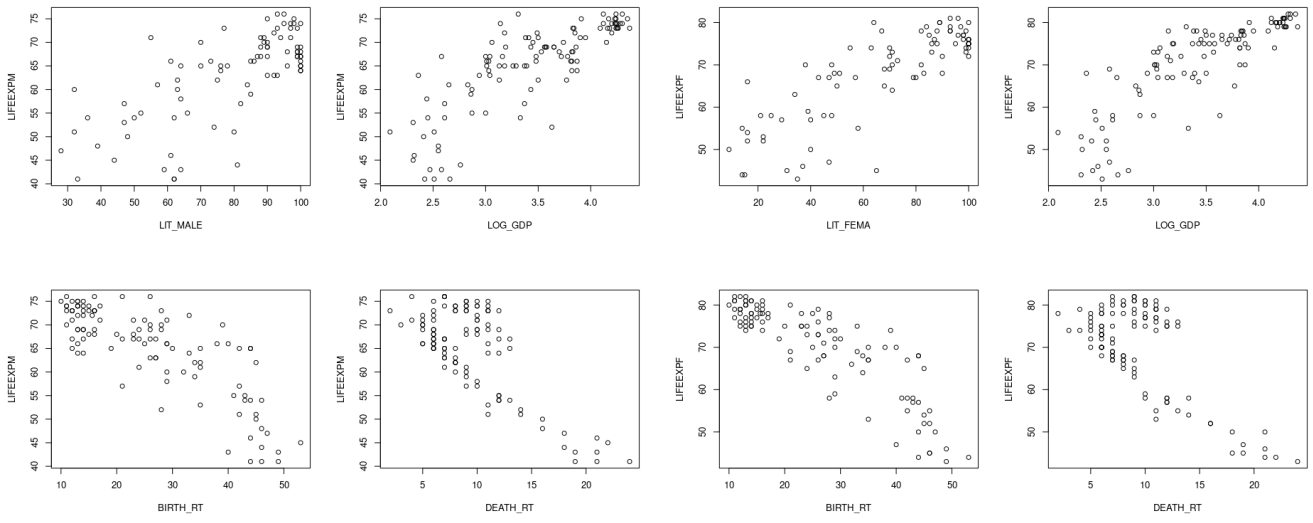

Рис. 2: Діаграми розсіювання тривалості життя (таблиці зліва – для чоловіків, справа – для жінок) та змінних типу грамотності, народжуваності, смертності, ВВП країни.

Коротко по рисункам, то справді відстежується залежність, схожа на лінійну (якщо придивитися, то для деяких пар слабко схожа на нелінійну). Для діаграм розсіювання тривалості від смертності бачимо зверху велику хмарину, яка відрізняється від загального тренду, однак зберігає знак кореляції. Пов'язати даний момент із тим що кожна хмарина відповідає конкретному фактору (беручи до уваги регіон / релігію / клімат тощо) неможливо, бо у кожній з двох хмарин розміщені точки з різних підгруп:

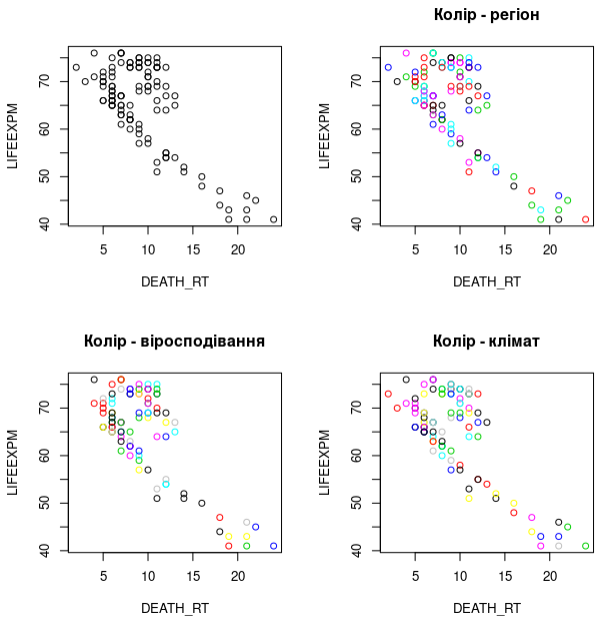

Рис. 3: Діаграма розсіювання тривалості життя чоловіків відносно смертності. На кожному з рисунків точки розфарбовано у відповідності до тієї підгрупи, до якої вона належить.

Як раніше зазначилося, спостереження, що утворюють другу (вищу) хмарину є впливовими, однак недоречно їх "викидати", а й з ними загальний тренд суттєво не зміниться (принаймні зі знаком все гаразд). Переходимо до побудови регресійної моделі.

Побудова моделі. Дослідження якості моделі.

Спочатку беремо до уваги лінійну регресійну модель вигляду:

$$\text{LIFEEXPM}_j = \beta_0 + \beta_1 \cdot \text{LIT\_MALE}_j + \beta_2 \cdot \text{LOG\_GDP}_j + \beta_3 \cdot \text{BIRTH\_RT}_j + \beta_4 \cdot \text{DEATH\_RT}_j + \varepsilon_j, \tag{1}$$

де ε_j є випадковою похибкою, розподіл якої буде формуватися на основі відповідних тестів для залишків прогнозу моделі. Бажано, щоб похибки мали гауссів розподіл з нульовим середнім та сталою дисперсією (тобто не залежить від j), а самі похибки були б некорельованими (щось сильніше відстежити важко, тому обмежимося лише цим).

