Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського»

Кафедра Математичних Методів Системного Аналізу

Розрахункова работа з дисципліни "Математическая статистика"

Перевірив
к.ф.м.н. Каніовская І. Ю.

Виконав
студент Панченко Є. С.

Київ 2021

Зміст

1. Задача 1

1.1. Постановка задачі

x	10	15	25	35	45	55	65	75	85	95
y	2600	2100	1300	1000	820	670	580	510	490	470

Основна мета - побудувати регресійну модель та зробити її аналіз.

1.2. Аналіз вибірки та вибір лінійної регресійної моделі

Для початку побудуємо точки на площині XOY.

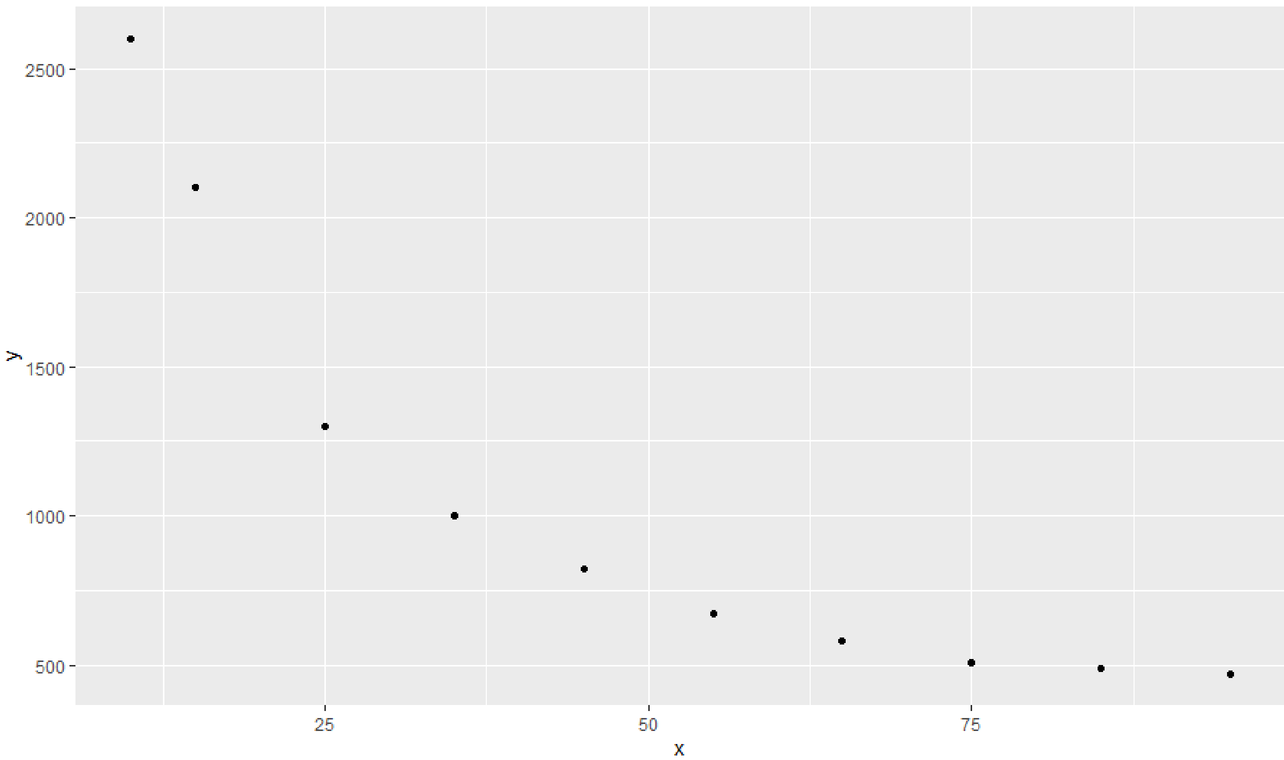

Рис. 1. Точки на площині

Помітимо, що це нагадує графік гіперболи, а тому зобразімо точки з координатами $(x, \frac{1}{y})$.

