Санкт-Петербургский Национальный
Исследовательский Университет Информационных
технологий, механики и оптики

Лабораторная работа #7

Построение и исследование моделей регрессионного анализа.

Выполнил: Канева
Тамара Игоревна
Группа № K3121
Проверила: Казанова
Полина Петровна

Санкт-Петербург
2021

Цель работы:

Изучить средства программы Microsoft Excel для регрессионного анализа данных.

Задачи:

Познакомиться с понятием корреляции и различными видами регрессии.

Ход работы:

Задание 1.

Для исходных данных (рис. 1) вычислим корреляцию попарно между признаками, применяя функцию “КОРРЕЛ”.

y	x
6	2
13	5
23	9
5	3
22	8

Рис. 1. Исходные данные.

В результате мы получим коэффициент корреляции, равный 0.98453529.

Сделаем то же самое, но теперь с помощью функции “Корреляция”, находящейся в пакете “Анализ данных”. Перед нами откроется окно “Корреляция” (рис. 2), в которое вводим необходимые параметры.

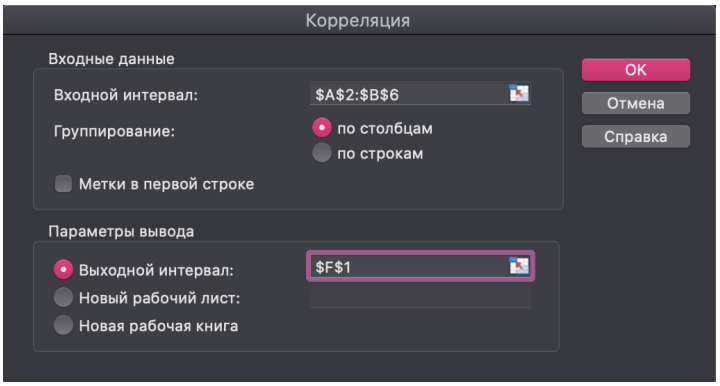

Рис. 2. Окно “Корреляция”.

На выходе получаем следующую таблицу (рис. 3), из которой мы видим, что коэффициент корреляции такой же как и при предыдущей попытки вычисления.

	Столбец 1	*Столбец 2*
Столбец 1	1	
Столбец 2	0,98453529	1

Рис. 3. Вывод функции “Корреляция”.

Использовать можно обе функции, но функция “Корреляция” даёт более полный ответ на запрос, нежели “КОРРЕЛ”.

Задание 2.

Подсчитаем коэффициент корреляции для данных, представленных в таблице (рис. 4).

x	12,1	14,7	20,5	16,6	19
y	53,2	44,2	51,4	45,5	34

Рис. 4. Исходные данные.

Воспользовавшись функцией “КОРРЕЛ”, получим коэффициент корреляции, равный -0.3636188.

Задание 3.

На основе данных, представленных в таблице (рис. 5), построим линейную модель и проведём её анализ.

7	483
6	489
7	486
8	563
9	570
9	559
9	594
9	575
6	464
9	647

Рис. 5. Исходные данные.

Определим параметры уравнения регрессии помощью метода наименьших квадратов, который реализован в Excel. Для этого используем функцию “Регрессия” в пакете “Анализ данных”. В результате ее выполнения появится несколько таблиц (рис. 6).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множествен	0,89009298
R-квадрат	0,79226552
Нормирован	0,76629871
Стандартная	0,62201726
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*начимость F*
Регрессия	1	11,8047562	11,8047562	30,5106985	0,00055798
Остаток	8	3,09524378	0,38690547		
Итого	9	14,9			

	Коэффициент	*дартная ош*	*статистик*	*P-Значение*	*Нижние 95%*	*ерхние 95%*	*ижние 95,0%*	*ерхние 95,0%*
Y-пересечен	-2,5567416	1,9032773	-1,3433364	0,21602168	-6,945707	1,83222368	-6,945707	1,83222368
Переменная	0,01925735	0,00348635	5,52364902	0,00055798	0,01121782	0,02729688	0,01121782	0,02729688

Рис. 6. Результат выполнения функции “Регрессия”.

