Практическая работа №1

Регрессионный анализ.

Уравнение линейной парной регрессии.

Уравнение линейной парной регрессии выглядит следующим образом: $Y=a_0+a_1X$

При помощи этого уравнения переменная Y выражается через константу a_0 и угол наклона прямой (или угловой коэффициент) a_1, умноженный на значение переменной X. Константу a_0 также называют свободным членом, а угловой коэффициент - коэффициентом регрессии. Параметры уравнения могут быть определены с помощью метода наименьших квадратов (МНК)

Метод наименьших квадратов

(в справочных системах англоязычных программ - Least Squares Method, LS) является одним из основных методов определения параметров регрессионных уравнений, дающий наилучшие линейные несмещенные оценки. Линейные – относится к характеру взаимосвязи переменных. Несмещенные значит, что ожидаемые значения коэффициентов регрессии должны быть истинными коэффициентами. То есть точки, построенные по исходным данным (x_i, y_i), должны лежать как можно ближе к точкам линии регрессии. Сущность данного метода заключается в нахождении параметров модели, при которых сумма квадратов отклонений эмпирических (фактических) значений результирующего признака от теоретических, полученных по выбранному уравнению регрессии, то есть:

$$S = \sum_{i=1}^{n} \left(y_i^{\text{ð}} - y_i \right)^2 = \sum_{i}^{n} (y_i^p - a_0 - a_1 x)^2 \to \min,$$

где y_i^p – значение, вычисленное по уравнению регрессии; $\left(y_i^p - y_i\right)$ – **отклонение** ε (ошибка, остаток) (рис. 1); n – количество пар исходных данных.

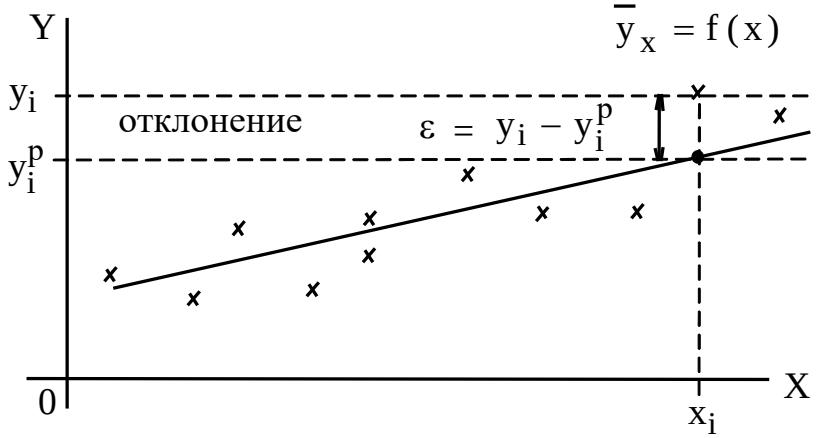

Рис. 1 Понятие отклонения ε для случая линейной регрессии

В регрессионном анализе предполагается, что математическое ожидание случайной величины ε равно нулю и ее дисперсия одинакова для всех наблюдаемых значений Y. Отсюда следует, что рассеяние данных возле линии регрессии должно быть одинаково при всех значениях параметра X. В случае, показанном на рис. 2 данные распределяются вдоль линии регрессии неравномерно, поэтому метод наименьших квадратов в этом случае неприменим.

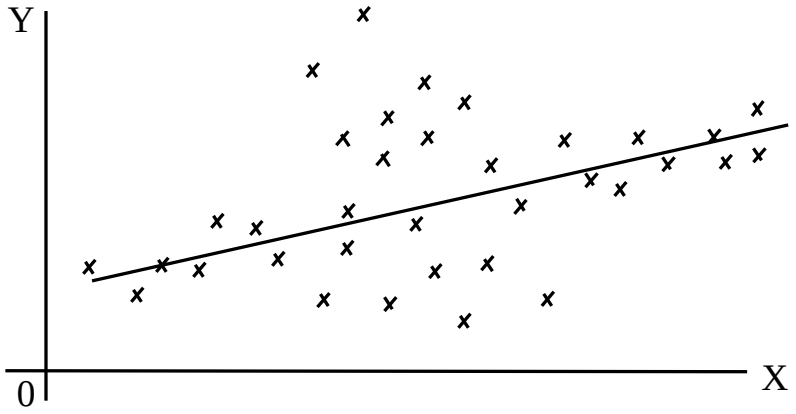

Рис.2. Неравномерное распределение исходных точек вдоль линии регрессии

Проведя необходимые преобразования, получим систему двух уравнений с двумя неизвестными a_0 и a_1, которые найдем решив систему.