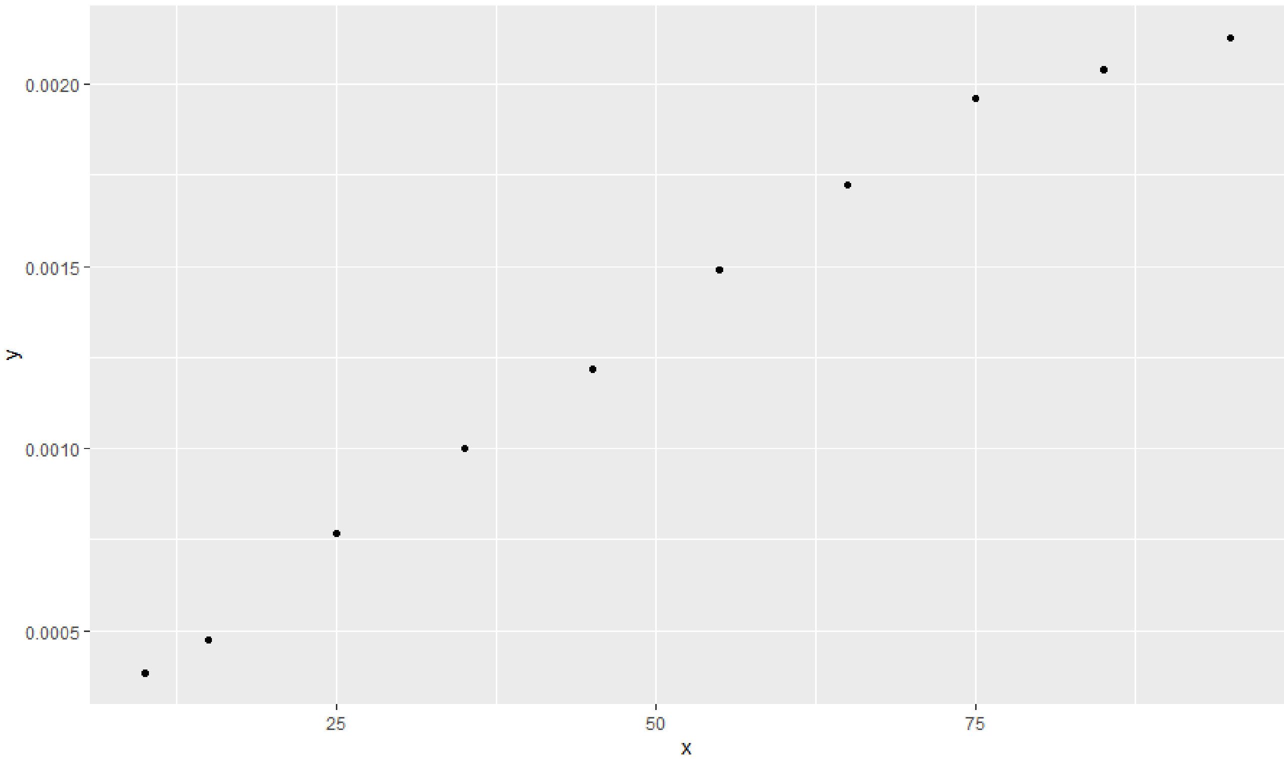

Рис. 2. Точки на площині, де координати y інверсовані

Бачимо, що точки розташовані майже на прямій. Саме тому можна обрати вигляд функції, яку оцінює регресійна модель, як

$$f(x) = \frac{1}{\beta_0 + \beta_1 x}.$$

Але цю модель можна спростити. Давайте будемо оцінювати функцію

$$g(x) = \frac{1}{f(x)}.$$

Тоді

$$g(x) = \beta_0 + \beta_1 x.$$

Одразу позначатимемо $\vec{\eta}^{(f)}$ - відклик або вихідна величина. Також введемо позначення $\eta_i^{(g)} = \frac{1}{\eta_i^{(f)}}$. Ці позначення дозволяють нам тимчасово забути про існування функції f. Тобто можна працювати з функцією g.