Зауважте, що в формулі фігурують показники для чоловіків. Судячи з графічних результатів, то результати мають вийти більш-менш однаковими і з показниками для жіночої статі. Можна було б формально застосувати тест про наявність розшарування у моделі вище, однак довіримося висунутим міркуванням на основі раніше отрмианих результатів. Відомості про підгонку параметрів, якість моделей за жіночими показниками покажемо в кінці роботи. Саму ж підгонку зробимо за допомогою методу найменших квадратів:

```
lm.based <- lm(LIFEEXPM ~ LIT_MALE + LOG_GDP + BIRTH_RT + DEATH_RT, data = dat)
```

Однак одного рядка не буде достатньо для подальшого аналізу моделі. Принаймні треба вивести звіт з підгонки параметрів у ній:

```
> summary(lm.based)

Call:
lm(formula = LIFEEXPM ~ LIT_MALE + LOG_GDP + BIRTH_RT + DEATH_RT,
    data = dat)

Residuals:
   Min     1Q Median     3Q    Max
-8.794 -1.777  0.258  1.617  5.939

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 63.02131    4.76422  13.228  < 2e-16 ***
LIT_MALE     0.04640    0.02587   1.793 0.076702 .
LOG_GDP      3.90366    0.82542   4.729 9.54e-06 ***
BIRTH_RT    -0.20393    0.05004  -4.076 0.000107 ***
DEATH_RT    -1.10052    0.08029 -13.706  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.881 on 80 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.9105,        Adjusted R-squared:  0.906
F-statistic: 203.5 on 4 and 80 DF,  p-value: < 2.2e-16
```

Припустимо, що похибки задовольняють умови гауссової функціональної регресії. Це необхідно для того щоб вважати, що застосування тестів Стьюдента і Фішера є законним (далі дослідимо розподіл залишків, звідки зробимо висновки про наші похибки). Тоді можна наївно вважати (тобто одразу з результатів звіту), що модель непогано підібрали. Частка дисперсії, що пояснюється прогнозом моделі є досить високою (грубо кажучи можна вважати модель зібрала достатню інформацію щодо розкиду відгуку – тривалості життя). Майже всі коефіцієнти у ній можна вважати значущо відмінними від нуля для "хорошого" рівня значущості. От коли стандартний рівень значущості дорівнює $\alpha = 0.05$, то для коефіцієнта при змінній LIT_MALE, зауважимо, досягнутий рівень значущості тесту Стьюдента більший за стандартний, що начебто дає підстави для прийняття гіпотези про те, що коефіцієнт дорівнює нулю і змінною можна знехтувати. Це можна зробити, однак хто гарантує, що модель не втратить вагому інформацію для прогнозу замість вилучення зайвих "збурень"? Далі покажемо, що вилучення цієї змінної до кращих результатів не призведе.

Повертаючись до звіту, то звернімо увагу на доцільність застосування цієї моделі (в тому сенсі, що запропонована форма залежності може мати місце): отримані результати за тестом Фішера дають підстави про те, що залежність виявлена (чого, справді кажучи, можна було очікувати).

Тепер потрібно перевірити, чи задовольняють похибки моделі умовам гауссової регресії. До цієї задачі підійдемо з двох сторін: за допомогою графічних тестів для перевірки якості моделі та використанням тестів на нормальність (на прикладі тесту Шапіро-Вілка).

Спочатку застосуємо графічні "тести". Побудуємо діаграми для залишків та прогнозу.

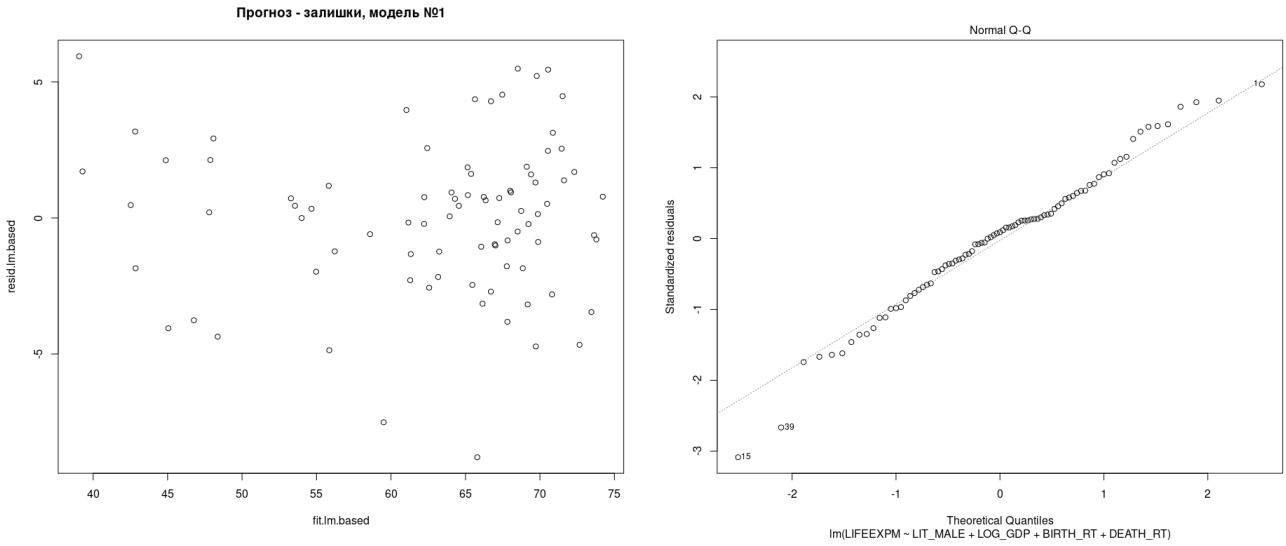

Рис. 4: Зліва – діаграма "Прогноз – залишки", справа – QQ-діаграма залишків.

