ПЕРВОЕ ВЫСШЕЕ ТЕХНИЧЕСКОЕ УЧЕБНОЕ ЗАВЕДЕНИЕ РОССИИ

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

«САНКТ-ПЕТЕРБУРГСКИЙ ГОРНЫЙ УНИВЕРСИТЕТ»

Кафедра маркшейдерского дела

Отчет по практической работе №2

«Регрессионный и корреляционный анализ»

Вариант 8

Выполнил: студент гр. ГГ-18-2 ________ /Шевченко А.С./
(подпись) (Ф.И.О.)

Проверил: доцент ________ /Выстрчил М.Г./
(должность) (подпись) (Ф.И.О.)

Санкт-Петербург
2022

Исходные данные:

Вариант№08

№	X	Y
1	27,75	390,10
2	9,29	111,65
3	6,02	97,43
4	22,70	322,70
5	1,97	33,26
6	0,34	5,64
7	27,51	376,53
8	6,12	146,42
9	21,59	296,81
10	21,11	336,07
11	24,73	381,32
12	14,15	260,30
13	14,48	233,05
14	16,51	198,82
15	18,72	258,48
16	11,59	166,43
17	7,86	118,54
18	17,54	292,66
19	12,79	215,77
20	0,43	30,56
21	21,24	304,19
22	13,90	189,97
23	17,29	232,08
24	26,45	389,07
25	1,96	46,43
26	15,80	213,49
27	11,05	138,73
28	14,26	245,40
29	19,65	290,06
30	10,28	124,37
31	14,44	178,15

Регрессионный анализ – статистический аналитический метод, позволяющий вычислить предполагаемые отношения между зависимой переменной одной или несколькими независимыми переменными; устанавливает вид функциональной зависимости и подбирает функцию таким образом, чтобы она наилучшим образом определяла зависимость y от x. Цель регрессионного анализа – с помощью уравнения регрессии предсказать ожидаемое среднее значение результирующей переменной.

Корреляционный анализ – статистический метод, позволяющий с использованием коэффициентов корреляции определить, существует ли зависимость между переменными и насколько она сильна (есть ли связь между x и y). Корреляционная зависимость – это согласованные изменения двух (парная корреляционная связь) или большего количества признаков (множественная корреляционная связь). Суть ее заключается в том, что при изменении значения одной переменной происходит закономерное изменение (уменьшение или увеличение) другой(-их) переменной(-ых).

1. Регрессионный анализ

По исходным данным можно заметить, что с увеличением X, растет У. И чтобы описать их зависимость какой-то функцией, необходимо провести прямую, которая бы наилучшим образом вписывалась в исходные значения. Критерием для такой прямой будет минимальная сумма квадратов отклонений [vv] = min.

Уравнение линейной регрессии имеет вид:

$$\hat{y} = \hat{k}x + \hat{b}$$

$\hat{k}, \hat{b}$ – уравненные параметры

$\hat{y}$ – уравненное значение функции

Уравненное значение функции можно представить в виде:

$$y_i + v_i = \hat{k}x + \hat{b}$$

v_i – поправки в исходные значения функции

Отсюда выразим поправки как:

$$v_i = \hat{k}x + \hat{b} - y_i$$

Уравненное значение функции можно так же представить в виде:

$$\hat{y} = y_0 + \partial y = (k_0 + \delta k)x + (b_0 + \delta b)$$

Где k_0 и b_0 – приближенные значения параметров, которые можно задать любыми значениями для линейной функции (для нелинейной их следует задавать максимально близкими к истинным)

$$v_i = x_i\delta k + \delta b + k_0 x + b_0 - y_i$$

$k_0 x + b_0$ - это свободный член (то, что хотим исправить), произвольно близкое к итоговому значению

Далее воспользуемся параметрическим способом уравнивания

Задача уравнивания состоит в определении поправок и параметров. Для этого применяем линеаризацию функций (приведение к линейному виду с помощью разложения в ряд Тейлора)

$$v_i = \frac{\partial y}{\partial k}\partial k + \frac{\partial y}{\partial b}\partial b + y_0 - y_i$$

Где первые два слагаемых это результат разложения в ряд Тейлора

$y_0 - y_i = l_i$ – вектор невязок свободных членов

Обозначим:

$$\frac{\partial y}{\partial k} = a = x, \quad \frac{\partial y}{\partial b} = b = 1$$

