Министерство образования Российской Федерации

Магнитогорский Государственный технический университет
имени Г. И. Носова

Кафедра технологий обработки материалов

МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Методические указания

Магнитогорск
2017

1.ЦЕЛИ РАБОТЫ

В реальных условиях функционирования технических объектов и организационно-технических систем зависимость результата функционирования (отклика Y) от управляемых и контролируемых воздействий (факторов $X_1, \ldots, X_j, \ldots, X_m$) проявляется как опосредованная разнообразными случайными причинами (возмущениями). Подобные зависимости принято называть стохастическими.

Множественный регрессионный анализ – это метод математической статистики, который позволяет найти наиболее точное и достоверное отображение (модель, аппроксимацию) стохастической зависимости между откликом Y и несколькими факторами $X_1, X_2, \ldots, X_j, \ldots, X_m$.

Цели работы:

Освоение методики множественного регрессионного анализа.

Приобретение навыков решения задачи множественного регрессионного анализа в среде электронных таблиц *MS Excel* с применением инструмента «Регрессия».

Приобретение навыков решения задачи множественного регрессионного анализа в среде электронных таблиц *MS Excel* с применением статистических функций.

2.КРАТКИЕ СВЕДЕНИЯ ИЗ СТАТИСТИКИ

2.1.Содержание и допущения регрессионного анализа

Понятие стохастической зависимости в некотором смысле является обобщением понятия о зависимости функциональной. Для последней характерно, что каждому значению фактора (аргумента) соответствует совершенно определенное значение отклика. В случае стохастической зависимости при определенном значении x_{ji} фактора X_j может наблюдаться множество значений отклика Y. В производственных условиях фактор также является случайной величиной, но при проведении регрессионного анализа полагают, что всякое его значение x_{ji} неслучайно [1].

Учитывая возможные отклонения, модель связи отклика с некоторым комплексом факторов $\vec{X} = \{X_1, \ldots, X_j, \ldots, X_m\}$ должна быть представлена в виде двух составляющих:

$$y = \varphi\left(\vec{X}\right) + \varepsilon, \qquad (1)$$

где $\varphi\left(\vec{X}\right)$ - систематическая (объясненная) составляющая. Она обусловлена существованием зависимости между откликом и комплексом факторов;

ε - случайная составляющая. Она обусловлена разнообразными возмущениями и вызывает отклонения y от значений, соответствующих реальной зависимости.

Для построения множественной регрессионной модели (иначе – множественного уравнения регрессии или просто множественной регрессии) необходимо решить следующие задачи:

1. Определить вид уравнения регрессии.
2. Оценить значимость коэффициентов регрессии.
3. Оценить допустимость отображения исследуемой зависимости выбранным уравнением регрессии.
4. Исследовать остатки (отклонения действительных значений отклика от предсказываемых по уравнению регрессии).

2.2.Определение вида уравнения множественной регрессии

Задача определения вида уравнения множественной регрессии состоит в нахождении систематической составляющей $\varphi\left(\vec{X}\right)$. Однако, поскольку используются выборки ограниченного объема ($n <<< \infty$), могут быть найдены лишь оценки истинных параметров.

Пусть, например, действительная зависимость отклика от комплекса факторов является линейной:

$$y = \varphi\left(\vec{X}\right) = \beta_0 + \sum_{j=1}^{m} \beta_j X_j . \qquad (2)$$

Оценкой (моделью, отображением, аппроксимацией) этой связи также может быть линейное выражение:

$$\hat{y} = \hat{\varphi}\left(\vec{X}\right) = b_0 + \sum_{j=1}^{m} b_j X_j . \qquad (3)$$

В выражении (3), которое и есть уравнение регрессии, коэффициенты регрессии b_0 и b_j ($j = 1, 2, \ldots, m$) представляют собой оценки коэффициентов истинной зависимости ($b_0 \approx \beta_0$ и $b_j \approx \beta_j$).

Для подбора уравнения $\hat{y} = \hat{\varphi}\left(\vec{X}\right)$, которое наилучшим образом отображает стохастическую связь между откликом и рассматриваемыми факторами, используют метод наименьших квадратов (МНК). Согласно МНК наилучшей оценкой исследуемой зависимости является та, которая дает наименьшую сумму квадратов отклонений наблюдаемых значений отклика y_i от рассчитанных по уравнению регрессии $\hat{y}_i$ при тех же значениях факторов $\left\{x_{1i},...,x_{ji},...,x_{mi}\right\}$. Это условие выражается следующим образом:

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = min\,. \qquad (4)$$

Исходя из условия (4) задача определения коэффициентов уравнения регрессии сводится практически к определению минимума функции нескольких переменных и решена математической статистикой для линейного уравнения. Значения коэффициентов регрессии в (3) вычисляются решением системы из n линейных уравнений с m неизвестными (здесь n - число наблюдений) [1, 2, 3].

В *MS Excel* коэффициенты линейной аппроксимации могут быть определены с использованием статистической функции ЛИНЕЙН(). Синтаксис функции и вопросы, связанные с ее использованием приведены в прил.2.

В *MS Excel* коэффициенты множественной линейной регрессии можно определить с использованием инструмента «Регрессия». Правила и особенности его использования приведены в приложении 1.