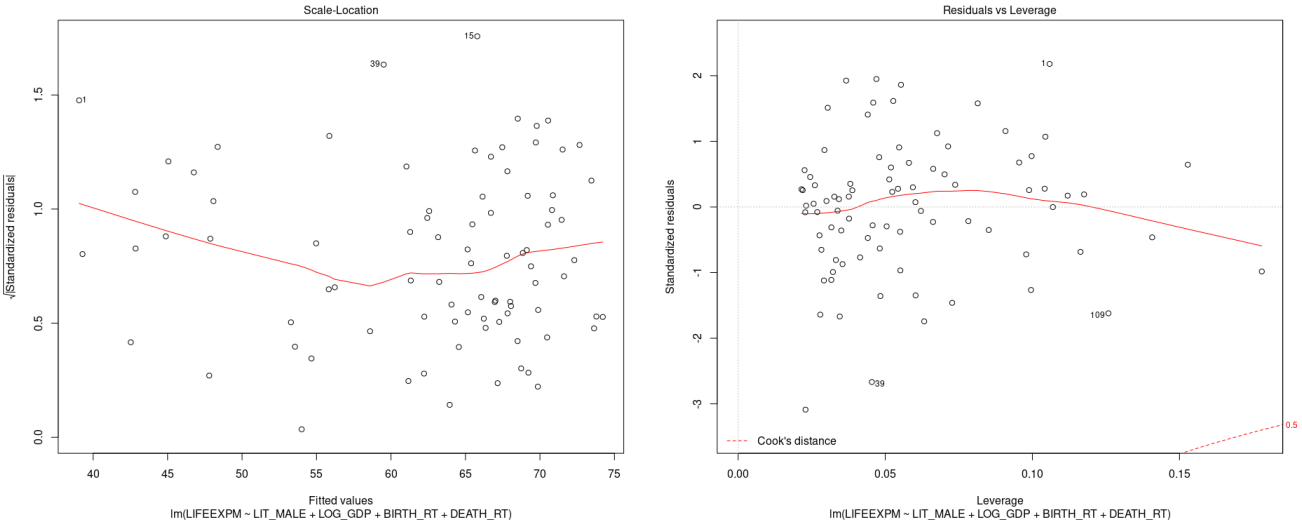

Рис. 5: Діаграма "Прогноз – стьюдентизовані залишки" зліва, та "Важіль – стьюдентизовані залишки" справа.

Верхня права діаграма показує узгодженість емпіричних квантилів залишків прогнозу із теоретичними квантилями нормального розподілу. Відхилення на кінцях тут, скоріше, пояснюється, нестачею спостережень, які б мали так звані "екстремальні" значення, що й призводить до більшої похибки. Спостереження з номерами 15 та 39 сильно відхиляються від загальної картини – це, зокрема, можна побачити на лівих діаграмах: для них відповідні залишки занадто високі, що дає підстави трактувати їх як викиди. Говорячи про верхню ліву діаграму, то закономірності відстуні (принаймні якщо дивитися неозброєним оком). Розкид залишків відносно однорідний. Судячи з нижнього правого рисунка, впливових спостережень досить мало (серед них або великі важелі, або великі стьюдентизовані залишки; аномальних відстаней Кука не відмічено). Тому можна припустити, що потрібні умови для похибок справді виконуються. Але ми підемо ще глибше.

З пакету gvlma застосуємо набір тестів для перевірки базових припущень про розподіл похибок у лінійній моделі. Вони базуються на детальнішому аналізі форми розподілу залишків:

- Global Stat – загальний висновок про модель (доречно взагалі брати до уваги чи ні);
- Skewness – коефіцієнт асиметрії;
- Kurtosis – коефіцієнт ексцесу;
- Link Function – коректність підбору функції зв'язку;
- Heteroscedasticity – перевірка гетероскедастичності.

```
> gvlma(lm.based)

Call:
lm(formula = LIFEEXPM ~ LIT_MALE + LOG_GDP + BIRTH_RT + DEATH_RT,
    data = dat)

Coefficients:
(Intercept)     LIT_MALE      LOG_GDP     BIRTH_RT     DEATH_RT
    63.0213       0.0464       3.9037      -0.2039      -1.1005

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.based)

                    Value p-value                   Decision
Global Stat         9.365 0.05259    Assumptions acceptable.
Skewness            1.678 0.19518    Assumptions acceptable.
Kurtosis            1.072 0.30043    Assumptions acceptable.
Link Function       4.681 0.03049 Assumptions NOT satisfied!
Heteroscedasticity 1.934 0.16437    Assumptions acceptable.
```

Як видно з результатів, то залишки модель задовольняє на рівні значущості $\alpha = 0.05$ майже всім умовам гауссової лінійної моделі, окрім гіпотези про коректність вибору функції зв'язку (напевно, пояснення випливає з міркувань про значення коефіцієнта при змінній LITERACY). Застосуємо тести Шапіро-Вілка, Бройша-Пагана та Дарбіна-Уотсона для перевірки гіпотез про гауссовість, сталу дисперсію та відстуність (простої) автокореляції залишків відповідно.

```
> shapiro.test(resid.lm.based)
        Shapiro-Wilk normality test
data:  resid.lm.based
W = 0.98141, p-value = 0.2594
...
```

```
...
> ncvTest(lm.based)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.1556937, Df = 1, p = 0.69315
> durbinWatsonTest(lm.based)
 lag Autocorrelation D-W Statistic p-value
   1      0.09305576      1.732092   0.182
 Alternative hypothesis: rho != 0
```