1.3. Знаходження оцінок параметрів за методом найменших квадратів

Метод найменших квадратів полягає у знаходженні таких значень параметрів β_0 та β_1, щоб мінімізувати значення

$$\sum_{i=1}^{10}(\eta_i^{(g)} - g(x_i))^2.$$

Відомо, що оцінкою методом найменших квадратів параметрів лінійної регресії є вектор

$$\vec{\beta}^* = (F^T F)^{-1} F^T \vec{\eta}^{(g)}, \qquad \text{де } F \text{ - матриця плану.}$$

В умовах нашої задачі

$$F = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 15 & 25 & 35 & 45 & 55 & 65 & 75 & 85 & 95 \end{pmatrix}^T$$

і

$$\vec{\eta}^{(g)} = \begin{pmatrix} 0.384 & 0.476 & 0.769 & 1 & 1.219 & 1.492 & 1.724 & 1.96 & 2.04 & 2.127 \end{pmatrix} \cdot 10^{-3}.$$

Виконавши всі розрахунки, отримаємо, що

$$\vec{\beta}^*_{val} = \begin{pmatrix} 2.185826 \\ 0.2180424 \end{pmatrix} \cdot 10^{-4}.$$

Отже, ми отримали оцінку параметрів регресійної моделі. Зобразімо на другому рисунку пряму, яку задають значення оцінок параметрів β_0 і β_1.

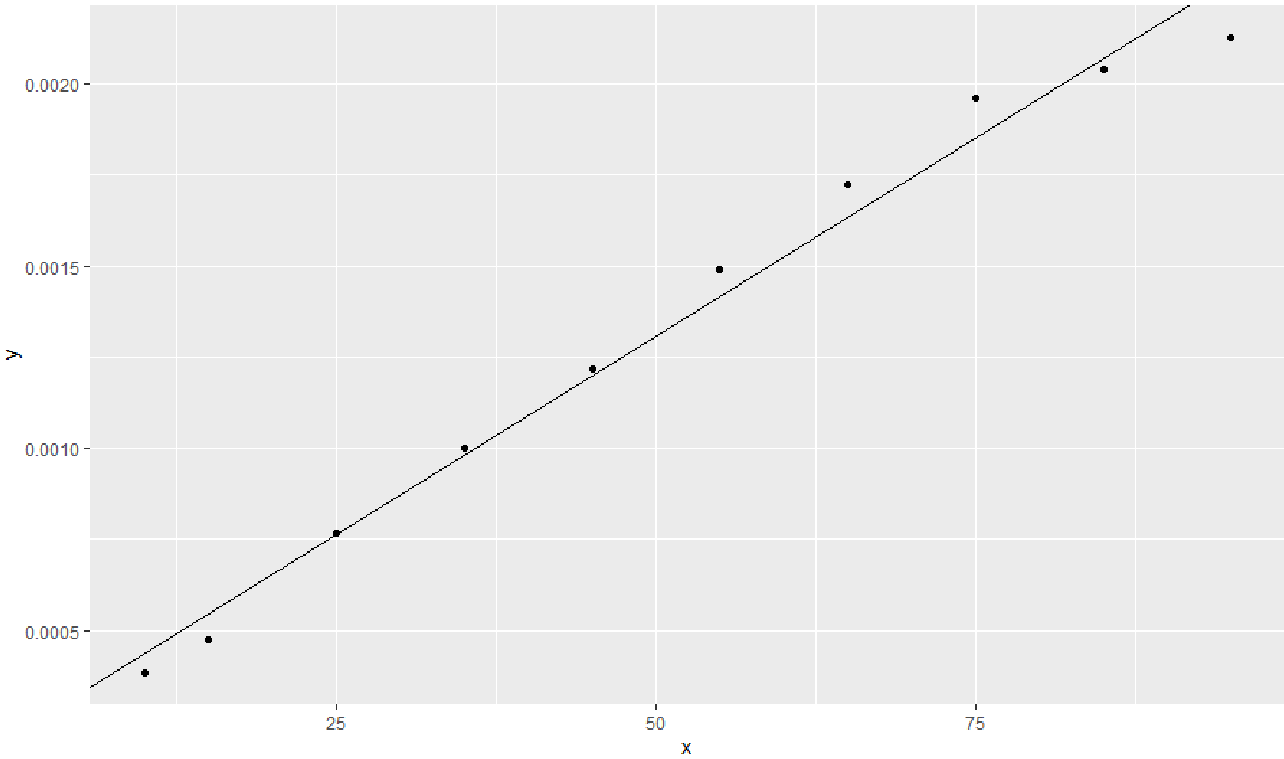

Рис. 3. Точки на площині та пряма

1.4. Перевірка адекватності побудованої моделі

Висунемо нульову гіпотезу, що константа та побудована модель не відрізняються. Альтернативною оберемо гіпотезу, що побудована модель краща за константу.

