Отчет
по лабораторной работе №8
по Анализу данных и основам Data Science
по теме: «Регрессионный анализ: модели и методы»

ИВТ 1.1.
Выполнил:
Тихонов А.С.

2024

Лабораторная работа №8.

Регрессионный анализ.

Цель работы: изучить методы регрессионного анализа и метод наименьших кевадратов, рассмотреть их использование на примерах конкретных задач.
Оборудование: MS Excel.

Задача 1.

Постановка задачи: имеются выборочные данные о стоимости квартир в их общей площади в некотором городе. Необходимо выполнить поставленные задачи.
Решение:
Начальные данные:

t	1	2	3	4	5	6	7	8	9	10	11	12
x	8,453	7,666	5,047	3,628	3,464	2,434	2,905	1,167	2,142	2,028	2,992	0,715

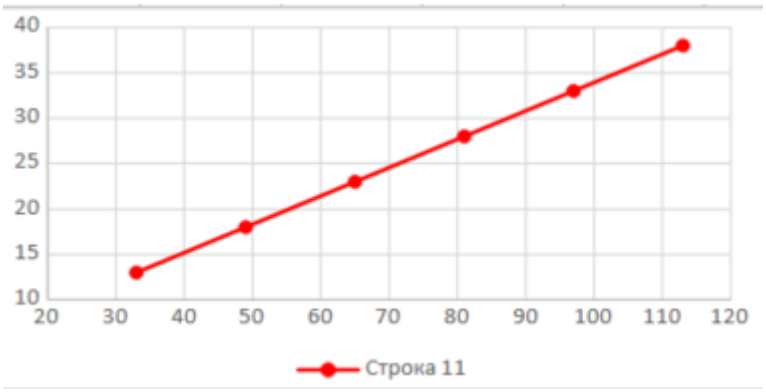
Задание 1. Построить график зависимости между переменными, по кеоторому необхходимо подобрать модель регрессии.
Расчет интервалов для x и y:

h_x	15,8	≈	16
h_y	4,82	≈	5

Получившиеся интервалы:

Таблица интервалов:						
x	33	49	65	81	97	113
y	13	18	23	28	33	38

График:



Промежуточный анализ: характер расположения точек на графике показывает, что связь между переменными может выражаться линейной зависимостью, то есть y’ = b0 + b1x

Задание 2. Рассчитать параметры уравнения регрессии методом наименьших квадратов.
Решение:
Таблица промежуточных вычислений:

i	x	y	x^2	y^2	x*y	y'	y - y'	(y-y')^2	A
1	33	13,8	1089	190,44	455,4	14,7341	-0,9341	0,8726	6,77%
2	40	13,8	1600	190,44	552	16,8468	-3,0468	9,2830	22,08%
3	36	14	1296	196	504	15,6396	-1,6396	2,6882	11,71%
4	60	22,5	3600	506,25	1350	22,8831	-0,3831	0,1467	1,70%
5	55	24	3025	576	1320	21,3740	2,6260	6,8959	10,94%
6	80	28	6400	784	2240	28,9193	-0,9193	0,8451	3,28%
7	95	32	9025	1024	3040	33,4465	-1,4465	2,0924	4,52%
8	70	20,9	4900	436,81	1463	25,9012	-5,0012	25,0119	23,93%
9	48	22	2304	484	1056	19,2613	2,7387	7,5004	12,45%
10	53	21,5	2809	462,25	1139,5	20,7704	0,7296	0,5324	3,39%
11	95	32	9025	1024	3040	33,4465	-1,4465	2,0924	4,52%
12	75	35	5625	1225	2625	27,4103	7,5897	57,6043	21,68%
13	63	24	3969	576	1512	23,7885	0,2115	0,0447	0,88%
14	112	37,9	12544	1436,41	4244,8	38,5773	-0,6773	0,4588	1,79%
15	70	27,5	4900	756,25	1925	25,9012	1,5988	2,5562	5,81%
Итого	985	368,9	72111	9867,85	26466,7	368,9	4,619E-14	118,62492	135,47%
Среднее	65,667	24,593	4807,400	657,857	1764,447	24,593	0,000	7,908	9,03%

$$b_1 = \frac{\overline{x * y} - \bar{x} * \bar{y}}{\overline{x^2} - (\bar{x})^2}$$
$$b_0 = \bar{y} - b_1 * \bar{x}$$

Тогда:

b1	0,3018
b2	4,7743

Уравнение регрессии имеет вид:

y' = 4,7743 - 0,3018*x

Промежуточный анализ: коэффициент регрессии указывает, что при увеличении квартиры на 1м^2 стоимость увеличивается на 0,3018.

Задание 3. Оценить качество уравнения с помощью средней ошибки аппроксимации.

$$\bar{A} = \frac{1}{n} \times \sum_{i}^{n} \left| \frac{y_i - y_i'}{y_i} \right| \times 100\%$$

В таблице 1 были представлены вычисления и получен результат: A = 9,03%

Задание 4. Найти коэффициент эластичности.

$$Э = b_1 \times \frac{\bar{x}}{\bar{y}}$$

Э = 0,8059

Промежуточный анализ: коэффициент эластичности показывает, что в среднем при увеличении общей площади квартиры на 1% ее стоимость поднимается на 0,806%.

