

ПРИМЕРНО ПРИЛОЖЕНИЕ НА ЕДИНИЧНА ЛИНЕЙНА РЕГРЕСИЯ И КОРЕЛАЦИЯ В ТРАНСПОРТНИТЕ ИЗСЛЕДВАНИЯ С ИЗПОЛЗВАНЕ НА MS EXCEL

Ставри Димитров

stavri_dimitrov@hotmail.com

*Висше транспортно училище “Тодор Каблешков”,
катедра “Технология, организация и управление на транспорта”,
ул. “Гео Милев” 158, София 1574, БЪЛГАРИЯ*

Ключови думи: приложение, статистически методи, корелационен анализ, регресионен анализ, единична (проста) линейна регресия, регресионно уравнение, научни изследвания, ms excel, транспорт

Резюме: Публикацията е допълнение към [1] и представя пример за приложението на статистическите методи корелационен и регресионен анализ в транспортните изследвания с използване на Microsoft® Excel[5]. Чрез разгледания пример са показани предимствата от използване на съвременни изчислителни програми и средства, каквито са електронните таблици, като възможна алтернатива за прилагане на обсъжданите в [1] и в настоящата публикация статистически методи при установяване наличието на връзки и зависимости между случайни величини.

1. Увод.

Заклучителен в провежданите научни изследвания се явява етапа на обработка на събраните от изследването данни и последващ анализ на получените резултати. За успешното приключване на този етап, изследователи и научни работници си служат с някои от съществуващите статистически методи[3]. Приложението на статистически методи като регресионен и корелационен анализ е свързано със значителен обем изчисления. Извършването на т.нар. ръчни изчисления в процеса на приложението на горните методи [2,3,4] е предпоставка за забавяне в извеждането и публикуването на крайните резултати от изследването, а оттам и до забавяне на внедряването на научните постижения в практиката. С течение на времето този факт породил острата необходимост от разработване и използване за нуждите на научната общност на програмни продукти, които значително улесняват събирането, обработката, изчисленията и съхраняването на статистическите данни.

2. Предпоставки и начини за разрешаване на проблема.

Развитието на електронно-изчислителната техника и софтуерните продукти непрекъснато продължава да разкрива нови възможности пред учени, докторанти и студенти. За успешното извършване на изчисленията в своите изследвания, вече не е необходимо да използват калкулатори, било то и научни, тъй като в зависимост от потребностите, възможностите, предпочитанията и преследваните цели, могат да избират измежду редица софтуерни офис приложения и специализирани софтуерни пакети. Една от масово използваните от широк кръг потребители програми за математически изчисления и анализи от различно естество е офис приложението Microsoft® Excel [5] в различните му версии. В електронната таблица MS Excel са вградени голям брой функции и са заложили разнообразни

функционалности, чиято мощ позволява с минимални усилия и за кратко време да се решават различаващи се по своята сложност и размерност задачи от различни области, в т.ч. и в областта на транспорта. Програмата *Excel* дава възможност с голяма лекота да се приложат на практика вградените в нея статистически методи – описателна статистика, тестове за съответствие на закони за разпределение, анализ на дисперсията, корелационен и регресионен анализ – и да се извършват свързаните с тях анализи на данните. Изброените предимства и достъпността на програмата я правят предпочитана от специалисти в различни научни области.

3. Решение на проучения проблем.

Пример за използването на *MS Excel* за обработка и анализ на данни в научните изследвания е приложението на статистическите методи еднофакторна линейна регресия и корелация [2,3] в областта на транспорта. За целите на примера, в настоящия доклад е обсъден въпроса по изследване за наличието на връзка и установяване на аналитична зависимост между двете променливи случайни величини Y – средно време за пътуване от местоживее до месторабота с личен лек автомобил на територията на даден град по време на час “пик” във функция от броя на регистрираните лични леки автомобили – X . Това е въпрос, който подробно е разгледан в [1], където корелационния и регресионния анализи и получените вследствие на тяхното приложение резултати са получени чрез извършване на ръчни изчисления.