Для перевірки адекватності побудованої моделі скористаємося F-критерієм - ми хочемо порівняти залишкову оцінку дисперсії з незміщеною оцінкою диспресії. Відомо, що статистика

$$\zeta = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(\eta_i^{(g)} - \overline{\eta^{(g)}})^2}{\frac{1}{n-m}\sum_{i=1}^{n}(\eta_i^{(g)} - g^*(x_i))^2} \sim F(n-1, n-m), \quad \text{де } m \text{ - кількість параметрів.}$$

Вирахуємо значення, якими будемо користуватися згодом:

$$(\mathbb{D}^{**}\eta^{(g)})_{val} = \frac{1}{9}\sum_{i=1}^{10}(y_i^{(g)} - \overline{y^{(g)}})^2 = 4.198214 \cdot 10^{-7}$$

і

$$(\sigma^2_{(g)})^{**}_{val} = \frac{1}{8}\sum_{i=1}^{10}(y_i^{(g)} - g^*(x_i))^2 = 7.550273 \cdot 10^{-9}$$

і

$$A^{-1} = F^T F = \begin{pmatrix} 0.426 & -0.0064 \\ -0.0064 & 0.00012 \end{pmatrix}.$$

Надалі елементи матриці A^{-1} позначатимемо маленькими літерами a з індексами, якщо не вказано інше. Вирахуємо значення статистики.

$$\zeta_{val} = \frac{(\mathbb{D}^{**}\eta^{(g)})_{val}}{(\sigma^2_{(g)})^{**}_{val}} = \; = \frac{4.198214 \cdot 10^{-7}}{7.550273 \cdot 10^{-9}} = 55.603.$$

На рівні значущості $\alpha = 0.05$ маємо $t_{cr} = 3.39$. Оскільки критична область правостороння і $\zeta_{val} > t_{cr}$, то нульову гіпотезу відхиляємо.

Отже, побудовану модель можна вважати адекватною.

1.5. Перевірка гіпотези про значущість найменшого значення параметра побудованої моделі

Оскільки ми з'ясували, що модель можна вважати адекватною, то перевіримо на значущість параметр β_0. Для цього висунемо нульову гіпотезу $H_0 : \beta_0 = 0$. Альтернативна гіпотеза $H_1 : \beta_0 > 0$.

Відомо, що статистика

$$\gamma^{(g)} = \frac{\beta_0^*}{\sqrt{(\sigma^2_{(g)})^{**} \cdot a_{00}}} \sim St_{n-m}.$$

Вирахуємо значення статистики.

$$\gamma^{(g)}_{val} = \frac{2.185826 \cdot 10^{-4}}{\sqrt{7.550273 \cdot 10^{-9} \cdot 0.426}} = 3.854.$$

На рівні значущості $\alpha = 0.05$ маємо $t_{cr} = 1.86$. Оскільки критична область правостороння і $\zeta_{val} > t_{cr}$, то нульову гіпотезу відхиляємо.

Перевіримо на значущість і параметр β_1.

Для цього висунемо нульову гіпотезу $H_0 : \beta_1 = 0$. Альтернативна гіпотеза $H_1 : \beta_1 > 0$.

Відомо, що статистика

$$\gamma^{(g)} = \frac{\beta_1^*}{\sqrt{(\sigma^2_{(g)})^{**} \cdot a_{11}}} \sim St_{n-m}.$$

Вирахуємо значення статистики.

$$\gamma^{(g)}_{val} = \frac{0.2180424 \cdot 10^{-4}}{\sqrt{7.550273 \cdot 10^{-9} \cdot 0.0001}} = 25.09.$$

На рівні значущості $\alpha = 0.05$ маємо $t_{cr} = 1.86$. Оскільки критична область правостороння і $\zeta_{val} > t_{cr}$, то нульову гіпотезу відхиляємо.

Отже, наша модель є адекватною і зменшити кількість параметрів не вдалося.

