Използване на MS Excel в обучението по статистика

Красимира Костадинова

Abstract: A Practical training of Statistics by MS Excel: *This paper presents an application of Regression analysis in teaching statistics with MS Excel.*

Key words: *Regression Analysis, Statistics, MS Excel.*

ВЪВЕДЕНИЕ

Обучението по Статистика изисква осъзнаването на голямо количество формули и статистически методи. То не е достатъчно ефективно без илюстрации на тяхното приложение. От своя страна приложението е свързано с много по обем изчислителни процедури. Това налага използването на подходящ софтуер.

Една от най-достъпните софтуерни среди е MS Office, в частност MS Excel. Тази софтуерна среда (MS Excel) позволява да бъдат представени теми като многомерен регресионен анализ, многомерен дисперсионен анализ и клъстърен анализ, които са особени важни в приложението на маркетинга, социологически и политически проучвания.

Масовото познаване и използване на продукта MS Office (MS Excel) дава възможност на всеки потребител лесно да продължи наученото в училище и да развие знанията си посредством използването на статистика в Excel.

Тук показваме как с помощта на MS Excel може да се реализира обучението по темата многомерен регресионен анализ (РА).

ИЗЛОЖЕНИЕ

Дефиниция (*РА*). РА служи за моделиране формата на зависимостта на един зависим (резултативен) признак от един или няколко фактор-признаци, като не се отчита, че изменението на разглежданите величини може да се дължи на външни, невключени в модела признаци.

Формата, свързваща резултативния признак с фактора-признак (или фактор-признаците) се нарича уравнение на регресия.

Ако фактор-признака е един, говорим за еднофакторен РА. А ако фактор-признака е повече от един – за многофакторен (многомерен) РА.

Ще покажем как се извършва многофакторен линеен РА, в частност при два фактора. При един фактор и с повече от два фактора се прави аналогично.

Уравнението на регресия при линейния многофакторен РА има вида:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + \ldots + a_mx_m, \qquad (1)$$

където

$\hat{y}$ е теоретичната (оценъчна) стойност на резултативния признак;

x_i, $i= 1, \ldots, m$ са измерените стойности на фактор-признаците;

a_i, $i= 0, 1, \ldots, m$ са коефициентите в уравнението на регресия.

Пример. Наблюдавани са 10 статистически единици (Табл. 1). Нека X_i, $i= 1, 2$ и Y са икономически показатели, като Y ще наречем резултативен (зависим) признак, а X_i, $i= 1, 2$ – фактор-признаци. По тези данни да се определят параметрите в уравнението на линейна регресия и да се направи РА.

Табл. 1

	A	B	C	D
1	№ на набюдаваната величина	зависима величина Y	фактор-признак X_1	фактор-признак X_2
2	1	150	138	400
3		...		
11	10	63	51	165

В този случай, уравнението на линейна регресия има вида:

$$\hat{y} = a_0 + a_1x_1 + a_2x_2. \quad (2)$$

За да определим коефициентите в уравнението на линейна регресия използваме менюто *Tools* и подменюто *Data Analysis*, откъдето избираме Regression и след това бутона *OK*. От отворелия се вече диалогов прозорец задаваме параметрите (Фиг. 1):

1. В полето *Input Y Range* – въвеждаме областта от клетки със стойностите на резултативния признак. В примера чрез селектиране въвеждаме *$B$1:$B$11*.

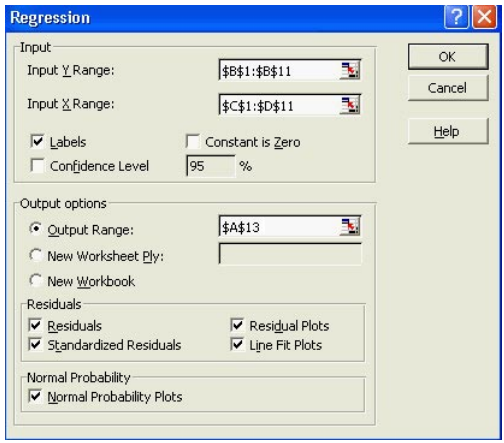
2. В полето *Input X Range* – въвеждаме областта от клетки със стойностите на фактор-признаците. В примера чрез селектиране въвеждаме *$C$1:$D$11*.

3. Ако в горните полета от области сме въвели и етикетите на стойностите, слагаме отметка в полето *Labels*.

4. В раздела *Output options* избираме мястото, където искаме да видим изходните данни. Ако искаме например изходните данни да са в същия работен лист избираме опцията *Output Range* и в полето вдясно указваме горната лява клетка на изходните данни, напр. *$A$13*.

5. В раздела *Residuals* слагаме отметки на *Residuals*, *Standardized Residuals*, *Residual Plots* и *Line Fit Plots*, ако искаме да видим съответно остатъците от регресията, стандартизираните остатъци от регресията, графиките на зависимостта между фактор-признаците и остатъците от регресията и между фактор-признаците и резултативният (зависим) признак.