2.3.Проверка значимости коэффициентов регрессии

Коэффициенты регрессии b_j являются случайными величинами с математическими ожиданиями β_j и дисперсиями, которым соответствуют стандартные отклонения S_{bj}. Значение b_j признается статистически значимым, если выполняется условие:

$$t_{bj} = \frac{\left|b_j\right|}{S_{bj}} > t\left[\alpha; n-k\right], \qquad (5)$$

где t_{bj} и $t\left[\alpha; n-k\right]$ - расчетное и табличное числа Стьюдента.

Если условие (5) не выполнено, то следует признать, что $b_j = 0$ и влияние фактора X_j на отклик несущественное. В таком случае рекомендуется повторить регрессионный анализ без учета фактора X_j.

2.4.Оценивание надежности аппроксимации

Из различных уравнений регрессии наилучшим в смысле МНК считают то, которое обеспечивает минимум дисперсии фактических (полученных экспериментально) значений отклика относительно линии регрессии. Эту дисперсию называют остаточной или дисперсией относительно регрессии и определяют по формуле:

$$S_e^2 = \frac{1}{n-k}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 , \qquad (6)$$

где k - число коэффициентов регрессии в уравнении.

Точность отображения (аппроксимации) исследуемой зависимости выбранным уравнением регрессии оценивают с помощью дисперсионного анализа. Для этого сравнивают дисперсию относительно регрессии (S_e^2) с оценкой дисперсии значений y_i относительно выборочного среднего фактических значений отклика $\bar{y}$:

$$S_E^2 = \frac{1}{k-1}\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2 = \frac{1}{m}\sum_{i=1}^{n}\left(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 . \qquad (7)$$

Величина S_E^2 характеризует рассеяние y_i, обусловленное зависимостью отклика от факторов в виде оцениваемого уравнения регрессии, и поэтому называется объясненной дисперсией. Остаточная дисперсия S_e^2 характеризует рассеяние y_i, вызванное случайными воздействиями (возмущениями). Аппроксимация связи между откликом и факторами в виде данного уравнения регрессии допустима (статистически надежна), если объясненная дисперсия существенно больше остаточной.

Сравнение дисперсий производится проверкой условия:

$$F = \frac{S_E^2}{S_e^2} > F\left[\alpha\,;\nu_E\,;\nu_e\right], \qquad (8)$$

где F - рассчитанное число Фишера; $F\left[\alpha,\nu_E,\nu_e\right]$ - табличное число Фишера для заданного уровня значимости α при степенях свободы $\nu_E = k-1$ и $\nu_e = n-k$. В *MS Excel* табличное число Фишера может быть найдено с применением статистической функции FРАСПОБР($\alpha\,;k-1\,;n-k$).

Если условие (8) выполняется, то объясненная дисперсия существенно больше остаточной. А это означает, что между откликом и факторами существует взаимосвязь, которую с вероятностью $p = 1-\alpha$ допустимо аппроксимировать рассматриваемым уравнением регрессии.

Из нескольких допустимых аппроксимаций наиболее точной, очевидно, будет та, для которой значение S_e^2 является наименьшим. Отсюда следует, что для наиболее точной модели $\hat{y} = \hat{\varphi}\left(\vec{X}\right)$ различие расчетного и табличного чисел Фишера будет максимальным.

2.5.Анализ остатков

Остатками принято называть отклонения действительных значений отклика от рассчитанных по уравнению регрессии (для i-го наблюдения остаток $e_i = y_i - \hat{y}_i$). Отклонения обусловлены наличием в (1) случайной составляющей ε, относительно которой делают следующие предположения:

1. Это нормально распределенная случайная переменная.
2. Математическое ожидание случайной составляющей равно нулю - $M(\varepsilon) = 0$. Считают, что данная гипотеза выполняется, если среднее выборочное остатков можно считать равным нулю:

$$\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i \approx 0\,. \qquad (9)$$

3. Дсперсия случайной составляющей постоянна - $D(\varepsilon) = Const$. Гипотеза может быть проверена, например, построением графиков остатков в зависимости от каждого фактора. Если на всех таких графиках остатки примерно равномерно рассеяны в пределах области, параллельной

оси X_j (рис. П1.1а), то гипотезу $D(\varepsilon) = Const$ считают справедливой.

4. В различных наблюдениях значения ε не зависят друг от друга. Для проверки гипотезы о независимости отклонений в различных наблюдениях оценивают автокорреляцию остатков с применением критерия Дарбина-Уотсона (критерия DW):

$$DW = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \qquad (10)$$

Строгое условие отсутствия автокорреляции DW =2 [2, 6]. Однако, с учетом особенностей распределения критерия Дарбина-Уотсона, ориентировочно можно считать, что автокорреляция остатков отсутствует при $1{,}2 \le DW \le 2{,}8$ [6]. В противном случае следует признать, что гипотеза о независимости остатков в рассматриваемом случае не верна.

Если анализ остатков обнаруживает несоответствия указанным гипотезам, то уравнение регрессии, относительно которого данные остатки получены, следует считать неудовлетворительным, т. к. правомерность применения МНК и указанных выше оценок и для множественного регрессионного анализа может быть поставлена под сомнение. В таком случае рекомендуют рассмотреть уравнение иного вида (например, нелинейное вместо линейного), включить неучтенные ранее факторы, выделить в области варьирования факторов различные подобласти.