1.6. Побудова прогнозованого довірчого інтервала для середнього значення відклику та самого значення відклику

Повернімося до функції f.

$$f^*(x) = \frac{1}{g^*(x)} = \frac{1}{\beta_0^* + \beta_1^* x}.$$

Вирахуємо залишкову оцінку диспресії для f:

$$(\sigma^2_{(f)})^{**}_{val} = \frac{1}{8}\sum_{i=1}^{10}(y_i^{(f)} - f^*(x_i))^2 = 21512.43.$$

Будемо будувати обидва довірчих інтервала для точки $\vec{x} = \begin{pmatrix} 1 \\ 50 \end{pmatrix}$.

Знайдемо довірчий інтервал для середнього значення відклику. Відомо, що статистика

$$\frac{f^*(x) - f(x)}{\vec{x}^\intercal A^{-1} \vec{x}} \sim St_{n-m}.$$

Тоді довірчий інтервал для середнього значення відклику має вигляд

$$f(x) \in \left(f^*(x) - t\sqrt{(\sigma^2_{(f)})^{**}_{val}\vec{x}^\intercal A^{-1}\vec{x}}, f^*(x) + t\sqrt{(\sigma^2_{(f)})^{**}_{val}\vec{x}^\intercal A^{-1}\vec{x}}\right).$$

Вирахуємо

$$\begin{aligned}
&\vec{x}^\intercal A^{-1}\vec{x} = 0.1,\\
&\sqrt{(\sigma^2_{(f)})^{**}_{val}\vec{x}^\intercal A^{-1}\vec{x}} = 46.38,\\
&f^*(50) = \frac{1}{\beta_0^* + \beta_1^* \cdot 50} = 764.1475.
\end{aligned}$$

При рівні надійності $\gamma = 0.95$ маємо $t = t_{cr} = 2.306$. Підставлючи усі знайдені значення маємо, що

$$f(x) \in (764.1475 - 2.306 \cdot 46.38, 764.1475 - 2.306 \cdot 46.38) \Leftrightarrow f(x) \in (657.19, 871.1).$$

Знайдемо довірчий інтервал для самого значення відклику. Відомо, що статистика

$$\frac{\eta - f^*(x)}{(\sigma^2_{(f)})^{**}(1 + \vec{x}^\intercal A^{-1}\vec{x})} \sim St_{n-m}.$$

Тому довірчий інтервал для середнього значення відклику має вигляд

$$\eta \in \left(f^*(x) - t\sqrt{(\sigma^2_{(f)})^{**}_{val}(1 + \vec{x}^\intercal A^{-1}\vec{x})}, f^*(x) + t\sqrt{(\sigma^2_{(f)})^{**}_{val}(1 + \vec{x}^\intercal A^{-1}\vec{x})}\right).$$

Акуратно підставивши значення отримаємо

$$\eta \in (764.1475 - 2.306 \cdot 153.83, 764.1475 - 2.306 \cdot 153.83) \Leftrightarrow \eta \in (409.42, 1118.88).$$

1.7. Висновок

На момент аналізу даних у мене були думки щодо двух моделей. Після деяких спроб модифікації даних я отримав, що якщо значення y замінити на обернені величини, то точки на графіку будут знаходитися майже на одній прямій. Таким чином можна було розглядати обернену функцію і будувати лінійну регресійну модель відштовхуючись від цього.

Після цих думок з'явилася думка, що можна задати вигляд функції f як $f = \beta_0 + \beta_1 \cdot \frac{1}{x}$. Перевагою цією моделі від першої є те, що тут не потрібно переходити між функціями. Але я побудував графіки обох оцінок функцій - і мені здалося, що перша модель більш точно виражає поведінку даних. Саме тому модель виду $f(x) = \frac{1}{\beta_0 + \beta_1 x}$ була обрана для оцінки.

2. Задача 2

2.1. Постановка задачі

2.2. Пошук оцінок параметрів двофакторної регресійної моделі за методом найменших квадратів

2.3. Перевірка адекватності побудованої моделі

2.4. Перевірка гіпотези про значущість найменшого значення параметра побудованої моделі

2.5. Побудова прогнозованого довірчого інтервала для середнього значення відклику та самого значення відклику

2.6. Висновок

Література