Лекция № 18

Парная линейная регрессия

Определение уравнения линейной регрессии

Как уже говорилось ранее, первым шагом при анализе зависимости между признаками является построение диаграммы рассеяния, которая позволяет качественно оценить наличие линейной (прямой или обратной) зависимости, криволинейной зависимости или отсутствие зависимости. На втором этапе рассчитывается коэффициент корреляции и проверяется его значимость. Если коэффициент значимый, то на третьем этапа переходят к построению модели (уравнения) регрессии, которое в линейном случае представляет собой прямую, наилучшим образом приближающую данные на диаграмме рассеяния. Основным назначением уравнения регрессии является выявление имеющейся в данных тенденции и прогнозирование - определение значения зависимой переменной для тех значений независимой, при которых наблюдения не проводились.

Наиболее распространенным способом построения уравнения регрессии является метод наименьших квадратов (МНК). Рассмотрим результаты наблюдений двух признаков (x_i, y_i), $i = 1, \ldots, n$ и предположим, что уравнение парной линейной регрессии $y' = a + bx$ построено. Величина y' обозначает расчетные значения зависимого признака, отличающиеся от истинных значений y. Тогда для каждой точки x_i можно из уравнения определить расчетное (прогнозируемое) значение $y_i' = a + bx_i$ и найти величину остатка $e_i = y_i - y_i' = y_i - (a + bx_i)$, рис. 1. Метод наименьших квадратов расчета уравнения линейной регрессии основан на условии минимизации суммы квадратов остатков:

$$S = y_i - y_i' = \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2 \to \min .$$

Дальнейшее решение сводится к задаче на экстремум, то есть определению, при каких значениях a и b функция двух переменных S достигнет своего минимума. Как известно из курса математического анализа, для этого нужно найти частные производные S по a и b, приравнять их к нулю и после элементарных преобразований решить систему двух уравнений с двумя неизвестными.

В соответствии с этим подходом найдем частные производные:

$$\begin{cases} \dfrac{\partial S}{\partial a} = 2\sum\limits_{i=1}^{n} \left(y_i - a - bx_i \right)(-1) = 0 \\ \dfrac{\partial S}{\partial b} = 2\sum\limits_{i=1}^{n} \left(y_i - a - bx_i \right)(-x_i) = 0 \end{cases}$$

Сократив каждое уравнение на -2, раскрыв скобки и перенеся члены с x_i; в одну сторону, а с y_i - в другую, получим:

$$\begin{cases} na + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i \\ a\sum\limits_{i=1}^{n} x_i + b\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} y_i x_i \end{cases}$$

Эта система называется системой нормальных уравнений МНК для линейного уравнения регрессии. Из нее можно в явном виде выписать соотношения для коэффициентов a и b:

$$a = \frac{\sum\limits_{i=1}^{n} y_i \sum\limits_{i=1}^{n} x_i^2 - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} x_i y_i}{n\left(\sum\limits_{i=1}^{n} x_i^2 \right) - \left(\sum\limits_{i=1}^{n} x_i \right)^2}, \; b = \frac{n\left(\sum\limits_{i=1}^{n} x_i y_i \right) - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\left(\sum\limits_{i=1}^{n} x_i^2 \right) - \left(\sum\limits_{i=1}^{n} x_i \right)^2}. \quad (*)$$

Пользуясь данными выражениями, можно вычислить коэффициенты

уравнения регрессии, зная размер выборки n и выборочные значения признаков x_i и y_i.