Основната цел на настоящия доклад е посредством разгледания в него пример, ползвайки примерните статистически данни от [1], показани в табл. 1, да представи алтернативни начини за разрешаване на проблема по установяване на връзки и зависимости между случайни величини. Акцентът в доклада пада главно не върху ръчното, както е в [1], а върху автоматизирано с помощта на *MS Excel* извършване на изчисленията, с които е съпроводено приложението на регресионния и корелационния анализи, като изцяло се набляга на практическото приложение на тези два метода в транспорта. Докладът представлява своеобразно допълнение към [1] относно употребата на съвременни средства и инструменти като възможни алтернативи за приложението на обсъжданите статистически методи при решаване на дефинираните задачи.

Таблица 1 - Примерни данни

Година по ред i	Средно време за пътуване до работа с индивидуален транспорт Y_i , минути	Лични леки автомобили X_i , хил. бр.
1	13,0	262,5
2	13,3	273,0
3	16,5	286,0
4	17,4	301,5
5	17,9	319,0
6	21,4	337,0
7	22,1	356,0
8	24,0	377,0
9	27,3	400,0
10	32,6	427,5

С данните от табл. 1, чрез *MS Excel* последователно са проведени корелационен и регресионен анализ за установяване наличието на връзка между изследваните променливи Y и X и изчисляване на коефициента на линейна корелация r_{YX} от една страна, и от друга – изчисляване на стойностите на регресионните коефициенти b_0 и b_1 на линейния еднофакторен регресионен модел имащ вида

$$(1) \hat{Y}_i = b_0 + b_1 \cdot X_i$$

избран за математическо описание на връзката между зависимата променлива Y и факторната променлива X :

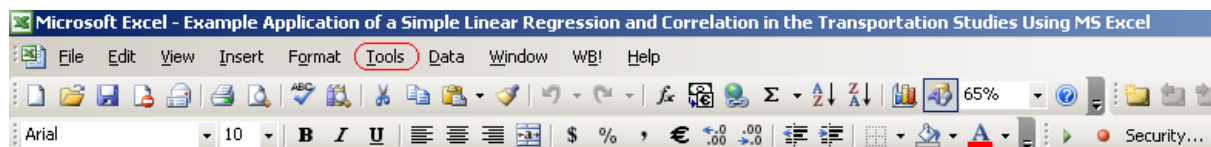
3.1. Корелационен анализ.

Стойностите на Y и X (табл. 1) служат за попълване на табл. 2, показана на фиг. 1 като част от използвания за изчисленията работен лист в средата на Excel.

i	X_i	Y_i	$(X_i - \bar{X})$	$(Y_i - \bar{Y})$	$(X_i - \bar{X}) \cdot (Y_i - \bar{Y})$	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	$X_i \cdot Y_i$	X_i^2	Y_i^2	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \hat{Y}_i)^2$
1	262,5	13,0	-71,5	-7,6	539,4	5105,103	57,003	3413	68906	12,661	62,229	0,115
2	273,0	13,3	-61,0	-7,3	441,9	3714,903	52,563	3631	74529	13,821	45,283	0,271
3	286,0	16,5	-48,0	-4,1	194,2	2299,203	16,403	4719	81796	15,256	28,026	1,547
4	301,5	17,4	-32,5	-3,2	102,2	1053,003	9,923	5246	90902	16,967	12,836	0,187
5	319,0	17,9	-15,0	-2,7	39,6	223,503	7,023	5710	101761	18,899	2,724	0,999
6	337,0	21,4	3,1	0,8	2,6	9,303	0,722	7212	113569	20,887	0,113	0,263
7	356,0	22,1	22,1	1,6	34,2	486,203	2,403	7868	126736	22,984	5,927	0,782
8	377,0	24,0	43,1	3,5	148,5	1853,303	11,903	9048	142129	25,303	22,591	1,698
9	400,0	27,3	66,1	6,8	445,8	4362,603	45,563	10920	160000	27,842	53,178	0,294
10	427,5	32,6	93,6	12,1	1127,3	8751,603	145,203	13937	182756	30,878	106,678	2,964
Сума:	3340	206			3075,775	27858,725	348,705	71703	1143085	205,500	339,585	9,120

