# Задача на построение модели линейной регрессии

### Постановка задачи

Пусть нам заданы значения входного признака $X = \{x_1, \ldots, x_{10}\}$ и значения выходного признака $Y = \{y_1, \ldots, y_{10}\}$

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| $x_i$ | 1.29 | 4.51 | 2.69 | 1.80 | 4.28 | 3.83 | 3.51 | 2.30 | 4.80 | 4.60 |
| $y_i$ | 1.11 | 3.53 | 2.42 | 1.99 | 3.18 | 3.31 | 2.90 | 2.27 | 4.09 | 3.47 |

Необходимо проверить, есть ли линейная зависимость между этими признаками. Если линейная зависимость присутствует, то построить модель линейной регрессии, проверить адекватность построенной модели, статистическую значимость коэффициентов и самой модели. Предполагается возможность оценки корреляции по Пирсону. Вычислить с помощью регрессионной модели значение выходного признака при $x = 5.1$.

### Решение задачи

Чтобы выяснить наличие линейной зависимости между $X$ и $Y$, нам нужно вычислить значение коэффициента корреляции. Оценивать значение коэффициента мы будем по Пирсону, то есть

$$r_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y},$$

где стандартная ошибка среднего

$$\sigma_{xy} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \overline{xy} - \overline{x} \cdot \overline{y},$$

$$\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x^2} - (\overline{x})^2,$$

$$\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \overline{y^2} - (\overline{y})^2,$$

а среднее

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i,$$

$$\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2,\ \overline{y^2} = \frac{1}{n}\sum_{i=1}^{n} y_i^2.$$

Таким образом, чтобы вычислить значение коэффициента корреляции Пирсона, нам нужно вычислить все 5 значений, которые определены по формулам выше. Начнем с вычисления средних значений:

$$\overline{x} = \frac{1}{10}(1.29 + 4.51 + 2.69 + 1.80 + 4.28 + 3.83 + 3.51 + 2.30 + 4.80 + 4.60) = 3.361,$$

$$\overline{y} = \frac{1}{10}(1.11 + 3.53 + 2.42 + 1.99 + 3.18 + 3.31 + 2.90 + 2.27 + 4.09 + 3.47) = 2.827,$$

$$\overline{xy} = \frac{1}{10}(1.29 \cdot 1.11 + \ldots + 4.60 \cdot 3.47) = 10.472.$$

$$\overline{x^2} = \frac{1}{10}(1.29^2 + \ldots + 4.60^2) = 12.728,$$

$$\overline{y^2} = \frac{1}{10}(1.11^2 + \ldots + 3.47^2) = 8.691,$$

Теперь вычислим стандартные ошибки

$$\sigma_x^2 = 12.728 - 3.361^2 \approx 1.431,$$

$$\sigma_y^2 = 8.691 - 2.827^2 \approx 0.699,$$

$$\sigma_{xy} = 10.472 - 3.361 \cdot 2.827 = 0.971.$$

Теперь мы можем вычислить коэффициент парной корреляции

$$r_{xy} = \frac{0.971}{\sqrt{1.431} \cdot \sqrt{0.699}} = 0.971.$$

Таким образом, мы имеем сильную прямую (положительную) линейную зависимость. Следовательно, имеет смысл построить модель линейной регрессии. Модель двумерной линейной регрессии имеет вид

$$y = ax + b,\ a = \frac{\sigma_{xy}}{\sigma_x^2},\ b = \overline{y} - a\overline{x}.$$