Основываясь на данных их таблиц на рисунке 6, можем сделать следующие выводы:

1. Уравнение регрессии имеет вид: y=-2.557x+0.019.

2. Значение коэффициента детерминации, равного 0,79 показывает, что количество работников существенно влияет на объём производства, что подтверждает правильность включения его в построенную модель.

3. Рассчитанный уровень значимости, равный 0.0006 и меньший 0,05 подтверждает значимость величины коэффициента детерминации.

4. *P-Значение* для переменной, равное 0,0006 и меньшее 0,05 подтверждает значимость коэффициента b_1.

5. *P-Значение* для коэффициента b_0 превышает 0,05, это означает, что данный коэффициент для модели не является значимым и его можно опустить, т.е. график модели будет проходить через точку начала координат.

Тогда раз коэффициент b_0 не является значимым, то модель можно построить без его учета (рис. 7).

ВЫВОД ИТОГОВ

Регрессионная статистика

Множествен	0,99702732
R-квадрат	0,99406348
Нормирован	0,88295237
Стандартная	0,64922476
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*начимость F*
Регрессия	1	635,206565	635,206565	1507,04018	2,1303E-10
Остаток	9	3,79343507	0,42149279		
Итого	10	639			

	Коэффициент	*дартная ош*	*-статистик*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*ижние 95,0%*	*ерхние 95,0%*
Y-пересечен	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
Переменная	0,01459909	0,00037607	38,8206154	2,4816E-11	0,01374837	0,01544981	0,01374837	0,01544981

Рис. 6. Результат выполнения функции “Регрессия” с включенным флагом “Константа-ноль”.

Тогда:

1. Уравнение регрессии имеет вид: y=0.016x.
2. Значение коэффициента детерминации, равного 0,99 показывает, что количество работников существенно влияет на объём производства, что подтверждает правильность включения его в построенную модель.
3. Рассчитанный уровень значимости, равный 2.1303E-10 и меньший 0,05 подтверждает значимость величины коэффициента детерминации.
4. *P-Значение* для срока службы, равное 2.4816E-11 и меньшее 0,05 подтверждает значимость коэффициента b_1.

Задание 4.

Построим модель зависимости величины заработной платы от стажа работы и пола сотрудника (рис. 7).

з/п	стаж	пол
38900	15	0
28700	2	1
31600	4	1
33800	13	0
31890	16	0
45000	35	0
24313	10	1
22700	8	1
36300	20	0
32350	7	0
31800	5	0

Рис. 7. Исходные данные.

ВЫВОД ИТОГОВ

Регрессионная статистика

Множествен	0,84116263
R-квадрат	0,70755457
Нормирован	0,63444321
Стандартная	3788,08022
Наблюдения	11

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	277743190	138871595	9,67776536	0,00731441
Остаток	8	114796414	14349551,8		
Итого	10	392539605			

	Коэффициенты	*Стандартная ош*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечен	30189,6179	2790,62977	10,8182096	4,7048E-06	23754,4141	36624,8217	23754,4141	36624,8217
стаж	348,762835	151,057805	2,30880381	0,04978196	0,42291287	697,102757	0,42291287	697,102757
пол	-5453,9449	2802,57739	-1,9460461	0,08752518	-11916,7	1008,81014	-11916,7	1008,81014

Рис. 8. Полученная линейная модель.

Построив модель заметим, что чем больше стаж работника, тем больше ему платят (коэффициент при переменной, отвечающей за стаж, положителен). В то же время женщинам платят меньше (коэффициент при переменной, отвечающей за пол, отрицателен, а сама переменная равна 1 в случае, когда речь идет о женщине).

Однако заметим, что коэффициент при переменной, отвечающей за пол, не существенен, то есть модель не адекватна. Поэтому построим модель без учета этой переменной (рис. 9).