3.ВЫПОЛНЕНИЕ РАБОТЫ

Порядок выполнения работы иллюстрируется (рис.1, 2 и 3) на примере построения зависимости предела текучести металла, прокатанного на широкополосном стане горячей прокатки (ШСГП) от температур конца прокатки (tкп) и смотки (tсм). Исходные данные заносятся на рабочий лист с клавиатуры (на рис. 1 и 3 они расположены в ячейках A1:C29).

3.1.Аппроксимация с применением инструмента «РЕГРЕССИЯ»

При выполнении работы настройки инструмента «РЕГРЕССИЯ» должны соответствовать указанным в прил. 1.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	ИСХОДНЫЕ ДАННЫЕ				ВЫВОД ИТОГОВ			УРАВНЕНИЕ РЕГРЕССИИ						
2	St, МПа	tкп,С	tсм,С					St = 1152,965 - 0,498*tкп -0,695*tсм						
3	364,0	788	560		*Регрессионная статистика*			ЗНАЧИМОСТЬ КОЭФФИЦИЕНТОВ						
4	341,2	788	584		Множественный R	0,937		t[0,05;n-k]	b0	Значим				
5	338,4	788	611		R-квадрат	0,879		2,0595	b1	Значим				
6	284,1	788	639		Нормированный R-квадрат	0,869			b2	Значим		АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ		
7	330,9	788	657		Стандартная ошибка	18,66		НАДЕЖНОСТЬ АППРОКСИМАЦИИ						
8	281,9	788	715		Наблюдения	28		F[0,05;k-1;n-k]	Аппроксимация			DW	2,3644	
9	265,9	788	728					3,3852	надежная			Отсутствует		
10	336,4	834	560		Дисперсионный анализ									
11	320,5	834	584			*df*	*SS*	*MS*	*F*	*Значимость F*				
12	336,8	834	611		Регрессия	2	62268,036	31134,018	90,4187	3,6E-12				
13	299,4	834	639		Остаток	25	8608,291	344,332						
14	269,3	834	657		Итого	27	70876,327							
15	240,3	834	715											
16	206,3	834	728			*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 90,0%*	*Верхние 90,0%*	
17	334,2	875	560		Y-пересечение	1152,695	73,484	15,6863	1,9E-14	1001,352	1304,039	1027,174	1278,217	
18	290,3	875	584		tкп,С	-0,498	0,073	-6,7941	4,03E-07	-0,649	-0,347	-0,623	-0,373	
19	318,5	875	611		tсм,С	-0,695	0,060	-11,6051	1,47E-11	-0,818	-0,572	-0,797	-0,593	
20	283,2	875	639											
21	281,0	875	657											
22	233,1	875	715											
23	200,0	875	728		ВЫВОД ОСТАТКА									
24	304,4	917	560											
25	307,5	917	584		*Наблюдение*	*Предсказанное St, МПа*	*Остатки*	$(e_i - e_{i-1})^2$						
26	268,8	917	611		1	371,3	-7,3	53,6						
27	225,4	917	639		2	354,6	-13,4	37,48						
28	270,3	917	657		3	335,9	2,5	254,80						

Рис. 1. Фрагмент рабочего листа с результатами множественного регрессионного анализа инструментом «РЕГРЕССИЯ»

В этом случае основные результаты (ячейки E1:M19 на рис. 1) будут дополнены таблицей остатков (ячейки E23:G23 на рис. 2).

	A	B	C	D	E	F	G	H
22	233,1	875	715					
23	200,0	875	728		ВЫВОД ОСТАТКА			
24	304,4	917	560					
25	307,5	917	584		*Наблюдение*	*Предсказанное Sm, МПа*	*Остатки*	$(e_i-e_{i-1})^2$
26	268,8	917	611		1	371,3	-7,3	53,6
27	225,4	917	639		2	354,6	-13,4	37,48
28	270,3	917	657		3	335,9	2,5	254,80
29	181,3	917	715		4	316,4	-32,3	1214,00
30	175,3	917	728		5	303,9	27,0	3517,48
31					6	263,6	18,3	75,61
32					7	254,6	11,3	48,53
33					8	348,4	-12,0	545,15
34					9	331,8	-11,3	0,60
35					10	313,0	23,8	1229,38
36					11	293,5	5,9	321,94
37					12	281,0	-11,7	309,47
38					13	240,7	-0,4	127,80
39					14	231,7	-25,4	623,31
40					15	328,0	6,2	996,21
41					16	311,3	-21,0	741,05
42					17	292,6	25,9	2205,48
43					18	273,1	10,1	250,99
44					19	260,6	20,4	106,26
45					20	220,3	12,8	57,69
46					21	211,3	-11,3	579,18
47					22	307,1	-2,7	73,28
48					23	290,4	17,1	391,16
49					24	271,7	-2,9	397,50
50					25	252,2	-26,8	573,25
51					26	239,7	30,6	3295,72
52					27	199,4	-18,1	2371,23
53					28	190,4	-15,1	9,20
54								

Рис. 2. Таблица остатков, полученная с использованием инструмента «РЕГРЕССИЯ»

Таблицу остатков дополнить столбцом, во всех строках которого, начиная со второй, вычислить значения $(e_i - e_{i-1})^2$. Эти данные будут в дальнейшем использованы для расчета критерия DW. Например, в ячейке H27:

=(G27-G26)^2.