В итоге, получаем параметрические уравнения поправок:

$$v_i = a\partial k + b\partial b + l_i$$

Применяя условия наименьших квадратов поправок, запишем:

$$[vv] = \sum_{i=1}^{n} (a_i \delta k + b_i \delta b + l_i)^2$$

Для соблюдения условия минимума квадратов поправок достаточно взять из выражения частные производные по каждому параметру в отдельности и приравнять их к нулю. После нахождения производных по всем параметрам будем иметь системунормальных уравнений:

$$\begin{cases} [aa]\partial k + [ab]\partial b + [al] = 0 \\ [ab]\partial k + [bb]\partial b + [bl] = 0 \end{cases}$$

Система параметрических уравнений поправок в матричной записи принимает вид

$$V = AT + L$$

Входящие в это выражение матрицы в развернутом виде равны:

Матрица коэффициентов параметрических уравнений поправок A:

$$A = \begin{pmatrix} a_1 & b_1 \\ a_2 & b_2 \\ \cdots & \cdots \\ a_n & b_n \end{pmatrix} = \begin{pmatrix} x_1 & 1 \\ x_2 & 1 \\ \cdots & \cdots \\ x_n & 1 \end{pmatrix}$$

Вектор невязок свободных членов L:

$$L = \begin{pmatrix} l_1 \\ l_2 \\ \cdots \\ l_n \end{pmatrix}$$

Вектор поправок T:

$$T = \begin{pmatrix} \delta k \\ \delta b \end{pmatrix}$$

Матричная запись условия наименьших квадратов для равноточных измерений (матрица весов Р равна 1) имеет вид:

$$V^T V = min$$

Матричная запись системы нормальных уравнений:

$$NT + A^T PL = 0$$

Где Т находится как:

$$T = -N^{-1} A^T L$$

N – матрица коэффициентов нормальных уравнений

$$N = A^T PA = A^T A = \begin{pmatrix} [aa] & [ab] \\ [ab] & [bb] \end{pmatrix}$$

$A^T PL$ - вектор свободных членов = $\begin{pmatrix} [al] \\ [bl] \end{pmatrix}$

Исправленные значения имеют вид:

$$k = k_0 + \delta k$$

$$b = b_0 + \delta b$$

№	X	Y	y0	li	ai	bi	[aa]	[ab]	[bb]	[al]	[bl]
1,00	27,75	390,10	28,75	-361,35	27,75	1,00	770,06	27,75	1,00	-10027,46	-361,35
2,00	9,29	111,65	10,29	-101,36	9,29	1,00	86,30	9,29	1,00	-941,63	-101,36
3,00	6,02	97,43	7,02	-90,41	6,02	1,00	36,24	6,02	1,00	-544,27	-90,41
4,00	22,70	322,70	23,70	-299,00	22,70	1,00	515,29	22,70	1,00	-6787,30	-299,00
5,00	1,97	33,26	2,97	-30,29	1,97	1,00	3,88	1,97	1,00	-59,67	-30,29
6,00	0,34	5,64	1,34	-4,30	0,34	1,00	0,12	0,34	1,00	-1,46	-4,30
7,00	27,51	376,53	28,51	-348,02	27,51	1,00	756,80	27,51	1,00	-9574,03	-348,02
8,00	6,12	146,42	7,12	-139,30	6,12	1,00	37,45	6,12	1,00	-852,52	-139,30
9,00	21,59	296,81	22,59	-274,22	21,59	1,00	466,13	21,59	1,00	-5920,41	-274,22
10,00	21,11	336,07	22,11	-313,96	21,11	1,00	445,63	21,11	1,00	-6627,70	-313,96
11,00	24,73	381,32	25,73	-355,59	24,73	1,00	611,57	24,73	1,00	-8793,74	-355,59
12,00	14,15	260,30	15,15	-245,15	14,15	1,00	200,22	14,15	1,00	-3468,87	-245,15
13,00	14,48	233,05	15,48	-217,57	14,48	1,00	209,67	14,48	1,00	-3150,41	-217,57
14,00	16,51	198,82	17,51	-181,31	16,51	1,00	272,58	16,51	1,00	-2993,43	-181,31
15,00	18,72	258,48	19,72	-238,76	18,72	1,00	350,44	18,72	1,00	-4469,59	-238,76
16,00	11,59	166,43	12,59	-153,84	11,59	1,00	134,33	11,59	1,00	-1783,01	-153,84
17,00	7,86	118,54	8,86	-109,68	7,86	1,00	61,78	7,86	1,00	-862,08	-109,68
18,00	17,54	292,66	18,54	-274,12	17,54	1,00	307,65	17,54	1,00	-4808,06	-274,12
19,00	12,79	215,77	13,79	-201,98	12,79	1,00	163,58	12,79	1,00	-2583,32	-201,98
20,00	0,43	30,56	1,43	-29,13	0,43	1,00	0,18	0,43	1,00	-12,53	-29,13
21,00	21,24	304,19	22,24	-281,95	21,24	1,00	451,14	21,24	1,00	-5988,62	-281,95
22,00	13,90	189,97	14,90	-175,07	13,90	1,00	193,21	13,90	1,00	-2433,47	-175,07
23,00	17,29	232,08	18,29	-213,79	17,29	1,00	298,94	17,29	1,00	-3696,43	-213,79
24,00	26,45	389,07	27,45	-361,62	26,45	1,00	699,60	26,45	1,00	-9564,85	-361,62
25,00	1,96	46,43	2,96	-43,47	1,96	1,00	3,84	1,96	1,00	-85,20	-43,47
26,00	15,80	213,49	16,80	-196,69	15,80	1,00	249,64	15,80	1,00	-3107,70	-196,69
27,00	11,05	138,73	12,05	-126,68	11,05	1,00	122,10	11,05	1,00	-1399,81	-126,68
28,00	14,26	245,40	15,26	-230,14	14,26	1,00	203,35	14,26	1,00	-3281,80	-230,14
29,00	19,65	290,06	20,65	-269,41	19,65	1,00	386,12	19,65	1,00	-5293,91	-269,41
30,00	10,28	124,37	11,28	-113,09	10,28	1,00	105,68	10,28	1,00	-1162,57	-113,09
31,00	14,44	178,15	15,44	-162,71	14,44	1,00	208,51	14,44	1,00	-2349,53	-162,71
							8352,06	449,52	31,00	-112625,38	-6143,96
k0=	1										
bo=	1										