Новая модель адекватна, а уравнение регрессии имеет вид y=504.9455x+26289.5776.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множествен	0,75439692
R-квадрат	0,56911472
Нормирован	0,52123858
Стандартная	4335,12192
Наблюдения	11

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*начимость F*
Регрессия	1	223400066	223400066	11,88723	0,00730018
Остаток	9	169139538	18793282		
Итого	10	392539605			

	Коэффициент	*дартная ош*	*статистик*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*ижние 95,0%*	*ерхние 95,0%*
Y-пересечен	26289,5776	2222,41648	11,8292758	8,6987E-07	21262,1222	31317,033	21262,1222	31317,033
стаж	504,945528	146,454998	3,44778625	0,00730018	173,641305	836,249751	173,641305	836,249751

Рис. 9. Полученная линейная модель, не учитывающая пол работника.

Задание 5.

Определим по данным (рис. 10) параметры уравнения линейной регрессии (рис. 11) и проведём его анализ.

прибыль	средства	фонды
188	129	510
78	64	190
93	69	240
152	87	470
55	47	110
161	102	420

Рис. 10. Исходные данные.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множествен	0,9972714
R-квадрат	0,99455024
Нормирован	0,99091707
Стандартная	5,05002558
Наблюдения	6

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*начимость F*
Регрессия	2	13962,3251	6981,16253	273,741469	0,00040231
Остаток	3	76,508275	25,5027583		
Итого	5	14038,8333			

	Коэффициент	*дартная ош*	*статистик*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*ижние 95,0%*	*ерхние 95,0%*
Y-пересечен	-1,943423	7,62541782	-0,2548612	0,81530285	-26,210906	22,3240598	-26,210906	22,3240598
средства	0,69499204	0,19685945	3,53039707	0,03862823	0,0684974	1,32148667	0,0684974	1,32148667
фонды	0,20234768	0,03519963	5,74857343	0,01045664	0,09032674	0,31436863	0,09032674	0,31436863

Рис. 11. Полученная линейная модель.

Из полученных таблиц мы видим, что свободный член уравнения не существенен. Построим другую линейную модель, но с флагом “Константа-ноль” (рис. 12).

Из полученных таблиц мы видим, что модель адекватна. Также мы видим, что полученная прибыль зависит только от оборотных средств и основных фондов, а все прочие факторы не вносят заметных изменений в поведение модели (и предприятий, соответственно).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множествен	0,99961724
R-квадрат	0,99923463
Нормирован	0,74904329
Стандартная	4,42054258
Наблюдения	6

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*начимость F*
Регрессия	2	102048,835	51024,4176	2611,12041	1,3757E-05
Остаток	4	78,1647868	19,5411967		
Итого	6	102127			

	Коэффициент	*дартная ош*	*статистик*	*P-Значение*	*Нижние 95%*	*ерхние 95%*	*ижние 95,0%*	*ерхние 95,0%*
Y-пересечен	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
средства	0,65646781	0,11039484	5,94654424	0,0040119	0,34996259	0,96297302	0,34996259	0,96297302
фонды	0,20666567	0,02700801	7,65201292	0,0015673	0,1316794	0,28165194	0,1316794	0,28165194

Рис. 12. Новая полученная линейная модель.

Вывод:

В ходе этой лабораторной работы мы научились решать практические задачи, связанные с различными видами регрессии.

Ответы на контрольные вопросы:

1. Регрессионная модель - модель отражающая влияние одной или нескольких независимых переменных на одну зависимую. Мы рассматриваем линейные регрессионные модели.

2. Общий вид регрессионной модели: y = f(x1, x2, …, xn), где y - зависимая переменная, x1, x2, …, xn - независимые переменные, а f - некоторая функция от n переменных, задающая необходимое соотношение.

3. Значимость коэффициента регрессии проверяется на основе P-значения: если оно меньше или равно 0.05, то коэффициент значимый.

4. При определении коэффициентов регрессионной модели обычно применяется метод наименьших квадратов.