$$a_1 = \frac{n\left(\sum y_i x_i\right) - \sum y_i \sum x_i}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}; \qquad (1)$$

$$a_0 = \frac{1}{n}\left(\sum y_i - a_1 \sum x_i\right) \qquad (2)$$

Направление связи между переменными определяется на основании знаков (отрицательный или положительный) коэффициента регрессии (коэффициента a_1).

Если знак при коэффициенте регрессии - положительный, связь зависимой переменной с независимой будет положительной. В нашем случае знак коэффициента регрессии положительный, следовательно, связь также является положительной.

Если знак при коэффициенте регрессии - отрицательный, связь зависимой переменной с независимой является отрицательной (обратной).

Для анализа общего качества уравнения уравнения регрессии используют обычно *множественный коэффициент детерминации* R^2, называемый также квадратом коэффициента множественной корреляции R. R^2 (мера определенности) всегда находится в пределах интервала [0;1].

Если значение R^2 близко к единице, это означает, что построенная модель объясняет почти всю изменчивость соответствующих переменных. И наоборот, значение R-квадрата, близкое к нулю, означает плохое качество построенной модели.

Коэффициент детерминации R^2 показывает, на сколько процентов ($R^2 \cdot 100\%$) найденная функция регрессии описывает связь между исходными значениями факторов X и Y

$$R^2 = \frac{\sum_{i=1}^{n}\left(y_i^{\delta} - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}$$

где $\left(y_i^p - \bar{y}\right)^2$ – объясненная вариация; $\left(y_i - \bar{y}\right)^2$ – общая вариация (рис.3).

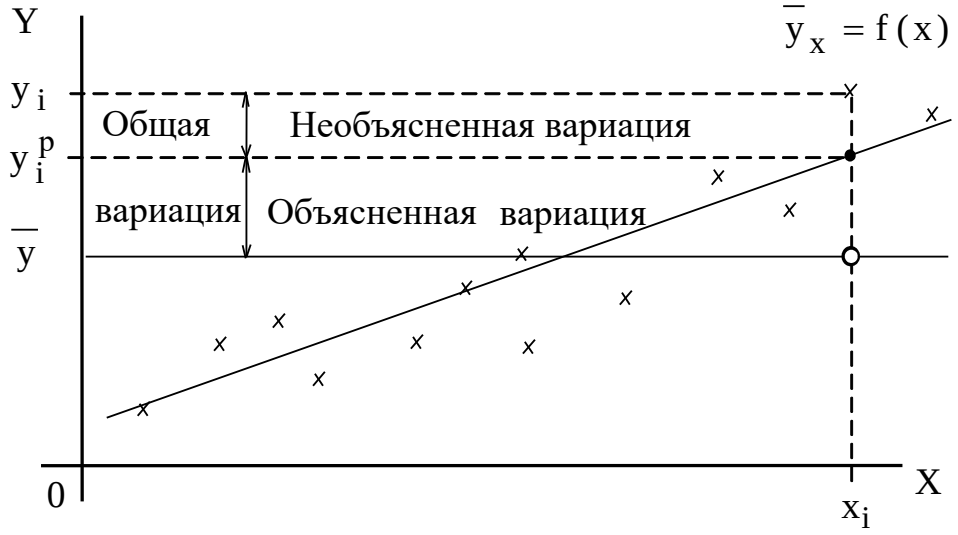

Рис. 3 Графическая интерпретация коэффициента детерминации для случая линейной регрессии Соответственно, величина $(1-R^2)\cdot 100\%$ показывает, сколько процентов вариации параметра Y обусловлены факторами, не включенными в регрессионную модель. При высоком ($R^2 \geq 75\%$) значении коэффициента детерминации можно делать прогноз $y^* = f(x^*)$ для конкретного значения x^*.