Рисунок 1 – Вычисление элементов в системе нормальных уравнений

N=	8352,06	449,52
	449,52	31,00
AtPL=	-112625,38	
	-6143,96	
Nобр=	0,00	-0,01
	-0,01	0,15
T=	12,83	
	12,09	
k	13,83	
b	13,09	

Рисунок 2 – Расчет вектора поправок из системы нормальных уравнений

Уравнение регрессии:

$$y = 13{,}83x + 13{,}09$$

2. Корреляционный анализ

Качество аппроксимации описывают через 2 параметра: коэффициент ковариации и коэффициент корреляции.

Ковариация оценивает силу линейной зависимости между двумя числовыми переменными X и Y. Знак ковариации указывает на вид линейной связи между рассматриваемыми величинами: если она > 0 - это означает прямую связь (при росте одной величины растет и другая), ковариация < 0 указывает на обратную связь. При ковариации = 0 линейная связь между переменными отсутствует.

$$Kov_{xy} = \frac{\sum\left(x_i - M(x) \cdot (y_i - M(y))\right)}{n} = \frac{25367{,}73}{31} = 818{,}31$$

Следовательно, с ростом X значения Y действительно увеличивается.

Коэффициент корреляции показывает тесноту линейной взаимосвязи и изменяется в диапазоне от -1 до 1. -1 означает полную (функциональную) линейную обратную взаимосвязь. 1 – полную (функциональную) линейную положительную взаимосвязь. 0 – отсутствие линейной корреляции (но не обязательно взаимосвязи).

$$r_{xy} = \frac{Kov_{xy}}{\sigma_x \cdot \sigma_y}$$

$$\sigma_x = \sqrt{\frac{(x_i - M(x))^2}{n}} = \sqrt{\frac{1833{,}73}{31}} = 7{,}69$$

$$\sigma_y = \sqrt{\frac{(y_i - M(y))^2}{n}} = \sqrt{\frac{369161{,}75}{31}} = 109{,}13$$

Коэффициент корреляции:

$$r_{xy} = \frac{818{,}31}{7{,}69 \cdot 109{,}13} = 0{,}975$$

Значение $r_{xy} > 0$ и близко к 1, следовательно, корреляция положительная, и значения X и Y обладают между собой тесной линейной связью.