На основании результатов работы инструмента записать уравнение регрессии в содержательной форме, оценить значимость коэффициентов регрессии, надежность аппроксимации и автокорреляцию остатков.

Уравнение регрессии в содержательной форме (ячейки H1:J2) записывается с клавиатуры.

Оценивание значимости коэффициентов регрессии (ячейки H3:J6). В ячейке H5 определяется табличное число Стьюдента. Для рассматриваемого примера (число коэффициентов регрессии k =3):

=СТЬЮДРАСПОБР(0,05;2;F8-3).

В ячейках J4:J6, с использованием функции ЕСЛИ() программируется вывод о значимости коэффициентов регрессии. Например, для ячейки J4:

=ЕСЛИ(ABS(H17)>$H$5;”Значим”;”Не значим”) .

Внимание! Если коэффициент регрессии при факторе Xj оказался не значимым, необходимо повторить регрессионный анализ, не включая во входной интервал X столбец со значениями Xj. Поскольку входной интервал X должен состоять из смежных столбцов, может оказаться необходимым перегруппировать столбцы со значениями факторов.

Оценивание надежности аппроксимации в примере на рис. 1 выполняется в ячейках H7:J9. В ячейке H9, с помощью статистической функции FРАСПОБР(), определяется табличное значение числа Фишера:

=FРАСПОБР(0,05;2;F8-3) .

Слово «Аппроксимация» в ячейки I8:J8 введено с клавиатуры. Собственно вывод («надежная» или «не надежная») формируется в ячейке I9 с применением функции ЕСЛИ() путем сравнения табличного (из ячейки H9) и рассчитанного (из ячейки I12) чисел Фишера:

=ЕСЛИ(H9>I12;”надежная”;”не надежная”) .

Оценивание автокорреляции остатков выполнено в ячейках L7:M9. Значение критерия Дарбина-Уотсона в ячейке M8 вычисляется по формуле (10), которая для рассматриваемого примера программируется следующим образом:

=СУММ(H27:H53)/G13 .

Вывод в ячейке L9 формируется с применением функции ЕСЛИ(), которая проверяет условие $1{,}2 \le DW \le 2{,}8$:

ЕСЛИ(M8<1,2;”Отсутствует”;ЕСЛИ(M8>2,8;”Отсутствует”;”Существует”)) .

3.2.Множественный регрессионный анализ с применением функции ЛИНЕЙН()

Пример множественного регрессионного анализа функцией ЛИНЕЙН() представлен на рис. 3. Синтаксис функции ЛИНЕЙН() и

особенности ее применения изложены в прил. 2.

	A	B	C	D	E	F	G	H	I	J	K	L
1	ИСХОДНЫЕ ДАННЫЕ				РЕЗУЛЬТАТЫ				ОСТАТКИ			
2	Sт, МПа	tкп,С	tсм,С		ЛИНЕЙН()				y^	e	(ei- ei+1)²	
3	364,0	788	560		-0,695	-0,498	1152,695		371,3	-7,3	-	
4	341,2	788	584		0,060	0,073	73,484		354,6	-13,4	37,5	
5	338,4	788	611		0,879	18,556	#Н/Д		335,9	2,5	254,8	
6	284,1	788	639		90,419	25	#Н/Д		316,4	-32,3	1214,0	
7	330,9	788	657		62268,036	8608,291	#Н/Д		303,9	27,0	3517,5	
8	281,9	788	715		РЕГРЕССИОННАЯ СТАТИСТИКА				263,6	18,3	75,6	
9	265,9	788	728		R^2	S_e	n-k		254,6	11,3	48,5	
10	336,4	834	560		0,879	18,556	25		348,4	-12,0	545,2	
11	320,5	834	584		ΣE^2	Σe^2	Fp		331,8	-11,3	0,6	
12	336,8	834	611		62268,036	8608,291	90,419		313,0	23,8	1229,4	
13	299,4	834	639		ТАБЛИЧНЫЕ ЗНАЧЕНИЯ				293,5	5,9	321,9	
14	269,3	834	657		P, %	t [0,05;n-k]	F[0,05;k-1;n-k]		281,0	-11,7	309,5	
15	240,3	834	715		95,00	2,0595	3,3852		240,7	-0,4	127,8	
16	206,3	834	728		ОЦЕНИВАНИЕ КОЭФФИЦИЕНТОВ				231,7	-25,4	623,3	
17	334,2	875	560		b2	b1	b0		328,0	6,2	996,2	
18	290,3	875	584		-0,695	-0,498	1152,695		311,3	-21,0	741,0	
19	318,5	875	611		Sb2	Sb1	Sb0		292,6	25,9	2205,5	
20	283,2	875	639		0,060	0,073	73,484		273,1	10,1	251,0	
21	281,0	875	657		tb2	tb1	tb0		260,6	20,4	106,3	
22	233,1	875	715		-11,6051	-6,7941	15,6863		220,3	12,8	57,7	
23	200,0	875	728		Значим	Значим	Значим		211,3	-11,3	579,2	
24	304,4	917	560		УРАВНЕНИЕ РЕГРЕССИИ				307,1	-2,7	73,3	
25	307,5	917	584		Sт = 1152,965 - 0,498*tкп - 0,695*tсм				290,4	17,1	391,2	
26	268,8	917	611		ОЦЕНКА АППРОКСИМАЦИИ				271,7	-2,9	397,5	
27	225,4	917	639		Аппроксимация надежная				252,2	-26,8	573,2	
28	270,3	917	657		АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ				239,7	30,6	3295,7	
29	181,3	917	715		DW		2,364		199,4	-18,1	2371,2	
30	175,3	917	728		Отсутствует				190,4	-15,1	9,2	
31												

