УДК 004:37

РЕГРЕССИОННЫЙ АНАЛИЗ В ЭЛЕКТРОННЫХ ТАБЛИЦАХ

Курзаева Л.В.

ФГБОУ ВО «Магнитогорский государственный технический университет им. Г.И. Носова», Магнитогорск, e-mail: lkurzaeva@mail.ru

Аналитическая статистика – одниз самых сложных разделов анализа данных в плане изучения, при этом регрессионный анализ является одним из самых информативных. Такой анализ производится при решения следующих задач: установление и оценка взаимосвязи признаков; прогнозирование и предсказание; управление процессами. Существует два вида анализа двумерных данных, представленных переменными: корреляционный и регрессионныйанализ, последнийпозволяет определить форму взаимосвязи между признаками. В статье описывается простой способ проведения регрессионного анализа в MicrosoftExcel. Материалы данной статьи представляют методическую и практическую ценность для преподавателей, занимающихся вопросами повышения эффективности обучения в области основ анализа данных с информационных технологий, и осуществляющие реализацию образовательного процесса в вузах и на курсах повышения квалификаций.

Ключевыеслова: анализ данных, электронные таблицы

REGRESSION ANALYSIS IN SPREADSHEETS

Kurzaeva L.V.

NosovMagnitogorsk State Technical University, Magnitogorsk, e-mail: lkurzaeva@mail.ru

Analytical statistics is one of the most difficult sections of the data analysis in terms of studying, while regression analysis is one of the most informative. Such analysis is performed under the following tasks: the establishment and evaluation of the relationship between signs; forecasting and prediction; process control. There are two types of analysis of two-dimensional data represented by variables: correlation and regression analysis, the latter allows to determine the form of the relationship between signs. This paper describes a simple method of regression analysis in Microsoft Excel.The contents of this article are of methodological and practical value to teachers working to increase the effectiveness of training in the area of foundations of data analysis with information technology, and implementing the educational process in universities and in courses of improvement of qualifications.

Keywords: dataanalysis, spreadsheets

Для реализации процедуры Регрессия необходимо: выбрать в меню Сервис команду Анализ данных. В появившемся диалоговом окне Анализ данных в списке Инструменты анализа выбрать строку Регрессия.

В появившемся диалоговом окне (рис.1) задать:

Входной интервал Y– диапазон (столбец), содержащий данные со значениями объясняемой переменной;

Входной интервал X– диапазон (столбцы), содержащий данные с заголовками.

Метки – флажок, который указывает, содержат ли первые элементы отмеченных диапазонов названия переменных (столбцов) или нет;

Константа-ноль– флажок, указывающий на наличие или отсутствие свободного члена в уравнении (а);

Уровень надежности– уровень значимости, (например, 0,05);

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона, в котором будет сохранен отчет по построению модели;

Новый рабочий лист– поставить значок и задать имя нового листа (Отчет – регрессия), в котором будет сохранен отчет.

Если необходимо получить значения и график остатков, а также график подбора (чтобы визуально проверить отличие экспериментальных точек от предсказанных по регрессионной модели), установите соответствующие флажки в диалоговом окне.

Рассмотрим результаты регрессионного анализа (рис. 2, 3).

Множественный R – коэффициент корреляции

R-квадрат – это коэффициент линейной детерминации. Коэффициент является одной из наиболее эффективных оценок адекватности регрессионной R^2модели, мерой качества уравнения регрессии в целом (или, как говорят, мерой качества подгонки регрессионной модели к наблюденным значениям.

Если R-квадрат > 0,95, говорят о высокой точности аппроксимации (модель хорошо описывает явление). Если R-квадрат лежит в диапазоне от 0,8 до 0,95, говорят об удов-

	A	B	C	D
1	Страны Восточной Европы	Доля городского населения, %	ВВП на душу населения с учетом паритета покупательной способности в 2008 г., (US$)	Число мобильных телефонов на 100 жителей
2	Белоруссия	74	12150	84
3	Болгария	71	11950	138
4	Чешская республика	74	22790	134
5	Венгрия	67	17790	122
6	Молдавия	41	3210	67
7	Польша	61	17310	115
8	Румыния	55	13500	115
9	**Россия**	**73**	**15630**	**141**
10	Словакия	55	21300	102
11	Украина	69	7210	121
12				
13				
14				
15				

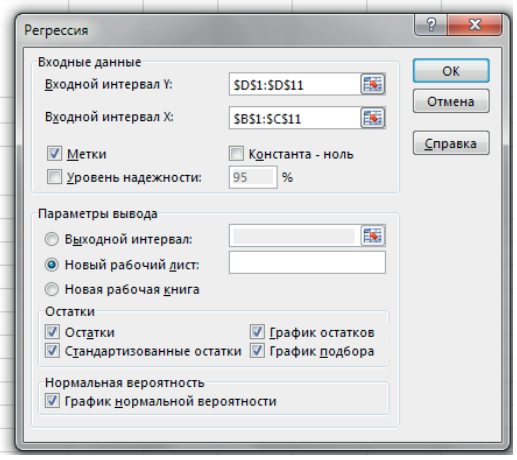

Рис.1. Окно «Регрессия»

летворительной аппроксимации (модель в целом адекватна описываемому явлению). Если R-квадрат < 0,6, принято считать, что точность аппроксимации недостаточна и модель требует улучшения (введения новых независимых переменных, учета нелинейностей и т. д.).