Тести дають підстави того, що залишки не мають аномальних властивостей. Як графічні тести, так і не графічні підтверджують висунуте припущення про адекватність моделі (ігноруючи незначну проблему з коефіцієнтом при LITERACY). Тепер покажемо, що вилучення змінної LITERACY з моделі не дасть хороших результатів. Тобто ми розглядаємо спрощену модель:

$$\text{LIFEEXPM}_j = \beta_0 + \beta_1 \cdot \text{LOG\_GDP}_j + \beta_2 \cdot \text{BIRTH\_RT}_j + \beta_3 \cdot \text{DEATH\_RT}_j + \varepsilon_j,$$

Продемонструємо звіт по моделі:

```
> summary(lm.based.red)

Call:
lm(formula = LIFEEXPM ~ LOG_GDP + BIRTH_RT + DEATH_RT, data = dat)

Residuals:
    Min      1Q  Median      3Q     Max
-9.6501 -1.1905  0.1227  1.5808  5.9976

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.99790    3.21670  20.517  < 2e-16 ***
LOG_GDP      4.74138    0.68737   6.898 4.24e-10 ***
BIRTH_RT    -0.29027    0.03397  -8.545 1.17e-13 ***
DEATH_RT    -1.02563    0.06944 -14.769  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.775 on 104 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9133,        Adjusted R-squared:  0.9108
F-statistic: 365.1 on 3 and 104 DF,  p-value: < 2.2e-16
```

З одного боку може здаватися, що результати вийшли в певному сенсі кращі, ніж у попередній моделі. Однак коли справа доходить до аналізу залишків прогнозу в моделі, картина стає менш вражаючою.

Подивимося на діаграми для залишків та прогнозу нової моделі:

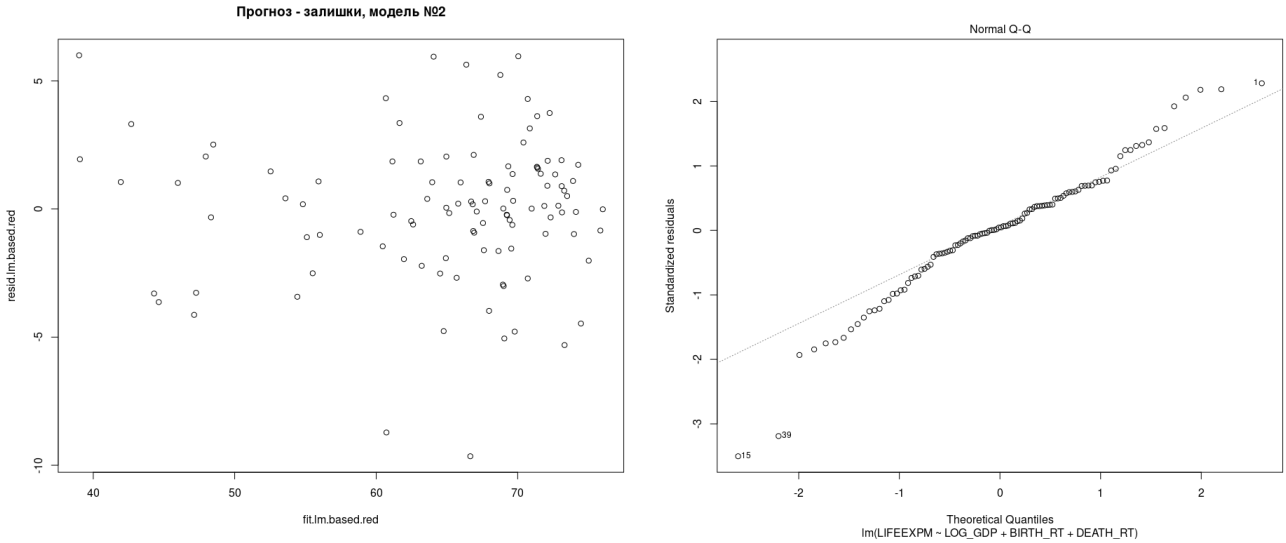

Рис. 6: Зліва – діаграма "Прогноз – залишки", справа – QQ-діаграма залишків.