Рис. 3. Фрагмент рабочего листа с результатами множественного регрессионного анализа функцией ЛИНЕЙН()

В рассматриваемом примере отыскивается уравнение регрессии относительно 2 факторов (число коэффициентов k =3) со всеми коэффициентами (т. е. b_0 должно быть определено) и требуется вывод регрессионной статистики. Поэтому предварительно должна быть выделена область смежных ячеек из 3 столбцов и 5 строк (на рис. 3 – ячейки E3:G7).

Запись функции для рассматриваемого примера имеет вид:

=ЛИНЕЙН(A3:A30;B3:C30;;ИСТИНА) .

Программирование функции должно быть завершено комбинацией клавиш <Ctr>+<Shift>+<Inter>. Результаты будут размещены в пределах предварительно выделенной области смежных ячеек E3:G7 (смысл выводимых параметров см. в прил. 2).

Используя ссылки на соответствующие результаты, сформировать блок данных «РЕГРЕССИОННАЯ СТАТИСТИКА» (ячейки E8:G12). Например, значение R^2 в ячейке E10 воспроизводится ссылкой ссылки на ячейку E5 (в E10 запрограммировано =E5).

Для оценивания результатов потребуются табличные значения чисел Стьюдента ($t\left[\alpha; n-k\right]$) и Фишера ($F\left[\alpha; k-1; n-k\right]$). В примере на рис. 3 они расположены в ячейках F15 и G15 блока «ТАБЛИЧНЫЕ ЗНАЧЕНИЯ». В ячейке F15 запрограммировано:

=СТЬЮДРАСПОБР(1-E15/100;G10),

а в ячейке G15:

=FРАСПОБР(1-E15/100;2;G10) .

При обращении к указанным функциям уровень значимости задан не числом, а рассчитывается через доверительную вероятность, которая вводится с клавиатуры в ячейку E15. Такая конструкция позволяет оценивать результаты регрессионного анализа при варьировании доверительной вероятности в широком диапазоне, но требует обеспечить изменчивость уровня значимости в обозначениях табличных чисел. Для этого может быть применена функция ФИКСИРОВАННЫЙ(). Например, для обозначения табличного числа Стьюдента в ячейке F14:

=”t [“&ФИКСИРОВАННЫЙ(1-E15/100;2)&”;n-k]” .

Оценивание коэффициентов регрессии выполняется в области E16:G23. Значения коэффициентов b_j воспроизводятся в ячейках E18:G18 с применением ссылок на соответствующие ячейки из массива результатов (например, в E18 записано =E3). Таким же образом в ячейках E20:G20 воспроизводятся ошибки коэффициентов регрессии S_{bj} (например, в E20 записано =E4). Расчетные числа Стьюдента t_{bj} определяются по формуле (5) и размещаются в ячейках E22:G22 (например, в E22 записано =E18/E20). Вывод о значимости коэффициентов должен формулироваться автоматически на основании проверки условия (5) с применением функция ЕСЛИ(). Например, в ячейке E23:

=ЕСЛИ(ABS(E22)>$F$15;"Значим";"Не значим") .

Внимание! Если коэффициент регрессии при факторе Xj оказался не значимым, необходимо повторить регрессионный анализ, не включая во входной интервал X столбец со значениями Xj. Поскольку входной интервал X должен состоять из смежных

столбцов, может оказаться необходимым перегруппировать столбцы со значениями факторов.

Если все коэффициенты значимы, записать уравнение регрессии в содержательной форме (на рис. 3 – ячейки E25:G25).

Оценка статистической надежности аппроксимации выполняется проверкой условия (8) в ячейке E27:

=ЕСЛИ(G12>G15;"Аппроксимация надежная";"Аппроксимация не надежная") .

Оценка автокорреляции остатков. Сначала создается таблица остатков (на рис. 3 – ячейки I2:J30). Здесь *y^* - оценки отклика при соответствующих значениях факторов, а *e* – отклонения от линиии регрессии. Например, в ячейках I3 и J3:

=$G$3+$E$3*C3+$F$3*B3 и =A3-I3 .

Затем к таблице остатков добавляется столбец с квадратами разностей смежных отклонений. Для первого наблюдения (ячейка K3) данный параметр не рассчитывается. Для остальных расчет выполняется подобно тому, как, например, в ячейке K4:

=(J4-J3)^2 .

Критерий Дарбина-Уотсона (10) вычисляется в ячейке G29:

=СУММ(K4:K30)/F12 .