6. В раздела *Normal Probability* слагаме отметка на *Normal Probability Plots*, за да се изведе на работния лист точната графика на зависимостта между съответните квантили на нормалното разпределение и предсказаните стойности на наблюдавания признак *Y*.



Фиг. 1

Избираме бутона *OK* и на екрана от клетка *A13* се появяват резултатите от РА. Да разгледаме първо таблица *Summary Output* (Табл. 2):

Табл. 2

	A	B
13	SUMMARY OUTPUT	
14		
15	*Regression Statistics*	
16	Multiple R	0,99
17	R Square	0,99
18	Adjusted R Square	0,98
19	Standard Error	7,72
20	Observations	10

Резултатите от тази таблица съответстват на следните статистически величини:

- в клетка *B16* (*Multiple R*) е изобразен корелационният коефициент *R* на Пирсън, който се изчислява по формулата $R = \sqrt{1 - \frac{S_O^2}{S_T^2}} = \frac{S_R}{S_T}$, където S_O^2 е дисперсията за остатъка от регресията. А S_T^2 е сумата от дисперсиите за регресията и за остатъка от регресията, т.е. $S_T^2 = S_R^2 + S_O^2$.
- в клетка *B17* (*R Square*) се намира коефициентът на детерминация R^2. Това е частта от дисперсията на *Y*, изразена чрез X_1 и X_2;
- в клетка *B18* (*Adjusted R Square*) е изобразен изгладеният коефициент на детерминация *R*;
- в клетка *B19* (*Standard Error*) е изобразена общата стандартната грешка на модела, изчислена по формулата $S_O = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2}{n-2}}$, където $\hat{y}_i$ са теоретичните стойности, а y_i са експерименталните (емпиричните) данни.
- в клетка *B20* (*Observations*) е указан броят на наблюденията.

Следващите две таблици (Табл. 3) са под общото название ANOVA (съкращение от *analysis of variance*), което в превод означава Дисперсионен анализ, който се използва за проверка за значимост на коефициента на детерминация R^2.

Табл. 3

	A	B	C	D	E	F	G	H	I
22	ANOVA								
23		*df*	*SS*	*MS*	*F*	*Significance F*			
24	Regression	2	28279,29	14139,64	237,01	3,71772E-07			
25	Residual	7	417,61	59,66					
26	Total	9	28696,90						
27									
28		*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95,0%*	*Upper 95,0%*
29	Intercept	6,08	5,42	1,12	0,30	-6,74	18,89	-6,74	18,89
30	фактор-признак X1	0,73	0,10	7,03	0,0002	0,48	0,97	0,48	0,97
31	фактор-признак X2	0,15	0,04	4,09	0,005	0,06	0,23	0,06	0,23

Стълбовете в първата таблица на табл. 3 могат да се обобщат в следната таблица (Табл. 4):

Табл. 4

	Степени на свобода *df*	Сума от квадратични отклонения *SS*	Дисперсия *MS*	*F* – критерий	Равнище на значимост *Significance F*
Регресия	*m*	$SS_R^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y}_n)^2$	$S_R^2 = \frac{SS_R^2}{m}$	$F_{емп} = \frac{S_R^2}{S_o^2}$	*=FDIST($F_{кр}$;df(Regression); df(Residual))*

Остатък (отклонение) от регресия-та	*n – m - 1*	$SS_o^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$S_o^2 = \frac{SS_o^2}{n - m - 1}$		
Общо:	*n - 1*	$SS_T^2 = \sum_{i=1}^{n} (y_i - \bar{y}_n)^2$ или $SS_T^2 = SS_R^2 + SS_o^2$	$S_T^2 = \frac{SS_T^2}{n - 1}$		

В табл. 4 означенията са:

m – броя на фактор-признаците в уравнението (1) на линейна регресия;

n - броят на наблюдаваните величини;

$\bar{y}_n$ - средната стойност от експерименталните данни.

Стълбовете във втората таблица на табл. 3 се интерпретират по следния начин:

- в стълб *Coefficients* (клетки *B29*, *B30* и *B31*) са разположени стойностите на коефициентите a_i, *i = 0, 1, 2* от уравнението (2).