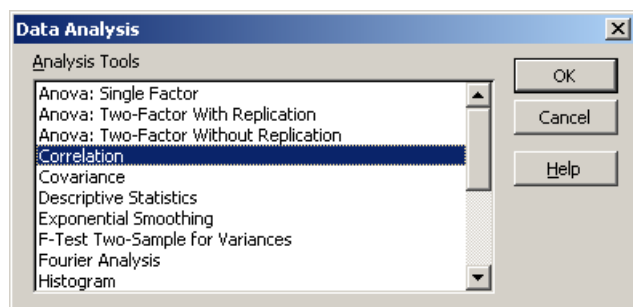
Фиг. 1. Изчислителна таблица

Изчисляването на коефициента на линейна корелация r_{YX} става, след като от меню "Tools" (Инструменти) (фиг. 2) се избере "Data Analysis..." (Анализ на данни).

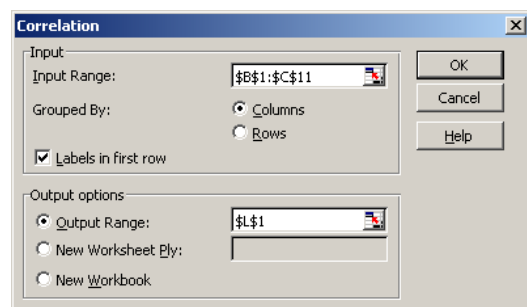


Фиг. 2. Избор на меню "Инструменти"

В отвореният се прозорец "Data Analysis" се избира функцията "Correlation" (фиг. 3), след което се натиска бутон "OK". В резултат на това действие се отваря прозореца "Correlation" (Корелация) (фиг. 4), в който в



Фиг. 3. Прозорец "Анализ на данни"



Фиг. 4. Прозорец "Корелация"

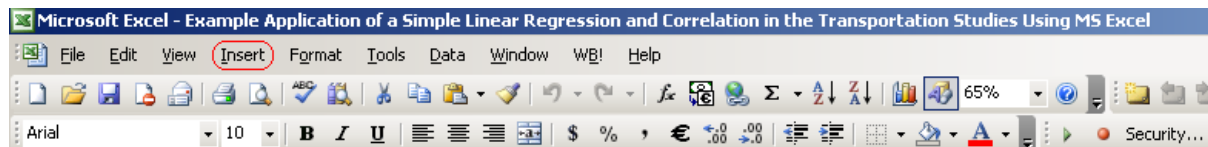
checkbox "Labels in first row" се поставя отметка, а в полето "Input Range" се избира обхвата от стойности за X_i и Y_i (табл. 2), включително със заглавния ред. В случай, че потребителят желае корелационната матрица да се изведе в текущата страница (worksheet), избира радио бутон "Output Range" и посочва адреса на клетката, в която ще се разположи корелационната матрица. Накрая, за извършване на изчисленията по пресмятане на стойността на коефициента r_{YX} и извеждане на резултатите във вид на таблица (табл. 3), се натиска бутона "OK".

Таблица 3 – Корелационна матрица

	X_i	Y_i
X_i	1	
Y_i	0,987	1

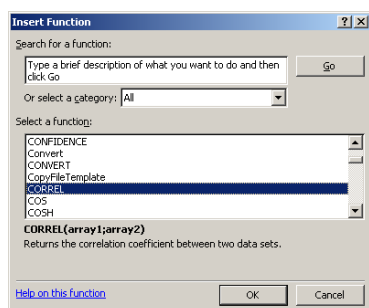
Полученият в *Excel*, чрез използване на функцията “*Correlation*” коефициент на линейна корелация е равен на $r_{YX} = 0,987$.

Коефициентът r_{YX} може да се пресметне и чрез използване на вградената в *Excel* функция “*CORREL*” по следния начин: След като предварително е маркирана клетката, в която програмата *Excel* ще запише изчислената стойност, от меню “*Insert*” (фиг. 5) се избира “*Function...*”(Функция).