Вывод об автокорреляции остатков запрограммирован в ячейку E30 следующим образом:

=ЕСЛИ(G29<1,2;"Существует";ЕСЛИ(G29>2,8;"Существует";"Отсутствует")).

4.СОДЕРЖАНИЕ ВЫВОДОВ ПО РАБОТЕ

В выводах по работе необходимо ответить на следующие вопросы:

1. Связь между какими величинами анализировалась?
2. Значимы ли коэффициенты регрессии?
3. Как выглядит уравнение множественной регрессии, полученное в результате выполнения работы?
4. Можно ли считать полученное уравнение множественной регрессии статистически надежной аппроксимацией анализируемой зависимости?

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Львовский Е.Н, Статистические методы построения эмпирических формул. - 2-е изд., перераб. и доп.- М.: Высш. шк.,, 1988. - 239 с.
2. Математическая статистика / Иванова В. М., Калинина В. Н., Нешумова Л. А. и др. – М.: Высшая школа, 1981. – 371 с.
3. Четыркин Е. М., Калихман И. Л. Вероятность и статистика. – М.: Финансы и статистика, 1982. – 320 с.

4. Ахназарова С. Л., Кафаров В. В. Организация эксперимента в химии и химической технологии - М.: Высш. шк., 1978. - 319 с.
5. Дьяконов В.П. Справочник по алгоритмам и программам на языке Бейсик для персональных ЭВМ: Справочник. - М., Наука. Гл. ред. физ.-мат. лит., 1987. - 240 с.
6. Замков О. О., Толстопятенко А. В., Черемных Ю. Н. Математические методы в экономике. – М.: МГУ им. М. В. Ломоносова, издательство «ДИС», 1997. – 368 с.

Содержание

ПРИЛОЖЕНИЕ 1

ИНСТРУМЕНТ «РЕГРЕССИЯ» И ЕГО ПРИМЕНЕНИЕ

П1.1.Диалоговое окно инструмента

Инструмент «РЕГРЕССИЯ» позволяет выполнить комплексный *линейный* регрессионный анализ с помощью метода наименьших квадратов. Он входит в пакет «Анализ данных», который относится к *надстройкам MS Excel*. Запуск инструмента «РЕГРЕССИЯ» осуществляется последовательным выбором пунктов из меню различных уровней:

<Сервис> / <Анализ данных> / <Регрессия>.

На экране откроется диалоговое окно (рис. П1.1).

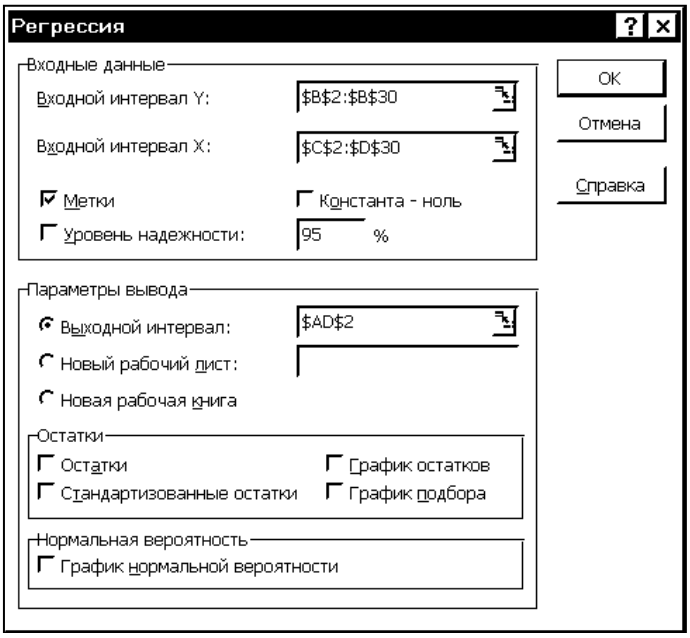


Рис. П1.1. Диалоговое окно «Регрессия»

Входной интервал Y. Здесь должна быть введена ссылка на диапазон ячеек, в которых записаны значения отклика. Диапазон должен состоять из одного столбца.

Входной интервал X. Здесь должна быть введена ссылка на диапазон ячеек, в которых записаны значения факторов. Максимальное число факторов 16. При отсутствии указателя «Метки»

(см. ниже) каждый фактор автоматически идентифицируются номером, соответствующим его расположению во входном интервале слева направо в порядке возрастания.

Метки. Указатель наличия условных обозначений отклика и факторов среди обрабатываемых данных. Если опция включена, то данные, записанные в первых ячейках входных диапазонов, воспринимаются как условное обозначение и при анализе не учитывается. Например, при группировке данных по столбцам условным обозначением будет считаться данное, расположенное в первой строке каждого из них. При группировке по строкам в качестве условного обозначения будут восприниматься данные, расположенные в первом столбце каждой из них. **Внимание!** Если анализируемые данные не снабжены условным обозначением, то задание опции «**Метки …** » приведет к потере данных из первых ячеек входного диапазона, что повлечет за собой ошибки в результатах регрессионного анализа.

Уровень надежности. По умолчанию доверительные границы коэффициентов регрессии определяются для доверительной вероятности 95%. Если данная опция будет включена, то будут также вычислены границы для доверительной вероятности, заданной исследователем.