В случая уравнението на линейна регресия има вида

$$\hat{y} = 6.08 + 0.73x_1 + 0.15x_2, \quad (3)$$

с което се изразява зависимостта на величината *Y* от фактор-признаците X_1 и X_2;

- в стълб *Standard Error* (клетки *C29*, *C30* и *C31*) са разположени стандартните грешки на коефициентите a_i, *i = 0, 1, 2* в уравнението (2);

- в стълб *t Stat* (клетки *D29*, *D30* и *D31*) са изобразени съответните стойности на *t* – статистика по формулата $t_{Stat} = \frac{Coefficient}{Standard\ Error}$;

- в стълб *P value* (клетки *E29*, E*30* и E*31*) е изчислена стойността на равнището на значимост, което съответства на съответната стойност от стълб *t Stat*. Изчислява се с помощта на функцията

$$=TDIST(t_{Stat}; n - m - 1; 2);$$

- в стълбове *Lower 95%* и *Upper 95%* (и съответните им *Lower 95.0%* и *Upper 95.0%*) са изчислени съответно долната и горната граница на доверителния интервал (ДИ) съответно за коефициентите a_i, *i = 0, 1, 2* от уравнението (2). Границите на доверителния интервал се изчисляват по формулата

(*Coefficent – Standard Error*$t_{кр}$; Coefficent + Standard Error*$t_{кр}$*),

където стойността на *t* – критерия $t_{кр}$ се изчислява чрез формулата

$$=TINV(0.05; n - m - 1).$$

Забележка: Стойностите за ДИ в стълбовете *Lower 95.0%* и *Upper 95.0%* са същите като стойностите в стълбовете *Lower 95%* и *Upper 95%*, т.к. сме приели указаното по подразбиране ниво на доверие 0,05, т.е. 95% ДИ.

Следват още две таблици – *Residual Output* и *Probability Output* (Табл. 5).

Табл. 5

	A	B	C	D	E	F	G
35	RESIDUAL OUTPUT					PROBABILITY OUTPUT	
36							
37	*Observation*	*Predicted зависима величина Y*	*Residuals*	*Standard Residuals*		*Percentile*	*зависима величина Y*
38	1	164,53	-14,53	-2,13		5	28
39	2	176,11	3,89	0,57		15	55
40		...				...	
46	9	101,28	-0,28	-0,04		85	180
47	10	67,14	-4,14	-0,61		95	200

Таблица *Residual Output* показва по колони съответно номерата на

наблюдаваните обекти, теоретичната (оценъчна) стойност на зависимия признак $\hat{y}$, остатъчните стойности $y_i - \hat{y}$ и стандартизираните остатъци от регресията.

В таблица *Probability Output* са показани процентите на ДИ и съответните им емпирични стойности Y.

Излизат и няколко графики. Две от тях са графиките на зависимостта между фактор-признаците и остатъците от регресията (*Residual Plots*), които без MS Excel е трудно да се анализират.

След това се прави анализ на получената таблица, като се прави проверка за адекватност на построеното уравнение на линейна регресия на няколко етапа.

Първи етап: Стойността на коефициента на детерминация $R^2=0.99$ (клетка B17 в табл. 2) показва, че 99% от общата вариация на резултативния признак се обяснява с фактор-признаците X_1 и X_2. Това означава, че избраните фактори X_1 и X_2 съществено влияят на резултативния признак Y, т.е. това потвърждава за правилността на техните включвания в построения модел. Изчисленото равнище на значимост (клетка *F24*) потвърждава значимостта на коефициента R^2.

Втори етап: проверка за значимост на коефициентите в уравнението (3) (табл. 3). Виждаме, че *P-value* за нулевия коефициент a_0 (клетка *E29*) е извън критичната област за нулевата хипотеза, т.е. a_0 не е статистически значим, а за коефициентите a_1 и a_2 (клетки *E30* и *E31*) сме в критичната област за нулевата хипотеза, т.е. те са статистически значими.

Трети етап: От диалоговия прозорец *Regression* поставяме отметка на *Constant is Zero*, а другите параметри ги задаваме същите. В случай, че незначим се окаже коефициент от фактор-признаците трябва да се преразгледа избора им в уравнението на линейна регресия. След избиране на бутона *OK* се появяват таблици, от които се вижда, че така полученото линейно регресионно уравнение $\hat{y} = 0.77x_1 + 0.15x_2$ е целесъобразно.

ЗАКЛЮЧЕНИЕ

Коефициентите a_1 и a_2 позволяват да се направят следните изводи: с увеличаването на фактор-признака X_1 с една единица (лв; млн. и др.) води до увеличаване на резултативния признак с 0.77 единици, а увеличаването на фактор-признака X_2 с една единица (лв; млн. и др.) води до увеличаване на резултативния признак с 0.15 единици.

Благодарности. Работата е финансирана по проект НИП №РД-05-285/11.03.2009 на Шуменски Университет.

ЛИТЕРАТУРА

[1] Levine,D., M.Berenson, D.Stephan. Statistics for Managers using Microsoft Excel. Prentice-Hall, New Jersey, 1999.

[2] Winston,W.. Microsoft Excel Data Analysis and Business Modeling. Microsoft Press, 2004.

За контакти:

Ас. Красимира Янкова Костадинова, Катедра “Икономика и моделиране”, ФМИ, Шуменски университет “Епископ Константин Преславски”, тел.: 054-830 495, вътр. 138, e-mail: kostadinova_kr@abv.bg

Докладът е рецензиран.