M(X)	M(Y)	xi-M(X)	yi-M(Y)	(xi-M(X))*(yi-M(Y))	Ков
14,50	213,69	13,25	176,41	2337,28	818,31
		-5,21	-102,04	531,71	
		-8,48	-116,26	985,98	
		8,20	109,01	893,79	
		-12,53	-180,43	2260,94	
		-14,16	-208,05	2946,16	
		13,01	162,84	2118,41	
		-8,38	-67,27	563,79	
		7,09	83,12	589,25	
		6,61	122,38	808,83	
		10,23	167,63	1714,72	
		-0,35	46,61	-16,34	
		-0,02	19,36	-0,40	
		2,01	-14,87	-29,88	
		4,22	44,79	188,97	
		-2,91	-47,26	137,57	
		-6,64	-95,15	631,88	
		3,04	78,97	240,01	
		-1,71	2,08	-3,55	
		-14,07	-183,13	2576,80	
		6,74	90,50	609,89	
		-0,60	-23,72	14,25	
		2,79	18,39	51,29	
		11,95	175,38	2095,64	
		-12,54	-167,26	2097,58	
		1,30	-0,20	-0,26	
		-3,45	-74,96	258,67	
		-0,24	31,71	-7,63	
		5,15	76,37	393,24	
		-4,22	-89,32	377,00	
		-0,06	-35,54	2,16	
	СУММ			25367,73	

Рисунок 3 – Расчет коэффициента ковариации

(xi-M(X))^2	сигмаX	(yi-M(Y))^2	сигмаУ	rxy
175,55	7,69	31119,46	109,13	0,98
27,15		10412,75		
71,92		13517,06		
67,23		11882,55		
157,02		32556,03		
200,52		43286,01		
169,24		26515,92		
70,24		4525,64		
50,26		6908,45		
43,68		14976,15		
104,64		28098,84		
0,12		2172,22		
0,00		374,70		
4,04		221,20		
17,80		2005,88		
8,47		2233,78		
44,10		9054,07		
9,24		6235,80		
2,93		4,31		
197,98		33537,66		
45,42		8189,72		
0,36		562,78		
7,78		338,09		
142,79		30757,13		
157,27		27976,88		
1,69		0,04		
11,91		5619,44		
0,06		1005,34		
26,52		5831,93		
17,81		7978,58		
0,00		1263,30		
1833,73		369161,74		

Рисунок 4 – Расчет коэффициента корреляции

Для оценки качества подбора уравнения регрессии определяется коэффициент детерминации.

Коэффициент детерминации – параметр, показывающий долю объясненной дисперсии. Чем ближе значение R^2 к 1, тем лучше регрессия описывает зависимость между результативным признаком и зависимой переменной.

$$R^2 = 1 - \frac{D_f(y)}{D(y)} = 1 - \frac{[vv]}{\left(y_i - M(y)\right)^2} = 1 - \frac{18226{,}01}{369161{,}745} = 0{,}95$$

Где v рассчитывается как разница между значением y, полученным по уравнению регрессии, и исходным значением.

Так как значение близко к 1, то доля объясненной дисперсии зависимой переменной высока.

Y
396,98
141,61
96,37
327,12
40,34
17,80
393,66
97,76
311,77
305,13
355,21
208,84
213,41
241,49
272,06
173,43
121,83
255,74
190,03
19,04
306,92
205,38
252,28
379,00
40,21
231,67
165,96
210,36
284,93
155,30
212,85

Рисунок 4 – Значения У из уравнения регрессии

v	vv	R^2
-6,88	47,39	0,95
-29,96	897,55	
1,06	1,12	
-4,42	19,56	
-7,08	50,19	
-12,16	147,75	
-17,13	293,56	
48,66	2368,23	
-14,96	223,70	
30,94	957,51	
26,11	681,98	
51,46	2647,92	
19,64	385,84	
-42,67	1820,75	
-13,58	184,50	
-7,00	48,96	
-3,29	10,80	
36,92	1363,15	
25,74	662,65	
11,52	132,70	
-2,73	7,48	
-15,41	237,58	
-20,20	408,07	
10,07	101,41	
6,22	38,73	
-18,18	330,44	
-27,23	741,30	
35,04	1227,53	
5,13	26,33	
-30,93	956,96	
-34,70	1204,36	
	18226,01	