Константа – ноль. Указатель на необходимость определения коэффициента регрессии b_0. Если опция выключена, анализ выполняется для уравнения, в котором априорно принимается b_0=0.

Выходной диапазон, Новый лист, Новая книга. Указатели расположения результатов работы инструмента «Регрессия».

При выборе опции **Выходной диапазон** результаты выводятся на активный рабочий лист активной книги *MS Excel*. Они будут размещены слева направо и сверху вниз, начиная с ячейки, адрес которой задан в окне. Порядок расположения и смысл результатов работы инструмента «Регрессия» поясняется на рис. П1.2. При выборе опций **Новый лист** или **Новая книга** результаты выводятся в новый рабочий лист активной книги или в первый лист другой книги *MS Excel*. В обоих случаях результаты будут размещены начиная с ячейки А1.

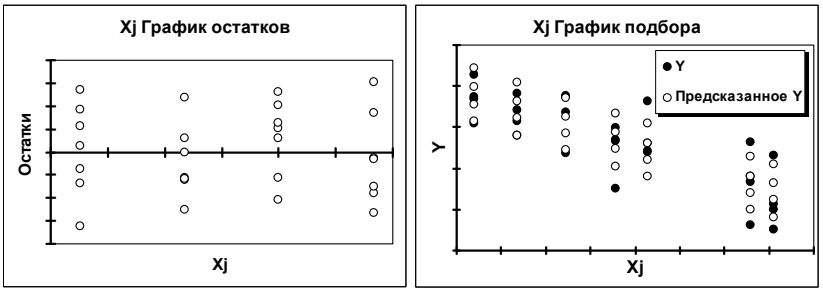
Остатки. Стандартизированные остатки. Если опции заданы, то в результаты работы инструмента «Регрессия» будет включена таблица остатков в следующем виде:

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*	*Стандартные остатки*
1	$\hat{y}_1$	$y_1 - \hat{y}_1$	$\approx (y_1 - \hat{y}_1)/S_e$
…	…	…	
i	$\hat{y}_i$	$y_i - \hat{y}_i$	
…	…	…	
n	$\hat{y}_n$	$y_n - \hat{y}_n$	

Таблица полезна, например, для проверки гипотезы о равенстве нулю математического ожидания случайной составляющей исследуемой зависимости. Значения остатков могут также быть использованы для оценивания их автокорреляции путем вычисления критерия Дарбина-Уотсона (10).

График остатков. Если опция задана, то для каждого из факторов будет выдан график (рис. П1.2а), которые позволяют визуально оценить выполнение гипотезы о постоянстве дисперсии случайной составляющей уравнения регрессии.



а б

Рис. П1.2. Общий вид графиков подбора и остатков

График подбора. Если опция задана, то для каждого из факторов будет выдан график (рис. П1.2б) с наблюдаемыми и предсказанными значениями отклика.

График нормальной вероятности. Если опция задана, будет построен график (рис.

Рис. П1.3. График нормальной вероятности

П1.3), который отображает функцию распределения отклика из предположения, что это распределение является нормальным.

П1.2. Содержание результатов работы инструмента «РЕГРЕССИЯ»

Ниже описано содержание результатов работы инструмента «РЕГРЕССИЯ», которые выводятся в область «Выходной диапазон» рабочего листа и разделены на 3 таблицы (рис. 1 на стр. 7).

Регрессионная статистика

Множественный R – коэффициент множественной корреляции

$$R = \sqrt{1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}}$$

R-квадрат – коэффициент множественной детерминации.

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}$$

Нормированный R-квадрат – несмещенный коэффициент множественной детерминации. Несмещенность достигается учетом числа степеней свободы $n - k = \nu_e$ и $n - 1 = \nu_Y$.

$$R^2_{норм} = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2} \cdot \frac{n-k}{n-1}$$

Стандартная ошибка – среднее квадратическое отклонение относительно линии регрессии (остаточное среднее квадратическое отклонение)

$$S_e = \sqrt{\frac{1}{n-k} \sum e_i^2} = \sqrt{\frac{1}{n-k} \sum (y_i - \hat{y})^2}$$

Наблюдения – число наблюдений n.

Дисперсионный анализ

Данные, позволяющие выполнить анализ составляющих вариации отклика (*Analysis of variance – ANOVA*).

Рассматриваются следующие составляющие вариации: *Регрессия* (вариация отклика, обусловленная существованием зависимости отклика от рассматриваемых факторов), *Остаток* (вари-

ация, обусловленная случайными причинами), *Итого* (общая вариация).

Для каждой из составляющих вариации выводится число степеней свободы *df* ():

Регрессия $df_E = k - 1 = \nu_E$

Остаток $df_e = n - k = \nu_e$

Итого $df_Y = \nu_E + \nu_e = (k-1) + (n-k) = n - 1 = \nu_Y$

Используются следующие характеристики вариации: *SS* – суммы квадратов отклонений (*Sums of Squares*) и *MS* – средние квадраты (*Mean Squares*).