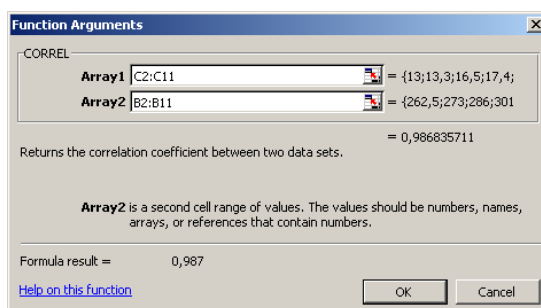


Фиг. 5.Избор на меню “Insert”

Това довежда до отваряне на прозорец “*Insert Function*” (фиг. 6), в който в текстовото поле “*Select a function:*” се избира желаната функция. В случая това е функцията “*CORREL*”. Натиска се бутон “*OK*”. В отвореният прозорец “*Function Arguments*” (фиг. 7) се въвеждат обхватите от клетки, съдържащи стойностите на променливите Y и X и се натиска бутон “*OK*”.



Фиг. 6.Прозорец “Въведи функция”



Фиг. 7.Прозорец “Аргументи на функцията”

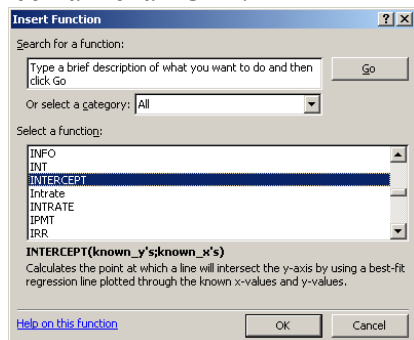
Както може да се види от фиг. 7, стойността $r_{YX} = 0,987$ съответства на тази, получена чрез ползване на функционалността “*Correlation*” (табл. 3).

3.2. Регресионен анализ.

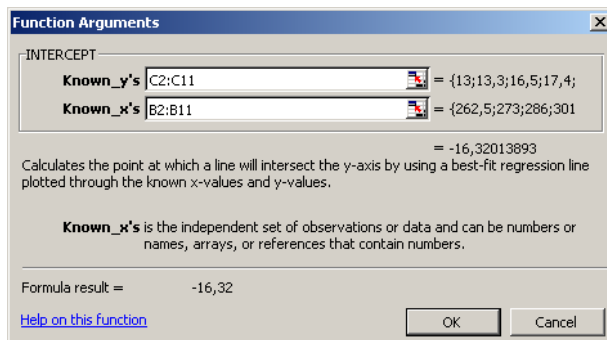
Изчисляването на стойностите на регресионните коефициенти b_0 и b_1 за избраното линейно уравнение на регресия, съдържащо една независима променлива, може да се извърши чрез използване на:

3.2.1. Функции “*INTERCEPT*” и “*SLOPE*”.

В средата на *Excel* се избира клетката, в която функцията ще върне като резултат изчислената от нея стойност b_0 , например клетка “*P6*”. От меню “*Insert*” се избира “*Function...*”. Отваря се прозорец “*Insert Function*” (Въведи Функция), в който се маркира функцията “*INTERCEPT*” (Отрязък) (фиг. 8). Натиска се бутон “*OK*” и в резултат на това се отваря прозорец “*Function Arguments*” (Аргументи на функцията), в който последователно се маркират, без да се включва заглавния ред, обхватите от клетки от табл. 2, съдържащи стойности за Y_i (*known_y's*) и X_i (*known_x's*) (фиг. 9), които са аргументи на функцията и накрая се натиска “*OK*”.



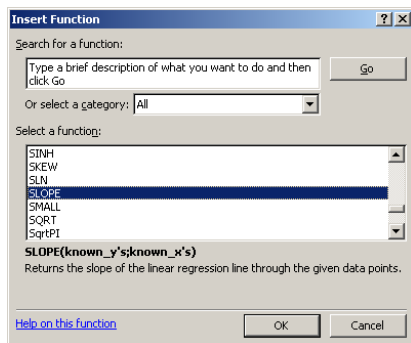
Фиг. 8.Прозорец “Въведи функция”



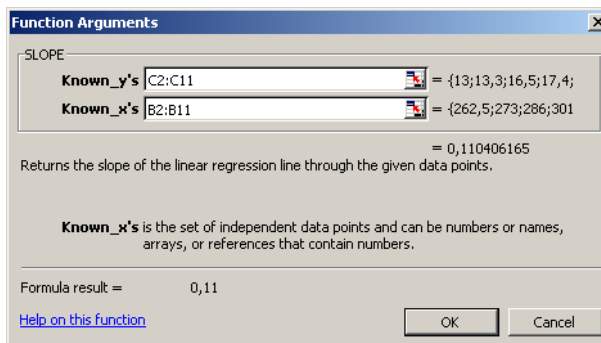
Фиг. 9.Прозорец “Аргументи на функцията”

Получената чрез функцията “*INTERCEPT*” стойност е $b_0 = -16,32$.