Нелинейная регрессия

Рассмотрим наиболее простые случаи *нелинейной* регрессии: гиперболу, экспоненту и параболу. При нахождении коэффициентов гиперболы и экспоненты используют прием приведения нелинейной регрессионной зависимости к линейному виду. Это позволяет использовать для вычисления коэффициентов функций регрессии выше приведенные формулы.

Гипербола. Для приведения уравнения вида $y = a_0 + \frac{a_1}{x}$ к линейному виду вводят новую переменную $z = \frac{1}{x}$, тогда уравнение гиперболы принимает линейный вид $y = a_0 + a_1 z$. После этого используют формулы (1) и (2) для нахождений линейной функции, но вместо значений x_i используются значения $z_i = \frac{1}{x_i}$:

$$a_1 = \frac{n\left(\sum y_i z_i\right) - \sum y_i \sum z_i}{n\left(\sum z_i^2\right) - \left(\sum z_i\right)^2}; \qquad a_0 = \frac{1}{n}\left(\sum y_i - a_1 \sum z_i\right). \qquad (3)$$

Экспонента. Для приведения к линейному виду уравнения **экспоненты** $y = a_0 e^{a_1 x}$ проведем логарифмирование:

$$\ln y = \ln\left(a_0 e^{a_1 x}\right);$$
$$\ln y = \ln a_0 + \ln\left(e^{a_1 x}\right);$$
$$\ln y = \ln a_0 + a_1 x.$$

Введем переменные $b_0 = \ln a_0$ и $b_1 = a_1$, тогда $\ln y = b_0 + b_1 x$, откуда следует, что можно применять формулы (1) и (2), в которых вместо значений y_i надо использовать $\ln y_i$:

$$b_1 = \frac{n\left(\sum [\ln y_i] x_i\right) - \sum \ln y_i \sum x_i}{n\left(\sum x_i^2\right) - \left(\sum x_i\right)^2}; \qquad b_0 = \frac{1}{n}\left(\sum \ln y_i - b_1 \sum x_i\right) \qquad (4)$$

При этом мы получим численные значения коэффициентов b_0 и b_1, от которых надо перейти к a_0 и a_1, используемых в модели экспоненты. Исходя из введенных обозначений и определения логарифма, получаем

$$a_0 = e^{b_0},\ a_1 = b_1.$$

Парабола. Для нахождения коэффициентов уравнения параболы $y = a_0 + a_1 x + a_2 x^2$ необходимо решить линейную систему из трех уравнений:

$$\begin{cases} n \cdot a_0 + \left(\sum x_i\right) a_1 + \left(\sum x_i^2\right) a_2 = \sum y_i, \\ \left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 + \left(\sum x_i^3\right) a_2 = \sum (y_i x_i), \\ \left(\sum x_i^2\right) a_0 + \left(\sum x_i^3\right) a_1 + \left(\sum x_i^4\right) a_2 = \sum \left(y_i x_i^2\right). \end{cases}$$

Сила регрессионной связи для гиперболы и параболы определяется непосредственно по той же формуле что и для линейной модели. При вычислении коэффициента детерминации для экспоненты все значения параметра Y (исходные, регрессионные, среднее) необходимо заменить на их логарифмы, например, y_i^p – на $\ln\left(y_i^p\right)$ и т.д.

Если функция регрессии определена, интерпретирована и обоснована, и оценка точности регрессионного анализа соответствует требованиям, можно считать, что построенная модель и прогнозные значения обладают достаточной надежностью.

Прогнозные значения, полученные таким способом, являются средними значениями, которые можно ожидать.

Методические рекомендации

Для проведения регрессионного анализа и прогнозирования необходимо:

1) ***построить график*** исходных данных и попытаться зрительно, приближенно определить характер зависимости;

2) ***выбрать вид функции*** регрессии, которая может описывать связь исходных данных;

3) ***определить численные коэффициенты*** функции регрессии методом наименьших квадратов;

4) ***оценить силу*** найденной регрессионной зависимости на основе коэффициента детерминации R^2;

5) ***сделать прогноз*** (при $R^2 \geq 75\%$) или сделать вывод о невозможности прогнозирования с помощью найденной регрессионной зависимости. При этом не рекомендуется использовать модель регрессии для тех значений независимого параметра X, которые не принадлежат интервалу, заданному в исходных данных.