Нормированный R-квадрат – скорректированный (адаптированный, поправленный) коэффициент детерминации.

Недостатком коэффициента детерминации *R-квадрат*является то, что он увеличивается при добавлении новых объясняющих переменных, хотя это и не обязательно означает улучшение качества регрессионной модели. В этом смысле предпочтительнее использовать *нормированный*, который в отличие от *R-квадрат* может уменьшаться при введении в модель новых объясняющих переменных, не оказывающих существенное влияние на зависимую переменную.

Наблюдения – число наблюдений (в нашем случае 10 стран).

Df– число степеней свободы связано с числом единиц совокупности и с числом определяемых по ней констант.

F и Значимость F позволяют проверить значимость уравнения регрессии, т.е. установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

SS – Сумма квадратов отклонений значений признака Y.

MS – Дисперсия на одну степень свободы.

F – Наблюдаемое (эмпирическое) значение статистики *F,* по которой проверяется гипотеза равенства нулю одновременно всех коэффициентов модели. *Значимость F* – теоретическая вероятность того, что при гипотезе равенства нулю одновременно всех коэффициентов модели *F*-статистика больше эмпирического значения *F.*

На уровне значимости α=0,05 гипотеза H_0:b_1=0отвергается, если *Значимость F<0.05*, и принимается, если *Значимость F≥0.05* . В нашем примере *Значимость F>0.05,* что говорит о неадекватности модели Следует понимать, что «плохой результат – тоже результат» – полученная оценка модели важна для ее последующего осмысления, т.к. дальнейший анализ может подсказать какие из независимых переменных незначимы и ухудшают качество модели.

Значения коэффициентов регрессии находятся в столбце Коэффициенты и соответствуют:

У-пересечение – *a*;

переменная XI – b_1;

переменная X2 – b_2 и т. Д.

Таким образом, получена следующая модель регрессии:

$$Y=1.2247X1+0.00108X2+19.9776$$

t-статистика соответствующего коэффициента.

P-Значение – вероятность, позволяющая определить значимость коэффициента регрессии. В случаях, когда *P-Значение*>0,05, коэффициент может считаться нулевым, что означает, что соответствующая независимая переменная практически не влияет на зависимую переменную.

1	ВЫВОД ИТОГОВ								
2									
3	*Регрессионная статистика*								
4	Множественный R	0,713385							
5	R-квадрат	0,508918							
6	Нормированный R-квадрат	0,368608							
7	Стандартная ошибка	18,86498							
8	Наблюдения	10							
9									
10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	2	2581,688	1290,844	3,627113	0,082993			
13	Остаток	7	2491,212	355,8875					
14	Итого	9	5072,9						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	19,97762	37,50242	0,532702	0,610725	-68,7015	108,6568	-68,7015	108,6568
18	Доля городского населения, %	1,224737	0,62508	1,959327	0,09091	-0,25334	2,702817	-0,25334	2,702817
19	ВВП на душу населения с учетом паритета покупательной способности в 2008 г., (US$)	0,001088	0,001126	0,966484	0,365998	-0,00157	0,00375	-0,00157	0,00375

Рис. 2. Вывод итогов регрессионного анализа

	A	B	C	D	E	F	G
23	ВЫВОД ОСТАТКА					ВЫВОД ВЕРОЯТНОСТИ	
24							
25	*Наблюдение*	*Число мобильных телефонов на 100 жителей*	*Остатки*	*Стандартные остатки*		*Персентиль*	*Число мобильных телефонов на 100 жителей*
26	1	123,8258391	-39,8258391	-2,393761204		5	67
27	2	119,9340537	18,06594634	1,085866925		15	84
28	3	135,4008436	-1,4008436	-0,084198729		25	102
29	4	121,3883047	0,611695297	0,036766393		35	115
30	5	73,68391246	-6,68391246	-0,40174145		45	115
31	6	113,5177037	1,482296312	0,089094505		55	121
32	7	102,0244744	12,97552561	0,779903462		65	122
33	8	126,3869121	14,61308788	0,878330341		75	134
34	9	110,50991	-8,50991001	-0,511494369		85	138
35	10	112,3280463	8,671953745	0,521234126		95	141