За изчисляване на стойността на регресионния коефициент b_1 пред фактора X_i , аналогично се отива в меню “*Tools*” -> “*Function...*”, от където в появилият се прозорец “*Insert Function*” се избира функцията “*SLOPE*” (Наклон) (фиг. 10), след което се натиска бутона “*OK*”. Отваря се прозореца “*Function Arguments*”, в който отново се посочват обхватите от стойности *known_y's* за Y_i и *known_x's* за X_i (фиг. 11), явяващи се аргументи на функцията, не включвайки при избора заглавния ред от табл. 2.



Фиг. 10. Прозорец “Въведи функция”

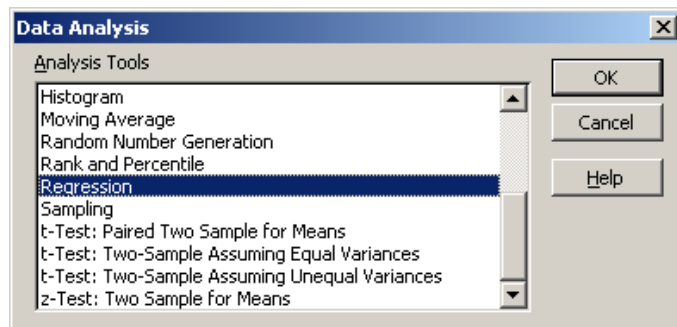


Фиг. 11. Прозорец “Аргументи на функцията”

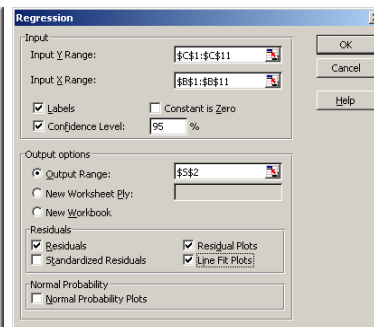
Върната от функцията “*SLOPE*” след завършване на изчисленията стойност на регресионния коефициент е $b_1 = 0,11$.

3.2.2. Функционалност “*Regression*”.

Регресионните коефициенти b_0 и b_1 за използвания за целите на примера линеен еднофакторен регресионен модел, коефициента на корелация R и коефициента на детерминация R^2 , както и емпиричната характеристика F_{em} могат да се определят с извършване на регресионен анализ с помощта на *Excel*, ползвайки функционалност “*Regression*” (Регресия), достъпна от меню “*Tools*”-> “*Data Analysis...*”->прозорец “*Data Analysis*” (фиг. 12). За целта се маркира функция “*Regression*” и натиска бутона “*OK*”. Следва отваряне на прозорец “*Regression*” (фиг. 13), в който в checkbox “*Labels*” се поставя отметка, а в полетата “*Input Y Range:*” и “*Input X Range:*” се указват обхватите от клетки от