Рисунок 5 – Расчет коэффициента детерминации

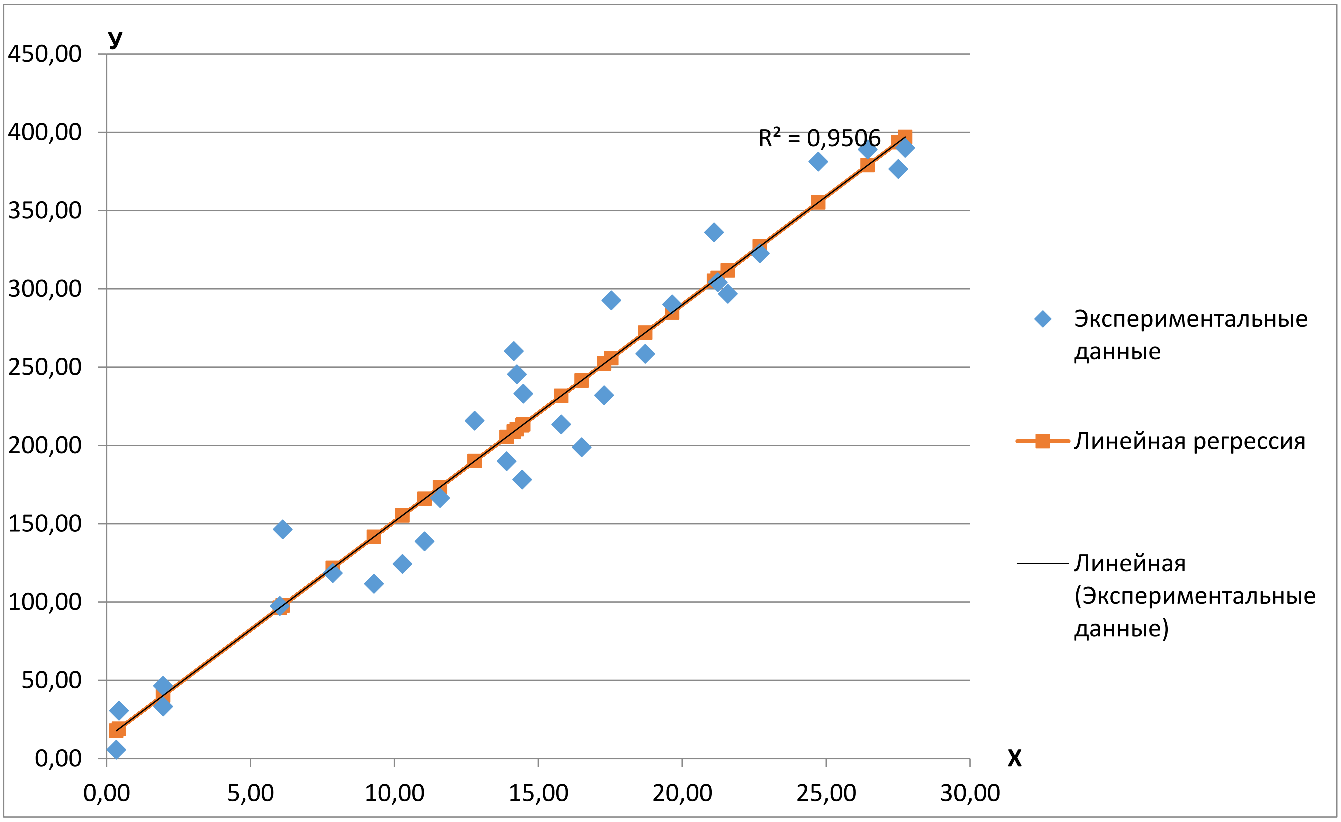

Рисунок 6 – Исходные данные, линейная регрессия и линия тренда с коэффициентом детерминации

<u>Вывод:</u> по исходным экспериментальным данным в результате регрессионного анализа было получено уравнение линейной регрессии:

$$y = 13{,}83x + 13{,}09$$

В результате корреляционного анализа был рассчитан коэффициент ковариации:

$$Kov_{xy} = 818{,}31,$$

который указывает на сильную линейную зависимость между Х и У, а так же был рассчитан коэффициент корреляции:

$$r_{xy} = 0{,}975,$$

который указывает на тесноту линейной зависимости между Х и У.

Для того чтобы выявить, какую долю дисперсии объясняет полученное уравнение регрессии, был рассчитан коэффициент детерминации:

$$R^2 = 0{,}95,$$

что означает, что подобранное уравнение линейной регрессии достаточно хорошо описывает поведение исходных данных.