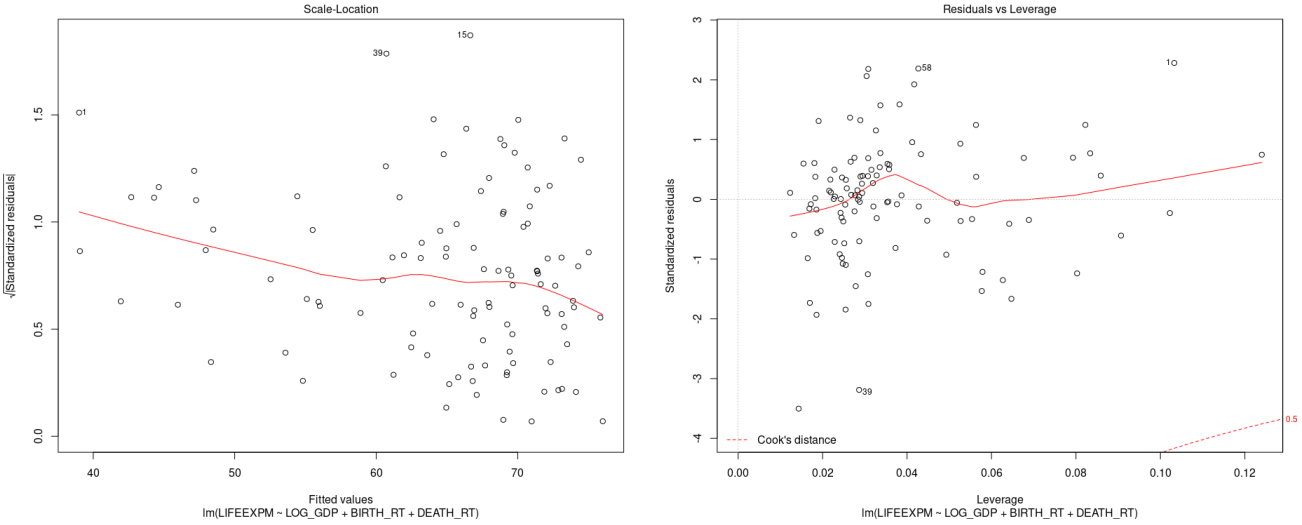

Рис. 7: Діаграма "Прогноз – стьюдентизовані залишки" зліва, та "Важіль – стьюдентизовані залишки" справа.

Більшість діаграм має приблизно таку ж саму природу, що й спостерігалося для першої моделі, але на QQ-діаграмі якось не зовсім чисто. Тільки відійшовши від "центральних" квантилів, маємо помітні відхилення точок у різні боки, що свідчить про неузгодженість залишків з нормальним розподілом. А це вже розхитує висунуту гіпотезу про гауссовість похибок. Аномалію також "відчули" базові тести з gvlma та безпосередньо тест Шапіро-Вілка:

```
> shapiro.test(resid(lm.based.red))
        Shapiro-Wilk normality test
data:  resid(lm.based.red)
W = 0.96603, p-value = 0.007331
...
```

```
...
> gvlma(lm.based.red)

Call:
lm(formula = LIFEEXPM ~ LOG_GDP + BIRTH_RT + DEATH_RT, data = dat)

Coefficients:
(Intercept)      LOG_GDP     BIRTH_RT     DEATH_RT
    65.9979       4.7414      -0.2903      -1.0256

ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = lm.based.red)

                     Value   p-value                   Decision
Global Stat         18.694 0.0009027 Assumptions NOT satisfied!
Skewness             4.698 0.0301921 Assumptions NOT satisfied!
Kurtosis             9.123 0.0025247 Assumptions NOT satisfied!
Link Function        3.359 0.0668332    Assumptions acceptable.
Heteroscedasticity   1.514 0.2185988    Assumptions acceptable.
```

З іншого боку бачимо, що підібрана функція зв'язку в другій моделі вважається коректно підібраною. Порушення припущення про нормальну розподіленість похибок призводить до того, що використання параметричних тестів типу тесу Стьюдента для коефіцієнтів чи тесту Фішера для перевірки загальних лінійних гіпотез є недоцільним. Хоча дисперсія оцінок коефіцієнтів зменшилася, однак це скоріше пояснюється зменшенням розмірності вектора невідомих параметрів у моделі.

Покращити першу модель (тобто забезпечити статистичнц значущість змінної LIT_MALE), можна замінити у формулі функції зв'язку для першої моделі замінити LIT_MALE на квадрат від нього, тобто $(\text{LIT\_MALE})^2$. Коротко кажучи, результати несуттєво зміняться, а вже на рівні значущості $\alpha = 0.05$ усі коефіцієнти "визнаються" значущо ненульовими.

Розшарування моделі.

Тепер перевіримо, чи відрізняється зв'язок у різних регіонах світу. Тобто у моделі вигляду:

$$\text{LIFEEXPM}_j = \sum_{k=1}^{6} \mathbb{1}\{\text{REGION\_KOD} = k\} \cdot P_{k,j} + \varepsilon_j,$$
$$P_{k,j} = \beta_0^k + \beta_1^k \cdot \text{LIT\_MALE}_j + \beta_2^k \cdot \text{LOG\_GDP}_j + \beta_3^k \cdot \text{BIRTH\_RT}_j + \beta_4^k \cdot \text{DEATH\_RT}_j \tag{2}$$

Потрібно перевірити гіпотези вигляду:

$$\mathbf{H}_0 : \beta_j^t = \beta_j^s,\ 1 \le t, s \le 6,\ j = \overline{0,4}$$
$$\mathbf{H}_1 : \text{Хоча б одна з рівностей з } \mathbf{H}_0 \text{ порушується}$$

Для цього застосуємо тест Фішера для перевірки загальної лінійної гіпотези.

Статистика і поріг тесту обчислюються за формулами:

$$F_{emp} = \frac{\frac{1}{p}(||U_{\mathbf{H}_0}||^2 - ||U_{\mathbf{H}_1}||^2)}{\frac{1}{n-d}||U_{\mathbf{H}_1}||^2},\ F_{theor} = Q^{F(p,n-d)}(1-\alpha),\ \alpha := 0.05$$

Тут $U_{\mathbf{H}_0}$, $U_{\mathbf{H}_1}$ – залишки прогнозу в моделях (1), (2) відповідно, $p = 6$ – кількість лінійних обмежень, $d = 30$ – кількість невідомих параметрів, що необхідно оцінити в (2), α – рівень значущості тесту. Підрахунок статистики тесту і порогу реалізується зовсім просто:

```
# Визначаємо ключові регіони та їхню кількість
regions <- unique(dat$REGION)
p <- length(regions)

# Підрахунок залишків в необмеженій моделі (за виконання H1)
lm.h1.subsets <- list()
for(j in 1:p)
{
  lm.h1.subsets[[j]] <- lm(LIFEEXPM ~ LIT_MALE + LOG_GDP + BIRTH_RT + DEATH_RT,
                           data = subset(dat, REGION == regions[j]))
}
resid.lm.h1 <- unlist(lapply(lm.h1.subsets, function(model) { resid(model) }))

# Підрахунок статистики тесту Фішера
alpha <- 0.05
params.h1 <- 5
# Поріг тесту
F.theor <- qf(1 - alpha, p, n - params.h1 * p)
# Сума квадратів залишків в обмеженій моделі (за виконання H0)
Sh0 <- sum(resid(lm.based)^2)
# Сума квадратів залишків в необмеженій моделі
Sh1 <- sum(resid.lm.h1^2)
# Статистика тесту
F.emp <- ((1 / p) * (Sh0 - Sh1)) / ((1 / (n - params.h1 * p)) * Sh1)
```