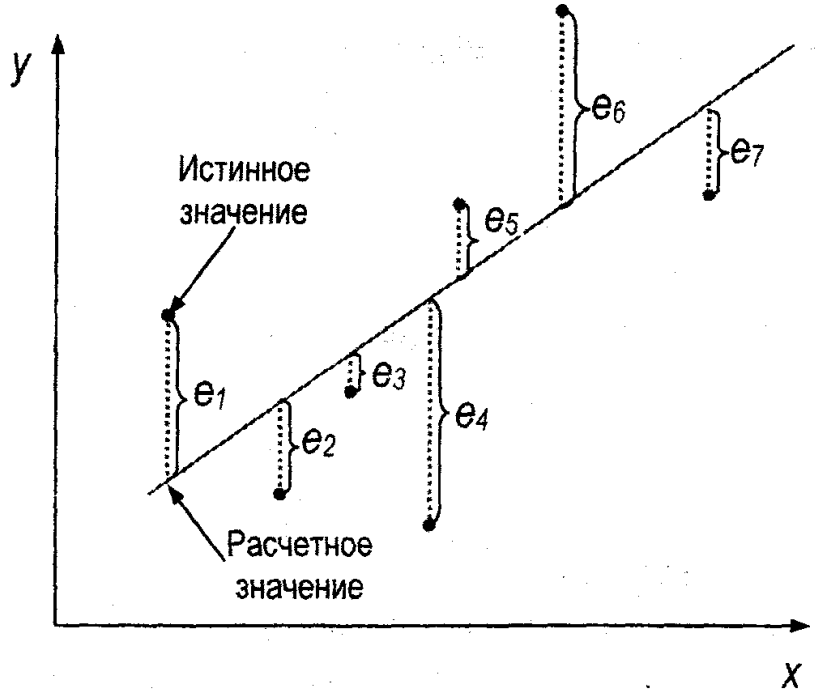

Рис. 1. *Линия регрессии, соотношение истинных, расчетных значений и остатков*

Пример. Построить уравнение регрессии для зависимости величин возраста и давления, табл. 1.

Решение. Для определения уравнения регрессии необходимо по данным наблюдений вычислить значения 4 сумм. Суммы x_i и y_i находятся непосредственно по столбцам 1 и 2, а для расчета сумм квадратов $\sum_{i=1}^{n} x_i^2$ и парных произведений $\sum_{i=1}^{n} x_i y_i$ удобно ввести в таблице дополнительных колонки и поместить в них суммируемые величины x_i^2, $x_i y_i$.

Таблица 1. *Расчет коэффициентов уравнения регресии*

Возраст x_i	Давление y_i	$x_i y_i$	x_i^2
43	128	5504	1849
48	120	5760	2304
56	135	7560	3136
61	143	8723	3721
67	141	9447	4489
70	152	10640	4900
$\sum_{i=1}^{6} x_i = 345$	$\sum_{i=1}^{6} y_i = 819$	$\sum_{i=1}^{6} x_i y_i = 47634$	$\sum_{i=1}^{6} x_i^2 = 20399$

Далее по формулам (*) находятся значения a и b:

$$a = \frac{819 \cdot 20399 - 345 \cdot 47634}{6 \cdot 20399 - (345)^2} = 81.048,\ b = \frac{6 \cdot 47634 - 345 \cdot 819}{6 \cdot 20399 - (345)^2} = 0.964.$$

Уравнение регрессии построено, рис. 2:

$$y' = a + bx = 81.048 + 0.964 \cdot x$$

Знак коэффициента b всегда совпадает со знаком коэффициента корреляции - он положительный для прямой зависимости между признаками и отрицательный для обратной. Коэффициенту может быть дана следующая интерпретация - он показывает, на сколько единиц изменится прогнозируемое значение результативного признака при изменении факторного признака на единицу. В данном случае $b = 0.964 \approx 1$, то есть давление повышается на единицу за год.

Модель линейной регрессии

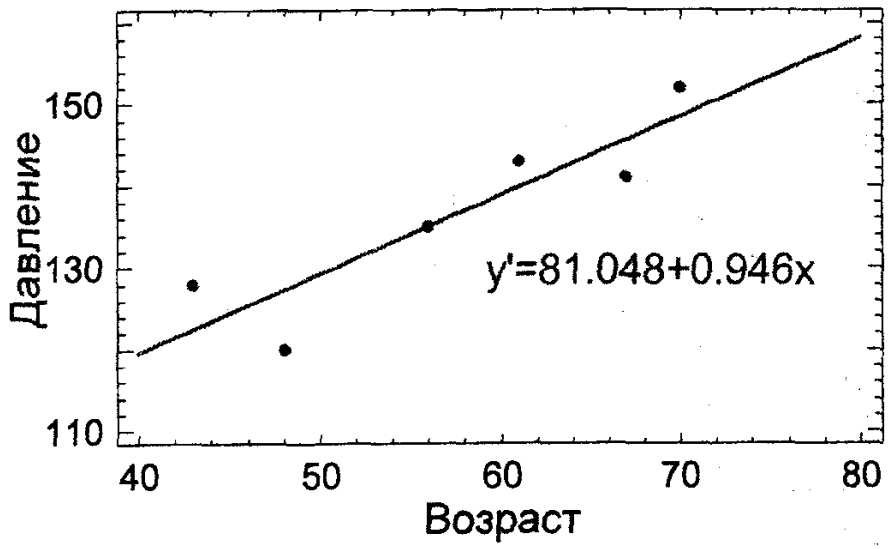

Рис. 2. *Уравнение линейной регрессии между возрастом и давлением*