Рис. 3. Вывод остатков и вероятности по результатам регрессионного анализа

Рис. 4. График остатков по значениям признака «Доля городского населения, %»

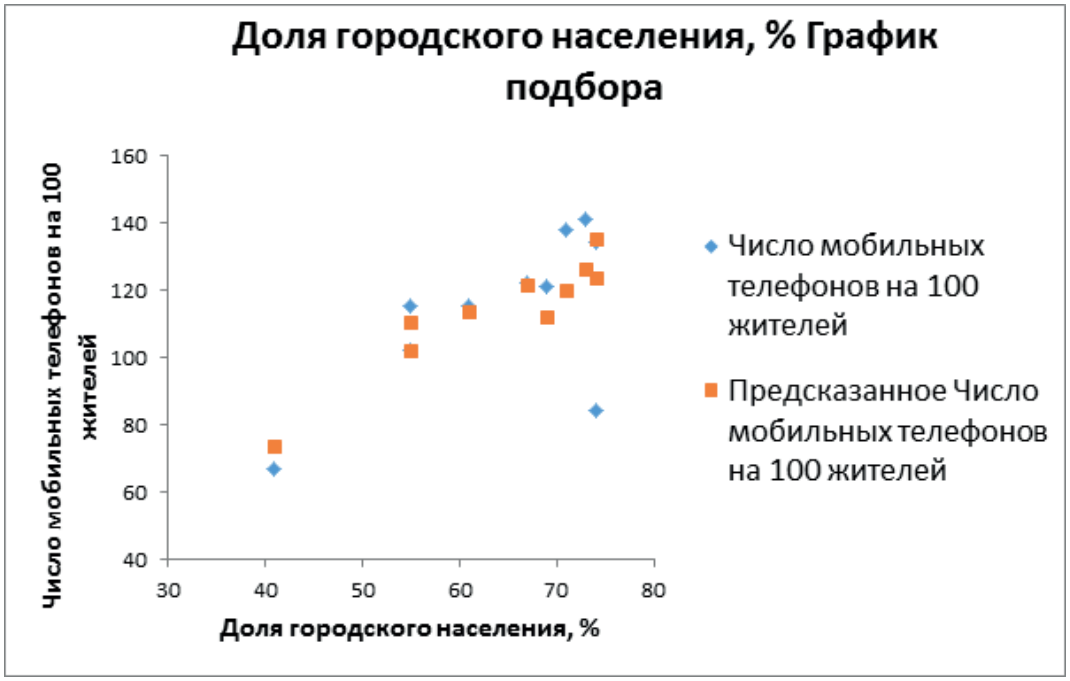

Рис. 5. График подбора для признаков «Доля городского населения, %» и «Число мобильных телефонов на 100 жителей»

В нашем случае оба коэффициента оказались «нулевыми», а значит обе независимые переменные не влияют на модель.

Нижние 95% – Верхние 95% – доверительный интервал для параметра , т.е. с надежностью 0.95 этот коэффициент лежит в данном интервале. Поскольку коэффициент регрессии в исследованиях имеют четкую интерпретацию, то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов. Так, например, «Доля городского населения, в %» не может лежать в интервале $-0,25 \geq b_1 \geq 2,7$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

Предсказанное Y - теоретические (расчетные) значения результативного признака.

Остатки – остатки по модели регрессии.

На основе данных об остатках модели регрессии был построен график остатков (рис. 4) и график подбора – поле корреляции фактических и теоретических (расчетных) значений результативной переменной (рис.5).

Рассмотрение графиков подбора позволяет предположить, что, возможно, качество модели можно усовершенствовать, исключив данные по Белоруссии как аномальные значения.

Список литературы

1. Овчинникова И.Г., Варфоломеева Т.Н., Гусева Е.Н. Учебно-методическое пособие для подготовки к вступительным экзаменам по информатике. -Магнитогорск, 2002. -С. 119

2. Овчинникова И.Г., Варфоломеева Т.Н., Корнещук Н.Г. Учебное пособие для подготовки к централизованному тестированию по информатике. -Магнитогорск, 2002. -С.205

3. Курзаева Л.В. Дистанционный курс «Основы математической обработки информации»: электронный учебно-методический комплекс // Хроники объединенного фонда электронных ресурсов Наука и образование. - 2014. -Т. 1. - № 12 (67). - С. 117

4. Курзаева Л.В. Введение в теорию систем и системный анализ: учеб. пособие/Л.В. Курзаева. -Магнитогорск: МаГУ, 2015. -211 с.

5. Курзаева Л.В. Введение в методы и средства получения и обработки информации для задач управления социальными и экономическими системами: учеб. пособие/Л.В. Курзаева, И.Г. Овчинникова, Г.Н. Чусавитина. -Магнитогорск:Магнитогорск. гос. техн. ун-та им. Г.И. Носова, 2016. -118 с.