Описательная часть расчетной работы №2 по математической статистике за 6 триместр

Вихляев Егор, ММТ-2

June 27, 2023

1 Задание №1

Для выбранного набора данных построить корреляционное поле.

Для построения корреляционного поля, определимся с исходными данными. В левом столбце имеем независимые переменные x_i, в правом столбце – зависимые переменные y_i. Строим поле по двум представленным столбцам через точечную диаграмму в Excel (все вычисления в расчетном файле).
Там же добавим и уравнение детерминированной части регрессии (в Excel оно называется «линия тренда»):

$$Y = 2.0814 \cdot X + 8.8277.$$

Обозначим график уравнения красной штрихованной линией.

2 Задание №2

Построить коэффициенты корреляции Пирсона и Спирмена. Проверить значимость коэффициентов при уровне значимости $\alpha = 0.1$. Сделать выводы о наличии корреляционной зависимости между переменными и характере их корреляционной зависимости, исходя из вычисленных значений коэффициентов.

1. Построить коэффициенты корреляции Пирсона и Спирмена.
   Коэффициент Пирсона считается по следующей формуле и равен:

$$r^*_{xy} = \frac{\overline{x \cdot y} - \overline{x} \cdot \overline{y}}{\sqrt{S_x^2 \cdot S_y^2}} \approx 0.531.$$

   В Excel его можно вычислить через функцию =ПИРСОН(X;Y).
   Ранговый коэффициент корреляции Спирмена считается по следующей формуле и равен:

$$r_s = 1 - \frac{6}{n \cdot (n^2 - 1)} \cdot \sum_{i=1}^{n} (rang(X_i) - rang(Y_i))^2 \approx 0.339.$$

   В Excel его можно вычислить либо по формуле, приведенной выше, расписав ее элементы по столбцам, либо найти ранги элементов, а затем применить к ним функцию =КОРРЕЛ(диап. рангов X; диап. рангов Y). Для ясности процесса, в расчетной работе мы привели первый вариант решения.

2. Проверить значимость коэффициентов при уровне значимости $\alpha = 0.1$.

   (a) Начнем с коэффициента Пирсона.
   Выдвинем следующую нулевую гипотезу:

$$H_0 : r^*_{xy} = 0.$$

   Для проверки значимости коэффициента Пирсона, потребуется сравнить по модулю две статистики, а именно:

$$t_{\text{набл}} = \frac{\sqrt{n-2} \cdot r^*_{xy}}{\sqrt{1 - (r^*_{xy})^2}} = \frac{\sqrt{50-2} \cdot 0.531}{\sqrt{1 - 0.531^2}} \approx 4.342,$$

$$t_{\text{кр}} = x_{1-\frac{\alpha}{2}}[St_{n-2}] = x_{0.95}[St_{48}] = 1.677.$$

   Исходя из вычисленных статистик, имеем следующее неравенство:

$$|t_{\text{набл}}| > t_{\text{кр}} \Rightarrow$$

   $\Rightarrow H_0$ – не принимается $\Rightarrow$ коэффициент Пирсона r^*_{xy} – значим!

(b) Проверим на значимость ранговый коэффициент Спирмена.

$$H_0 : r_s = 0.$$

Находим необходимые статистики:

$$t_{\text{набл}} = \frac{\sqrt{n-2} \cdot r_s}{\sqrt{1-r_s^2}} = \frac{\sqrt{50-2} \cdot 0.339}{\sqrt{1-0.339^2}} \approx 2.496,$$

$$t_{\text{кр}} = x_{1-\frac{\alpha}{2}}[St_{n-2}] = x_{0.95}[St_{48}] = 1.677.$$

Исходя из вычисленных статистик, имеем следующее неравенство:

$$|t_{\text{набл}}| > t_{\text{кр}} \Rightarrow$$

$\Rightarrow H_0$ – не принимается $\Rightarrow$ ранговый коэффициент Спирмена r_s – значим!

3. Сделать выводы о наличии корреляционной зависимости между переменными и характере их корреляционной зависимости, исходя из вычисленных значений коэффициентов.
Исходя из вычисленных значений коэффициентов, можно сделать следующие выводы:

   (a) Согласно коэффициенту Пирсона $0.3 < r^*_{xy} \approx 0.531 < 0.7$, переменные X_i и Y_i линейно зависимы, связь средней тесноты.

   (b) Согласно ранговому коэффициенту Спирмена $0.3 < r_s* \approx 0.339 < 0.7$, переменные X_i и Y_i линейно зависимы, связь средней тесноты.

   Как следствие слабой зависимости, мы не можем явно сказать, что имеется прямая («чем больше, тем больше») или обратная («чем больше, тем меньше») зависимости. Это потверждает и корреляционное поле из задания №1.

3 Задание №3

Записать линейную регрессионную модель. Выписать оценки неизвестных параметров модели.