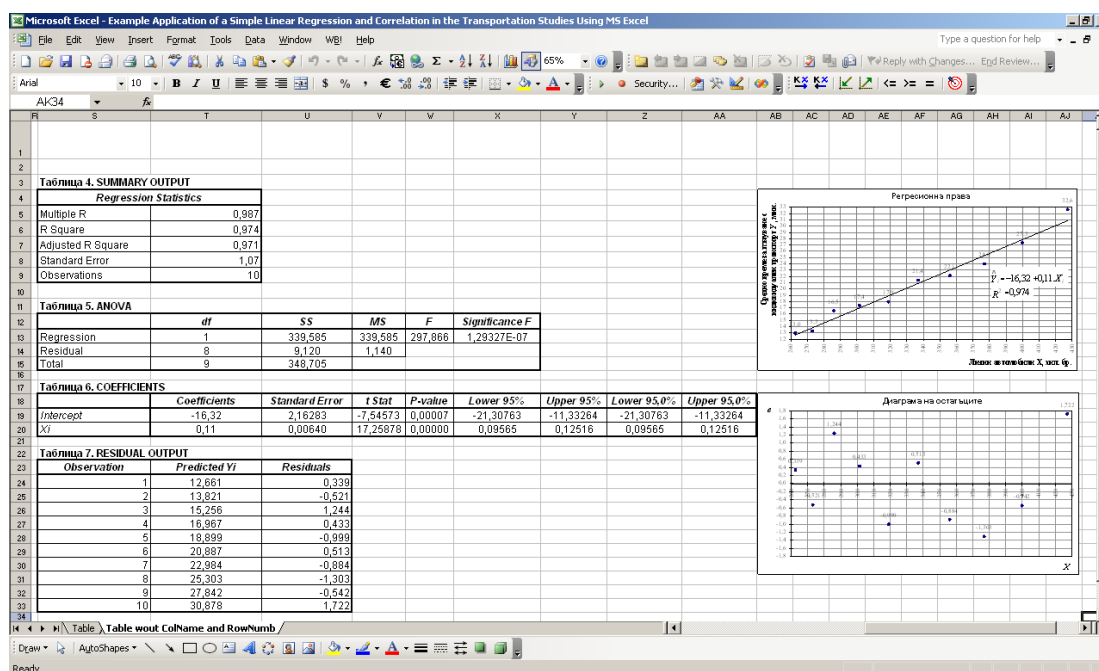


Фиг. 12. Прозорец “Анализ на данни”



Фиг. 13. Прозорец “Регресия”

табл. 2, в т.ч. и заглавния ред, съдържащи стойности за X_i и Y_i . Поставя се и отметка за извеждане на данни за избрания доверителен интервал (“*Confidence Level:*”). В рамката “*Residuals*” (остатъци) се поставят отметки съответно за “*Residuals*” (за таблица с остатъците), “*Residual Plots*” (за диаграма на остатъците) и “*Line Fit Plots*” (за диаграма с емпирични и теоретични стойности). Последното действие се състои в избиране на радио бутон “*Output Range:*” (обхват с резултати), в който се указва, че резултатите от изчисленията да бъдат изведени в текущата страница (*worksheet*), в посочената клетка (\$D\$2). За стартиране на изчислителните процедури и извеждане на крайните резултати от изчисленията във вид на таблици и диаграми (фиг. 14) се натиска бутон “*OK*”.



Фиг. 14. Резултати от регресионния анализ

Последователно в таблици от 4 до 7 са визуализирани следните резултати:

- в табл. 4. *SUMMARY OUTPUT* – стойностите на коефициента на корелация R , коефициента на детерминация R^2 (R Square), коригирания коефициент на детерминация ($Adjusted$ R Square), стандартната грешка на оценката на модела ($Standard$ $Error$) и брой на наблюденията n ($Observations$);

- в табл. 5. *ANOVA* – степените на свобода ν и w от числителя и знаменателя на F – отношението [3] (колона “ df ” – *Degrees of Freedom*), обяснената SSR и остатъчната девиация SSE (колона “ SS ” – *Sum of Squares*) [3], както и тяхната сума SS_{YY} (колона “ SS ”), числителят и знаменателят на F – отношението (*Mean Squares*), емпиричната характеристика F_{em} на F – теста и наблюдаваното равнище на значимост, съответстващо на F (*Significance F*);