Задание 5. Оценить тесноту связи между переменными с помощью показателей корреляции и детерминации.
Коэффициент корреляции:

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sigma_x \sigma_y}$$

Коэффициент детерминации равен коэффициенту корреляции в квадрате.

σ_x =	22,2551
σ_y =	7,2818
r =	0,9224
r^2 =	0,8509

Промежуточный анализ: так как значение коэффициента корреляции очень близко к единице, можно сделать вывод, что между признаками существует очень тесная связь,близкая к линейной зависимости. Коэффициент детерминации показывает, что 85% различий в стоимости квартир объясняется разницей в их площади, а 15% – другими неучтенными факторами.

Задание 6. Оценить значимость коэффициентов корреляции и регрессии по критерию t-Стьюдента при уровне значимости $\alpha = 0.05$.
Выдвинем гипотезу H0, о том, что коэффициент корреляции в генеральной совокупности равен нулю, и изучаемый признак не оказывает существенного влияния на результативный признак.
Для коэффициента корреляции:

$$t_{расч} = \frac{|r|}{\sqrt{\frac{1-r^2}{n-2}}}$$

t_расч =	8,6118

По таблице:

t_кр =	2,16

t_расч > t_кр, что означает, что гипотеза отвергается, и общая площадь квартир оказывает статистически существенное влияние на стоимость квартир.
Для регрессии:

$$t_{расч} = \frac{b_1}{m_{b_1}}, \quad m_{b_1} = \sqrt{\frac{\sum (y - y')^2}{(n-2) * \sigma_x^2 * n}}$$

m_b_1 = 0,035

t_расч = 8,61

По таблице:

t_кр = 2,16

t_расч > t_кр, что подтверждает вывод значимости влияния общей площади квартиры на ее стоимость.

Задание 7. Охарактеризовать прогнозное значение результативного признака, если возможное значение факторного признака составит 1.2 от его среднего уровня по совокупности.

$$x_p = \bar{x} * 1,2$$

x_p = 78,8

Тогда прогнозная стоимость квартиры:

(y')_p = 28,5571

Задача 2. Лекционная задача.

Постановка задачи: даны данные о распаде радиоактивного вещества. Необходимо показать, что процесс подчиняется экспоненциальному закону.

Исходные данные:

t	1	2	3	4	5	6	7	8	9	10	11	12
x	8,453	7,666	5,047	3,628	3,464	2,434	2,905	1,167	2,142	2,028	2,992	0,715

Решение:

Предположим, что функция регрессии имеет вид

$$x = ae^{bt}$$

Прологарифмировав эту функцию, обозначим следующее:

$$y = \ln(x), \quad a' = \ln a$$

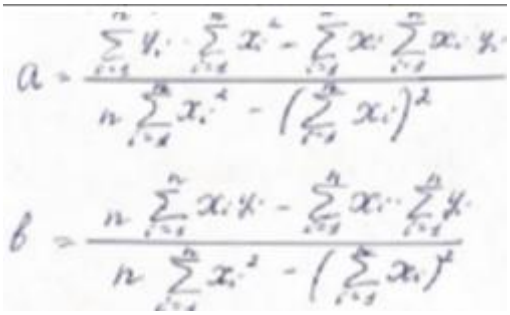
Тогда уравнение примет вид:

$$y = a' + bt$$

Это — линейное уравнение регрессии.

Вычисления расположим в таблице:

i	t	x	y	t^2	y^2	y*t
1	1	8,453	2,135	1	4,556	2,135
2	2	7,666	2,037	4	4,149	4,074
3	3	5,047	1,619	9	2,620	4,856
4	4	3,628	1,289	16	1,661	5,155
5	5	3,464	1,242	25	1,544	6,212
6	6	2,434	0,890	36	0,791	5,337
7	7	2,905	1,066	49	1,137	7,465
8	8	1,167	0,154	64	0,024	1,235
9	9	2,142	0,762	81	0,580	6,856
10	10	2,028	0,707	100	0,500	7,071
11	11	2,992	1,096	121	1,201	12,055
12	12	0,715	-0,335	144	0,113	-4,026
Итого	78	42,641	12,661	650	18,876	58,425
Среднее	6,5	3,553	1,055	54,167	1,573	4,869



a' = 2,140
b = -0,167

В таком случае получаем:

$$y = 2{,}14 - 0{,}167t$$

Находим a:

a = 8,50

Находим уравнение регрессии:

$$x = 8{,}5e^{-0{,}167t}$$

Вывод: во время выполнения лабораторной работы были сделаны следующие задачи: были получены уравнения линейной регрессии для двух задач, а также была проведена оценка качества регрессионной модели для первой задачи: найдены коэффициенты корреляции и детерминации, оценена их значимость по

критерию t-Стьюдента; охарактеризована статистическая надежность полученной регрессионной модели с помощью F-критерия Фишера; найдено прогнозное значение результативного признака.