Линейная регрессионная модель имеет вид:

$$Y = a + b \cdot X + \epsilon,$$

где Y – отклик (зависимая переменная), X – фактор (независимая переменная), ϵ – случайная компонента (суммарная ошибка). Из задания №1 мы знаем, что наша детерминированная часть регрессионой модели имеет вид:

$$Y = 2.0814 \cdot X + 8.8277.$$

Сделаем оценки $\hat{a}$ и $\hat{b}$ неизвестных параметров a и b с помощью метода наименьших квадратов (МНК), используя заведомо выведенные формулы:

$$\hat{a} = \overline{Y} - \hat{b} \cdot \overline{X} = 9.33 - 2.0398 * 0.24 \approx 8.8376,$$

$$\hat{b} = \frac{\overline{X \cdot Y} - \overline{X} \cdot \overline{Y}}{S_x^2} = \frac{12.177 - 0.24 * 9.33}{74.99} \approx 2.0398.$$

Итак, оценки неизвестных параметров a и b: $\hat{a} \approx 8.8376$, $\hat{b} \approx 2.0398$. Как можем наблюдать, наши оценки неизвестных параметров примерно совпадают со значениями неизвестных параметров, выведенных с помощью Excel в задании №1: $a = 2.0814, b = 8.8277$.

4 Задание №4

Для нелинейных моделей $y = a + b \cdot x^2 + \epsilon, y = a + \frac{b}{x} + \epsilon$ найти МНК-оценки коэффициентов и коэффициент детерминации. Выбрать лучшую из нелинейных моделей и выписать ее. Сравнить выбранную нелинейную модель с линейной моделью

1. Найдем МНК-оценки для первой модели $y = a + b \cdot x^2 + \epsilon$.

$$W = \sum_{i=1}^{n} (y_i - a - b \cdot x_i^2)^2 \to min.$$

$$\frac{\partial W}{\partial a} = \sum_{i=1}^{n} 2 \cdot (y_i - a - b \cdot x_i^2) \cdot (-1) = 0,$$

$$\frac{\partial W}{\partial b} = \sum_{i=1}^{n} 2 \cdot (y_i - a - b \cdot x_i^2) \cdot (-x_i^2) = 0$$

$$\begin{cases} \sum_{i=1}^{n} y_i - a \cdot n - b \sum_{i=1}^{n} x_i^2 = 0 \ / : n \\ \sum_{i=1}^{n} y_i \cdot x_i^2 - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i^4 = 0 \ / : n \end{cases}$$

$$\begin{cases} \overline{y_i} - a - b \cdot \overline{x^2} = 0 \\ \overline{x^2 \cdot y} - a \cdot \overline{x^2} - b \cdot \overline{x^4} = 0 \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} \hat{a} = \overline{y} - b \cdot \overline{x^2}, \\ \hat{b} = \frac{\overline{x^2 \cdot y} - \overline{y} \cdot \overline{x^2}}{\overline{x^4} - (\overline{x^2})^2} \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} \hat{a} = 2.6562, \\ \hat{b} = 1.3804 \end{cases}$$

Коэффициент детерминации

$$R_1^2 = 0.793.$$

2. Найдем МНК-оценки для второй модели $y = a + \frac{b}{x} + \epsilon$.

$$W = \sum_{i=1}^{n} (y_i - a - \frac{b}{x_i})^2 \to min.$$

$$\frac{\partial W}{\partial a} = \sum_{i=1}^{n} 2 \cdot (y_i - a - \frac{b}{x_i}) \cdot (-1) = 0, / : (-2)$$

$$\frac{\partial W}{\partial b} = \sum_{i=1}^{n} 2 \cdot (y_i - a - \frac{b}{x_i}) \cdot (-\frac{1}{x_i}) = 0 / : (-2)$$

$$\begin{cases} \sum_{i=1}^{n} y_i - a \cdot n - b \sum_{i=1}^{n} \frac{1}{x_i} = 0 \ / : n \\ \sum_{i=1}^{n} \frac{y_i}{x_i} - a \sum_{i=1}^{n} \frac{1}{x_i} - b \sum_{i=1}^{n} \frac{1}{x_i^2} = 0 \ / : n \end{cases}$$

$$\begin{cases} \overline{y} - a - b \cdot \overline{(\frac{1}{x})} = 0 \\ \overline{(\frac{y}{x})} - a \cdot \overline{(\frac{1}{x})} - b \cdot \overline{(\frac{1}{x^2})} = 0 \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} \hat{a} = \overline{y} - b \cdot \overline{(\frac{1}{x})}, \\ \hat{b} = \frac{\overline{\frac{y}{x}} - \overline{\frac{1}{x}} \cdot \overline{y}}{\overline{\frac{1}{x^2}} - (\overline{\frac{1}{x}})^2} \end{cases} \Rightarrow$$

$$\Rightarrow \begin{cases} \hat{a} = 9.3479, \\ \hat{b} = 0.1386 \end{cases}$$

Коэффициент детерминации

$$R_2^2 = 0.002.$$

Сравнивая две нелинейных модели по коэффициенту детерминации, можно явно сказать, что первая нелинейная модель $y = a + b \cdot x^2 + \epsilon = 2.6562 + 1.3804 \cdot x^2 + \epsilon$ лучше, чем вторая.

В то же время, сравнивая первую нелинейную модель с линейной моделью, основываясь на соответствующих коэффициентах детерминации, можно сказать, что первая нелинейная модель $y = a + b \cdot x^2 + \epsilon = 2.6562 + 1.3804 \cdot x^2 + \epsilon$ все же лучше, чем линейная, поскольку ее коэффициент детерминации $R_1^2 = 0.793$ значительно больше, чем коэффициент детерминации линейной модели $R_0^2 = 0.531$