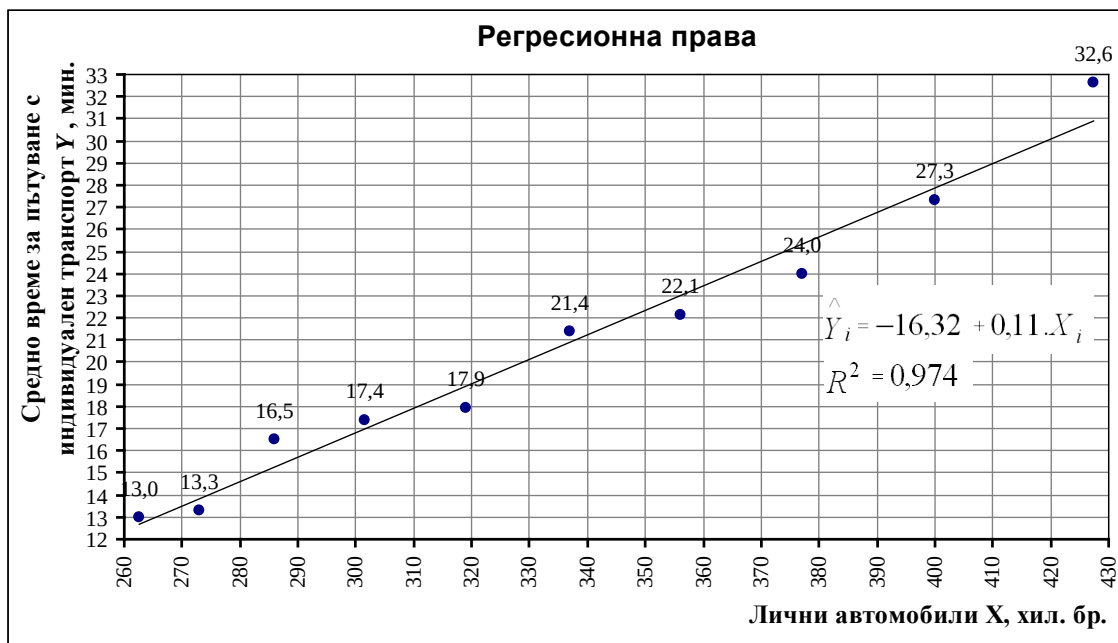
- в табл. 6. *COEFFICIENTS* – стойностите на регресионните коефициенти b_0 и b_1 (*Coefficients*), стандартните грешки на оценките на регресионните коефициенти (*Standard Error*), емпиричните характеристики за проверка на хипотезите за статистическа значимост на регресионните коефициенти (t Stat), наблюдаваните равнища на значимост (P -value), съответстващи на емпиричните характеристики за проверка на хипотезите за b_0 и b_1 , долна и горна граници на доверителните интервали (*Lower 95 %* и *Upper 95 %*) на коефициентите b_0 и b_1 ;

- табл. 7. *RESIDUAL OUTPUT* – наблюдавани стойности Y_i (*Observations*), предвидени стойности $\hat{Y}_i$ (*Predicted*) и стойностите на остатъците e_i (*Residuals*).

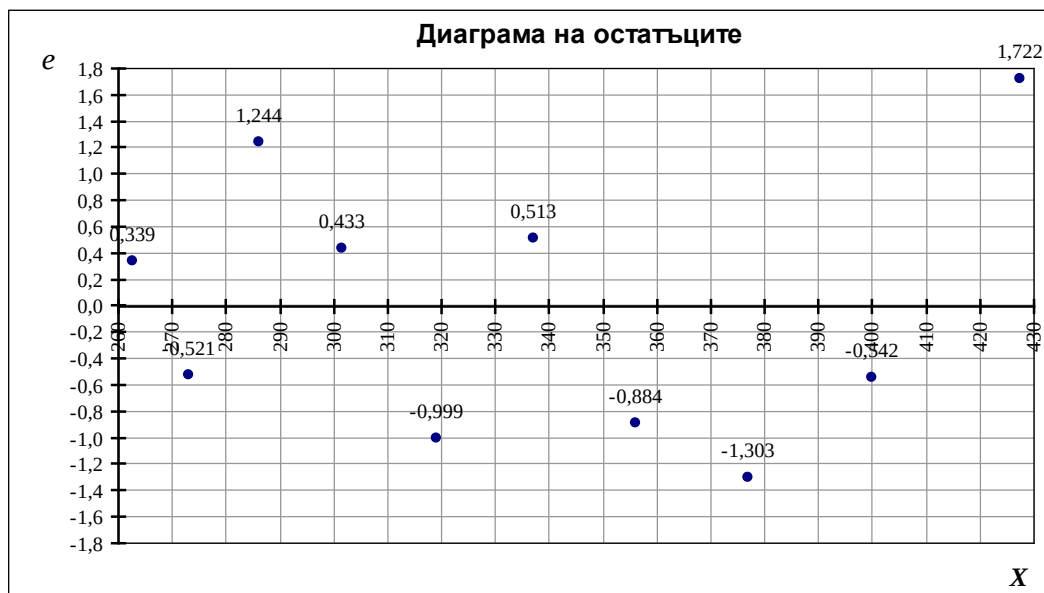
След заместване на изчислените регресионни коефициенти в линейното еднофакторно уравнение на регресия получаваме:

$$(2) \quad \hat{Y}_i = -16,32 + 0,11 \cdot X_i$$

Диаграмите с предвидените стойности (*Line Fit Plots*) и остатъците (*Residual Plots*) освен на фиг. 14, са показани съответно и на фиг. 15 и фиг. 16.



Фиг. 15.Регресионна права



Фиг. 16. Остатъци

4. Резултати и дискусия.

В резултат на извършените с използване на *MS Excel* корелационен и регресионен анализ на взетите от [1] примерни статистически данни се:

1. Установи наличие на корелация между средното време за пътуване с индивидуален транспорт Y и броят на личните автомобили X , за силата на която може да се съди от стойността на коефициента на линейна корелация $r_{YX} = 0,989$.

2. Установи видът и определиха стойностите на регресионните коефициенти $b_0 = -16,32$ и $b_1 = 0,11$ на уравнението на регресия, отразяващо връзката между зависимата променлива Y и фактора X .

3. Установи, че избрания линеен еднофакторен регресионен модел адекватно и със задоволително висока точност моделира връзката между зависимата променлива Y и независимата променлива X , за което сочат резултатите от извършения от *Excel F* – тест за адекватност на модела [3] и близката до единица стойност на коефициента на детерминация $R^2 = 0,974$.

5. Заключение.

Резултатите, получени от приложението на статистическите методи корелация и регресия, чрез използване на програмата *MS Excel* при решаване на задачата в разгледания в настоящия доклад пример, съответстващи на получените в [1] резултати, дават основание да се направят следните изводи:

1. Използването в научните изследвания на офис приложения като *Excel*, за извършване на изчисления и анализи, има следните предимства:

- те са общодостъпни и лесни за работа програмни продукти;
- вероятността от допускане на човешки грешки е малка, тъй като изчисленията се извършват по заложи в програмата изчислителни алгоритми;
- изчисленията се извършват бързо, от порядъка на няколко секунди;
- постига се висока точност на получаваните резултати;
- крайните резултати могат да се извеждат във вид на таблици и диаграми;
- вградените в тях функции предоставят възможност на потребителя, използвайки различни техники по няколко начина да достигне до един и същи резултат.

2. При използване на офис приложения за извършване на статистически анализи трябва да се имат в предвид и някои особености:

- за да е в състояние да интерпретира получените на изхода на “черната кутия” резултати, още преди да въведе данните на входа ѝ, потребителят трябва да е наясно с теоретичната постановка на провеждания от него анализ, както и в какво се състоят и как се извършват използваните изчислителните процедури;

- възможно е в някои програми с общо предназначение, каквато е *Excel*, да не е предвидено вграждане на функционалности за приложение на познати и използвани в математическата статистика методи за анализ на данни. Затова, когато използваните офис приложения не успяват да удовлетворят потребностите на изследователите и специалистите, алтернатива могат да бъдат създадените за тази цел специализирани статистически софтуерни пакети.

ЛИТЕРАТУРА

[1] Димитров, С. Д., “Примерно приложение на единична линейна регресия и корелация в транспортните изследвания”, научно списание “Механика, транспорт, комуникации”, ВТУ “Тодор Каблешков”, София, 2010 г. (под печат)

[2] Качаунов, Т. Т., “Моделиране и оптимизация на транспортните процеси”, второ преработено издание, Печатница при ВТУ “Тодор Каблешков”, София, 2005 г.

[3] Манов, А., “Статистика със SPSS”, второ издание, Издателство “Тракия – М”, София, 2001 г.

[4] Рихтер, К. Ю., Фишер, П., Шнейдер, Г., “Статистические методы в транспортных исследованиях”, перевод от немски, Москва, “Транспорт”, 1982г.

[5] Microsoft Excel 2003 Help, Microsoft Corporation