Виявилося, що значення статистики тесту перевищує заданий поріг:

```
> F.emp
[1] 10.2784
> F.theor
[1] 2.215694
```

А тому маємо підстави прийняти гіпотезу про залежність зв'язку від конкретного регіону. Однак до цього можна було б прийти з інтуїтивних міркувань (різний клімат у різних регіонах, переважний рівень розвитку країн, рівень життя населення у них тощо). Нижче покажемо отримані коефіцієнти при заданих змінних на кожній з підмножин:

```
> lapply(lm.h1.subsets, function(model) coef(model))
[[1]]
(Intercept)    LIT_MALE     LOG_GDP    BIRTH_RT    DEATH_RT
58.01612849  0.05454108  3.79289901 -0.21922736 -0.59524790

[[2]]
(Intercept)    LIT_MALE     LOG_GDP    BIRTH_RT    DEATH_RT
 56.8605508   0.1575747   3.3021249  -0.1705759  -1.3650516

[[3]]
(Intercept)    LIT_MALE     LOG_GDP    BIRTH_RT    DEATH_RT
62.09143053  0.05575176  3.09464546 -0.11715999 -1.02408929

[[4]]
(Intercept)    LIT_MALE     LOG_GDP    BIRTH_RT    DEATH_RT
44.80136855  0.38476080 -1.77735359 -0.15773819  0.09288977

[[5]]
(Intercept)    LIT_MALE     LOG_GDP    BIRTH_RT    DEATH_RT
290.4690131  -2.1529401  -4.4823121   0.7554346  -0.2652952

[[6]]
(Intercept)    LIT_MALE     LOG_GDP    BIRTH_RT    DEATH_RT
75.19757470 -0.03119226  0.36251635 -0.15150245 -1.18466248
```

А на десерт залишили пророблення аналогічних кроків для побудови моделі лінійної залежності середньої тривалості життя жінок від аналогічних показників. Спробуємо пройтися більш стисло, демонструючи в основному конкретні результати.

Для моделі вигляду:

$$\text{LIFEEXPF}_j = \beta_0 + \beta_1 \cdot \text{LIT\_FEMA}_j + \beta_2 \cdot \text{LOG\_GDP}_j + \beta_3 \cdot \text{BIRTH\_RT}_j + \beta_4 \cdot \text{DEATH\_RT}_j + \varepsilon_j$$

Маємо звіт:

```
> summary(lm.based)

Call:
lm(formula = LIFEEXPF ~ LIT_FEMA + LOG_GDP + BIRTH_RT + DEATH_RT,
    data = dat)
Residuals:
    Min      1Q  Median      3Q     Max
-7.6429 -1.3647  0.4462  1.8363  4.9576
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 65.15360    4.03900  16.131  < 2e-16 ***
LIT_FEMA     0.07543    0.01835   4.111 9.47e-05 ***
LOG_GDP      4.53936    0.73585   6.169 2.66e-08 ***
BIRTH_RT    -0.26559    0.04822  -5.507 4.29e-07 ***
DEATH_RT    -0.98473    0.07218 -13.643  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.57 on 80 degrees of freedom
  (24 observations deleted due to missingness)
Multiple R-squared:  0.9453,        Adjusted R-squared:  0.9426
F-statistic: 345.7 on 4 and 80 DF,  p-value: < 2.2e-16
```

```
> gvlma(lm.based)

Call:
lm(formula = LIFEEXPF ~ LIT_FEMA + LOG_GDP + BIRTH_RT + DEATH_RT,
    data = dat)
Coefficients:
(Intercept)     LIT_FEMA      LOG_GDP     BIRTH_RT     DEATH_RT
   65.15360      0.07543      4.53936     -0.26559     -0.98473
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05
Call:
 gvlma(x = lm.based)
                     Value  p-value                   Decision
Global Stat        10.0470 0.039643 Assumptions NOT satisfied!
Skewness            7.4486 0.006349 Assumptions NOT satisfied!
Kurtosis            0.1688 0.681208    Assumptions acceptable.
Link Function       1.7407 0.187052    Assumptions acceptable.
Heteroscedasticity  0.6890 0.406510    Assumptions acceptable.
```

Покажемо діаграми:

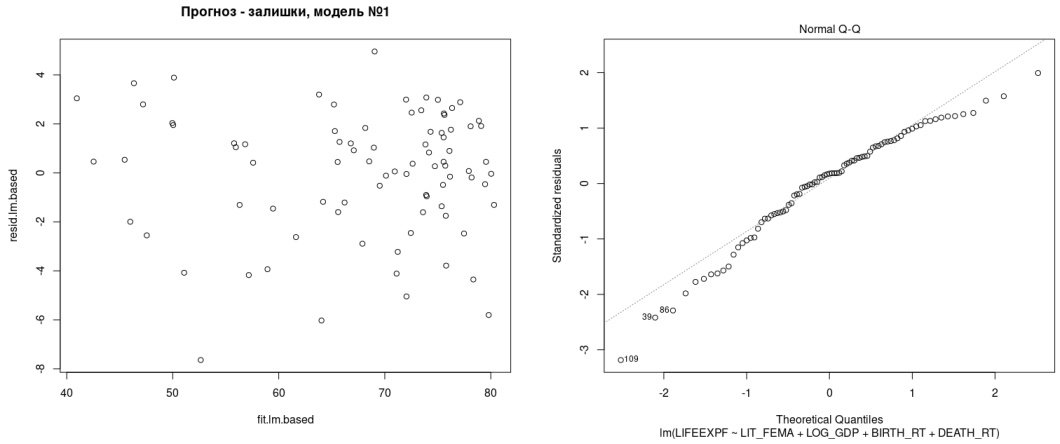

Рис. 8: Зліва – діаграма "Прогноз – залишки", справа – QQ-діаграма залишків.

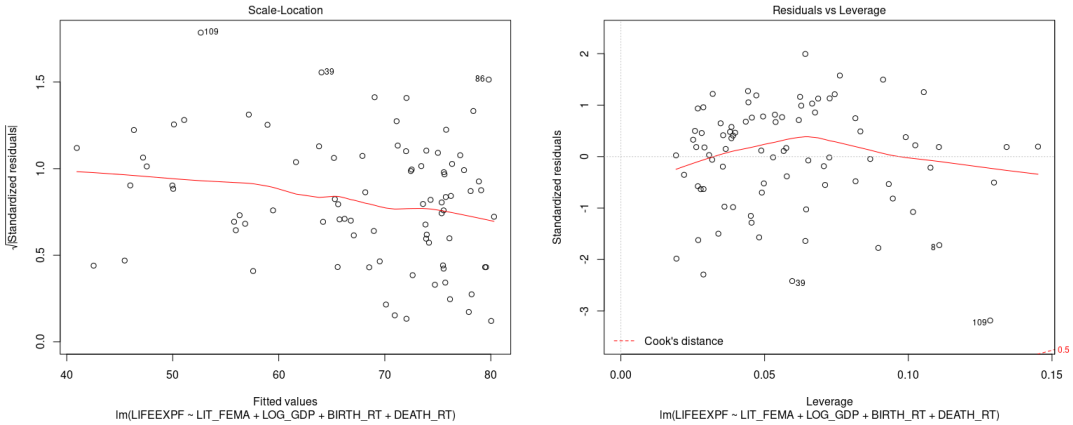

Рис. 9: Діаграма "Прогноз – стьюдентизовані залишки" зліва, та "Важіль – стьюдентизовані залишки" справа.

Тести Шапіро-Вілка, Бройша-Пагана, Дарбіна-Уотсона:

```
> shapiro.test(resid.lm.based)
        Shapiro-Wilk normality test
data:  resid.lm.based
W = 0.96009, p-value = 0.01004

> ncvTest(lm.based)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.631623, Df = 1, p = 0.10475

> durbinWatsonTest(lm.based)
 lag Autocorrelation D-W Statistic p-value
   1     -0.03442446      1.940745   0.714
 Alternative hypothesis: rho != 0
```

Для моделі вигляду:

$$\text{LIFEEXPF}_j = \beta_0 + \beta_1 \cdot \text{LOG\_GDP}_j + \beta_2 \cdot \text{BIRTH\_RT}_j + \beta_3 \cdot \text{DEATH\_RT}_j + \varepsilon_j$$

Маємо звіт:

```
> summary(lm.based.red)

Call:
lm(formula = LIFEEXPF ~ LOG_GDP + BIRTH_RT + DEATH_RT, data = dat)
Residuals:
    Min      1Q  Median      3Q     Max
-7.9047 -1.3077  0.3927  1.6121  6.6032
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.24346    2.97040  24.658  < 2e-16 ***
LOG_GDP      5.02850    0.63474   7.922 2.72e-12 ***
BIRTH_RT    -0.41703    0.03137 -13.295  < 2e-16 ***
DEATH_RT    -0.99396    0.06413 -15.500  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 2.562 on 104 degrees of freedom
  (1 observation deleted due to missingness)
Multiple R-squared:  0.9431,        Adjusted R-squared:  0.9415
F-statistic: 574.9 on 3 and 104 DF,  p-value: < 2.2e-16
```

```
> gvlma(lm.based.red)

Call:
lm(formula = LIFEEXPF ~ LOG_GDP + BIRTH_RT + DEATH_RT, data = dat)
Coefficients:
(Intercept)      LOG_GDP     BIRTH_RT     DEATH_RT
     73.243        5.028       -0.417       -0.994
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05
Call:
 gvlma(x = lm.based.red)
                     Value p-value                   Decision
Global Stat        7.74495 0.10138    Assumptions acceptable.
Skewness           6.01661 0.01417 Assumptions NOT satisfied!
Kurtosis           1.17827 0.27771    Assumptions acceptable.
Link Function      0.53933 0.46271    Assumptions acceptable.
Heteroscedasticity 0.01074 0.91746    Assumptions acceptable.
```

Покажемо діаграми:

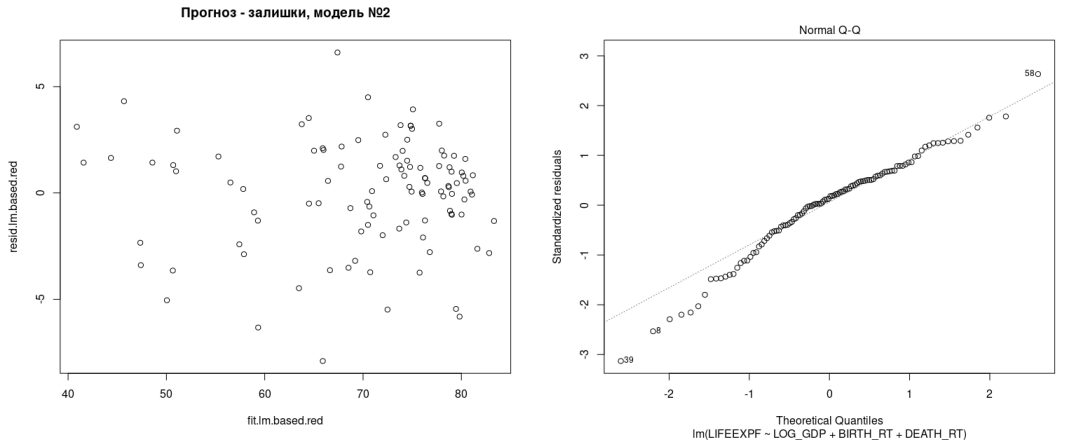

Рис. 10: Зліва – діаграма "Прогноз – залишки", справа – QQ-діаграма залишків.

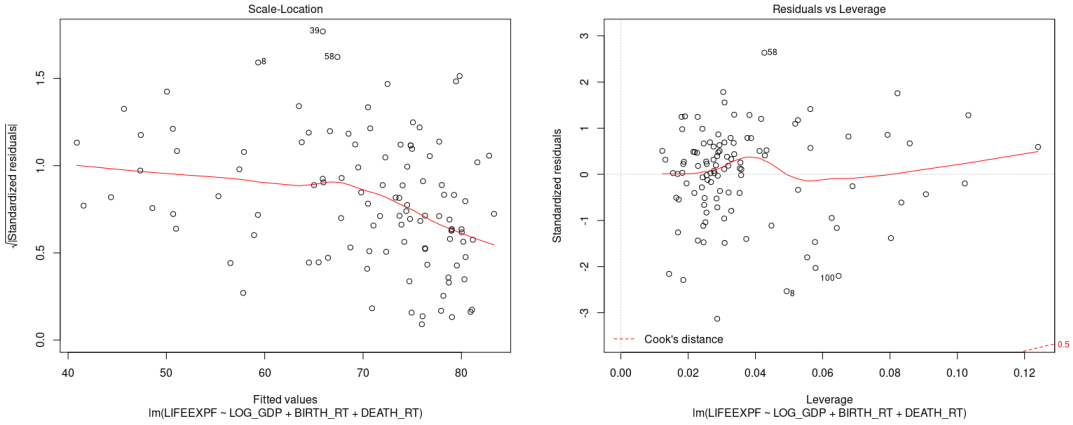

Рис. 11: Діаграма "Прогноз – стьюдентизовані залишки" зліва, та "Важіль – стьюдентизовані залишки" справа.

Тести Шапіро-Вілка, Бройша-Пагана, Дарбіна-Уотсона:

```
> shapiro.test(resid.lm.based.red)
        Shapiro-Wilk normality test
data:  resid.lm.based.red
W = 0.97234, p-value = 0.0237

> ncvTest(lm.based.red)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.393629, Df = 1, p = 0.036073

> durbinWatsonTest(lm.based.red)
 lag Autocorrelation D-W Statistic p-value
   1     0.002387594      1.943749   0.716
 Alternative hypothesis: rho != 0
```

Як виявилося, то похибки попередніх моделей гірше імітують гауссовість на жіночій статі. Це можна виправити, задіявши нелінійні перетворення на відповідні змінні (як-от, піднести до квадрату LIT_FEMA). Закриваючи очі на проблеми із гауссовістю, маємо з питання про розшарування за регіоном:

```
> F.emp
[1] 13.68445
> F.theor
[1] 2.215694
```

Тобто приймається гіпотеза про наявність розшарування в моделі в залежності від регіону. Коефіцієнти за кожною підгрупою такі:

```
> lapply(lm.h1.subsets, function(model) coef(model))
[[1]]
(Intercept)    LIT_FEMA     LOG_GDP    BIRTH_RT    DEATH_RT
57.38223798  0.07636572  5.30577677 -0.18396967 -0.78432406

[[2]]
(Intercept)    LIT_FEMA     LOG_GDP    BIRTH_RT    DEATH_RT
 55.6524527   0.1265939   6.0278845  -0.1528127  -1.2650643

[[3]]
(Intercept)    LIT_FEMA     LOG_GDP    BIRTH_RT    DEATH_RT
 66.3272916   0.1076122   2.2993594  -0.1100718  -1.0363399

[[4]]
(Intercept)    LIT_FEMA     LOG_GDP    BIRTH_RT    DEATH_RT
 76.2245770   0.1152447   2.1466254  -0.7274422  -0.7497726

[[5]]
 (Intercept)     LIT_FEMA      LOG_GDP     BIRTH_RT     DEATH_RT
141.11692925  -0.68123346  -0.89184029   0.37823835   0.04280875

[[6]]
(Intercept)    LIT_FEMA     LOG_GDP    BIRTH_RT    DEATH_RT
92.27956350 -0.05279824 -1.00351957 -0.41316597 -1.11230012
```

Висновки.

За даними можна побудувати нескладну лінійну модель залежності тривалості життя від доступних нами показників, причому її результати є відносно непоганими. Розглянутий аналіз та побудова моделі не виявився повноцінним: потрібно було спробувати моделі із врахуванням інших змінних, які мають природній вплив на досліджувану величину. Можливо, тоді можна було б отримати кращі моделі.