Регрессия $SS_E = \sum E_i^2 = \sum (\hat{y}_i - \bar{y})^2 ; MS_E = \dfrac{\sum E_i^2}{\nu_E} = S_E^2$

Остаток $SS_e = \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 ; \; MS_e = \dfrac{\sum e_i^2}{\nu_e} = S_e^2$

Итого $SS_Y = SS_E + SS_e = \sum (y_i - \bar{y}_i)^2$

Для оценивания статистической надежности аппроксимации исследуемой зависимости линейным уравнением регрессии выводятся:

F - расчетное число Фишера

$$F = \frac{MS_E}{MS_e} = \frac{S_E^2}{S_e^2} = \frac{R^2}{1 - R^2} \cdot \frac{n-k}{k-1};$$

Значимость F - вероятность того, рассчитанное число Фишера не соответствует гипотезе $S_E^2 > S_e^2$. Отыскивается по распределению Фишера $\alpha_F = f_F[F; \nu_E; \nu_e]$ с использованием статистической функции FРАСП($F; k-1; n-k$). Величина $p = 1 - \alpha_F$ представляет собой доверительную вероятность утверждения о статистической надежности полученной аппроксимации.

Коэффициенты регрессии

Значения коэффициентов регрессии и данные для оценивания их значимости выводятся в таблицу без названия. Число строк таблицы равно числу коэффициентов в уравнении регрессии. В строке *Y-пересечение* размещены данные для коэффициента b_0,

а в остальных – для коэффициентов b_j. Если массив исходных данных содержит обозначения факторов и анализ выполнялся с заданной опцией «Метки», то строки будут содержать обозначения факторов. В противном случае каждая строка будет помечена идентификатором, соответствующим расположению факторов во входном интервале слева направо в порядке возрастания.

Коэффициенты – значения коэффициентов регрессии.

Стандартная ошибка – среднее квадратическое отклонение коэффициента регрессии:

$$S_{bj} = \sqrt{\frac{S_e^2}{\sum \left(x_{ji} - \bar{x}_j\right)^2}}.$$

t-статистика – расчетное число Стьюдента для коэффициента регрессии b_j (знак коэффициента сохраняется):

$$t_{bj} = \frac{b_j}{S_{bj}}.$$

P-значение – вероятность того, что t_{bj} не соответствует гипотезе о значимости коэффициента регрессии b_j. Отыскивается по распределению Стьюдента $p_S = f_S\left[t_{bj}; \nu_e; \nu_E\right]$ с использованием статистической функции СТЬЮДРАСП($t_{bj}; k-1; n-k$). Величина $p = 1 - p_S$ представляет собой доверительную вероятность утверждения о статистической значимости анализируемого коэффициента регрессии.

Нижние P% и *Верхние P%* - нижняя и верхняя границы доверительного интервала для коэффициента регрессии b_j при доверительной вероятности *P*:

$$b_j - S_{bj} t\left[1-p; n-k\right] \text{ и } b_j + S_{bj} t\left[1-p; n-k\right].$$

ПРИЛОЖЕНИЕ 2

СТАТИСТИЧЕСКАЯ ФУНКЦИЯ «ЛИНЕЙН()» И ЕЕ ПРИМЕНЕНИЕ

Статистическая функция ЛИНЕЙН() позволяет определить коэффициенты линейной регрессии и некоторые параметры, необходимые для ее оценивания. Может применяться как для множественного, так и для парного регрессионного анализа.

Синтаксис функции:

ЛИНЕЙН(<Y>; <X>; К$_1$; К$_2$),

где <Y> - ссылка на значения отклика;

<X> - ссылка на значения факторов;

К$_1$ - указатель на необходимость определения коэффициента регрессии b_0. Если задано «ЛОЖЬ», априорно принимается b_0=0;

К$_2$ - указатель на необходимость вывода расчетных параметров для оценивания регрессии.

Функция ЛИНЕЙН() относится к функциям массива. Поэтому при ее программировании следует помнить следующие правила.

1. Сначала на рабочем листе необходимо выделить область ячеек, размеры которой зависят от варианта использования функции. Ширина области (число столбцов) должно соответствовать числу коэффициентов, которое будет содержаться в уравнении регрессии. Например, для уравнения относительно m факторов, при необходимости определения коэффициента b_0 (К$_1$ задано «ИСТИНА» или пропущено) число коэффициентов регрессии $k = m + 1$. Высота области (число строк) зависит от указателя К$_2$. Если К$_2$ задано «ИСТИНА», то область должна включать 5 строк, а в противном случае (К$_2$ задано «ЛОЖЬ» или пропущено) область вывода результатов работы функции должна содержать 1 строку.
2. Программирование функции должно быть завершено комбинацией клавиш <Ctr>+<Shift>+<Inter>.

В самом общем случае (рассматривается m факторов, К$_1$ задано «ИСТИНА» или пропущено и К$_2$ задано «ИСТИНА») результаты работы функции ЛИНЕЙН() будут представлены следующим образом (обозначения результатов соответствуют принятым в разделе 2 данных методических указаний):

b_m	b_{m-1}	…	b_j	…	b_1	b_0
S_{bm}	S_{bm-1}	…	S_{bj}	…	S_{b1}	S_{b0}
R^2	S_e	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
F_p	$n-k$	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
$\sum_{i=1}^{n} E_i^2$	$\sum_{i=1}^{n} e_i^2$	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д

Символы #Н/Д, которые в *MS Excel* обычно указывают на ошибку при обращении к функции, в данном случае содержательного значения не имеют.