Варианты задач

Задача №1

Постройте регрессионную модель (линейную) для исходных данных приведенных в таблице. Для облегчения расчетов исходные данные содержат только четыре пары значений (x_i, y_i).

Исходные данные задачи №1

№ варианта	Координаты	Точки				x*
1	X	1	2	3	4	1.6
	Y	30	7	8	1	?
2	X	1	2	3	4	2.3
	Y	25	7	7	2	
3	X	9	5	2	3	2.9
	Y	25	7	7	2	?
4	X	1	2	3	4	2.6
	Y	15	10	7	0.5	?
5	X	10	3	6	4	8
	Y	25	7	7	2	?
6	X	9	5	2	3	2.5
	Y	15	8.5	7.5	5	?
7	X	2	3	7	8	7.5
	Y	11	8.5	6.5	5	?
8	X	10	3	6	4	9
	Y	15	7	8	6	?
9	X	2	3	4	5	4.5
	Y	13	9	8	7	?
10	X	1	2	3	4	1.5
	Y	7.5	7	5	3.5	?
11	X	1	2	3	4	3.6
	Y	13	9	8	7	?
12	X	3	4	6	10	8
	Y	7.5	7	6.5	3.5	?
13	X	3	4	5	6	7.8
	Y	9	7	5	3	?
14	X	7	5.6	13	14.7	15
	Y	7.5	7	5	3.5	?
15	X	9	5	2	3	5.7
	Y	13	9	8	7	?
16	X	3	4	6	8	5
	Y	7.5	7	6.5	5	?
17	X	2	3	7	8	7.5
	Y	9	9	8	7	?
18	X	9	10	11	12	10.5
	Y	13	9	8	7	?

19	X	1	2	3	4	3.5
	Y	5	4.5	3	3	?
20	X	11	12	13	16	13.6
	Y	7.6	8	6.5	4.2	?
21	X	5	6	7	8	6.5
	Y	5	4.5	3	3	?
22	X	9	10	12	14	12.5
	Y	8	7	6.5	4.2	?
23	X	7	8	9	10	9.6
	Y	8	7	6	4.2	?
24	X	1.5	2.5	3.5	4.5	3.9
	Y	5	4.5	3	3	?
25	X	1	2	5	6	3.9
	Y	5	4	3	3	?
26	X	1.5	2.4	3.8	6.9	4.1
	Y	5.5	5.5	4.8	1.1	?
27	X	1	2	3	4	3.6
	Y	12	3	9	5	?
28	X	1	2	3	7	2.8
	Y	5	5.5	4.8	1.1	?
29	X	11	12	13	16	14.1
	Y	0.25	0.19	5.2	8	?
30	X	1	2	3	4	3.4
	Y	13	4	10	6	?

Задача № 2

Для исходных данных, представленных в таблице, были построены следующие регрессионные модели:

- $y = 6{,}067 - 0{,}085x$;
- $y = -2{,}017 + 3{,}957x - 0{,}367x^2$;
- $y = 5{,}918e^{-0{,}043x}$.

Исходные данные задачи №2

X	3	8	5	10	7	6	4	9	1	2
Y	6	5	9	1	8	9	8	4	2	4

С помощью графика отклонений выберите удовлетворительную модель и проверьте свой выбор с помощью соответствующего расчета.

Задача №3

В таблице представлены данные о ценах на комплектующие для ПЭВМ. Комплектующие производятся различными компаниями-производителями и разбиты на группы по своим функциональным возможностям.

Исходные данные задачи №3

Группа	1	1	2	2	2	3	3	3	4	4
Цена, $	50	60	70	80	95	100	115	120	105	120
Группа	4	5	5	5	6	6	6	7	7	7

Цена, $	130	110	150	190	120	130	220	145	265	270

Постройте график исходных данных и с его помощью проанализируйте применимость метода наименьших квадратов. Подтвердите свои выводы с помощью расчета (для линейной модели). Прокомментируйте экономические причины полученного результата.