Таким образом, подставляя в эти формулы известные нам значения, получим

$$a = \frac{0.971}{1.431} = 0.679,$$

$$b = 2.827 - 0.679 \cdot 3.361 = 0.484.$$

Таким образом, модель линейной регрессии имеет вид

$$y = f(x) = 0.679x + 0.484.$$

Проверим адекватность построенной линейной модели. Чтобы модель была адекватна, необходимо, чтобы

$$\sigma_\varepsilon = \sqrt{\sigma_y^2(1 - r_{xy}^2)} < 0.67\sigma_y.$$

Таким образом, подставляя известные значения, получим

$$\sqrt{0.699 \cdot (1 - 0.971^2)} < 0.67 \cdot \sqrt{0.699},$$

$$0.199 < 0.560,$$

что верно. Следовательно, модель можно считать адекватной. Также адекватность модели нам позволяет определить коэффициент детерминации

$$R^2 = r_{xy}^2,$$

а именно если $R^2 > 0.7$, то линейную модель можно считать адекватной. Вычислим коэффициент детерминации

$$R^2 = 0.971^2 = 0.942 > 0.7,$$

то есть в силу этого мы также можем считать модель адекватной.

Теперь проверим статистическую значимость коэффициентов построенной линейной модели. Для этого вычислим стандартное отклонение параметров линейной модели

$$\sigma_a = \frac{\sigma_\varepsilon}{\sigma_x\sqrt{n-2}} = \frac{0.199}{\sqrt{1.431}\sqrt{8}} = 0.059,$$

$$\sigma_b = \frac{\sigma_\varepsilon}{\sqrt{n-2}}\sqrt{1+\frac{\overline{x^2}}{\sigma_x^2}} = \frac{0.199}{\sqrt{8}}\cdot\sqrt{1+\frac{12.728}{1.431}} = 0.221.$$

Определим пороговый уровень значимости $\alpha = 0.05$. Тогда оценим значимость коэффициентов при заданном уровне значимости $\alpha$:

$$T_a = \frac{a}{\sigma_a} = 11.5,$$

$$T_b = \frac{b}{\sigma_b} = 2.19.$$

По таблице $t$-распределения Стъюдента можно найти

$$t(n-2,\alpha) = t(8, 0.05) = 2.306.$$

Тогда, сравнивая полученные статистики с табличным значением, получим

$$|T_a| = 11.5 > 2.306 = t(8, 0.05),$$

то есть коэффициент $a$ является статистически значимым.

$$|T_b| = 2.19 < 2.306 = t(8, 0.05),$$

то есть коэффициент $b$ не является статистически значимым. Таким образом, отбросив из модели коэффициент $b = 0.484$ адекватность модели не сильно ухудшится, а может даже и улучшится.

Оценим статистическую значимость уравнения регрессии с помощью $F$-критерия Фишера. $F$-статистика вычисляется по формуле

$$F = \frac{(n-2)\overline{\delta}^2}{\overline{D}},$$

где остаточная (необъясненная) дисперсия, которая характеризует отклонение от выбранной модели регрессии,

$$\overline{D} = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2,$$

а объясненная дисперсия, при которой вариация обусловлена уравнением регрессии,

$$\overline{\delta}^2 = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \overline{y})^2,$$

где $f(x_i)$ – это значения модели, найденные при известных $x_i$ по уравнению регрессии. Вычислим значение $F$-статистики

$$\overline{D} = \frac{(1.11 - 0.679\cdot 1.29 - 0.484)^2 + \ldots + (3.47 - 0.679\cdot 4.60 - 0.484)^2}{10} \approx 0.251,$$

$$\overline{\delta}^2 = \frac{(0.679 \cdot 1.29 + 0.484 - 2.827)^2 + \ldots + (0.679 \cdot 4.60 + 0.484 - 2.827)^2}{10} \approx 0.501,$$

отсюда

$$F = \frac{8 \cdot 0.501}{0.251} = 15.968.$$

По статистическим таблицам критических значений $F$-критерия Фишера, по заданному уровню значимости $\alpha = 0.05$ и степеням свободы $df_1 = 1,\ df_2 = n-2$, найдем критическую точку

$$F(1, 8, 0.05) = 5.7.$$

Следовательно, поскольку

$$F = 15.968 > 5.7 = F(1, 8, 0.05),$$

то построенная регрессионная модель является статистически значимой.

Остается лишь найти неизвестное значение $y$ при $x = 5.1$. Подставляя это значение $x$ в регрессионную модель, получим

$$0.679 \cdot 5.1 + 0.484 = 0.830.$$