Statistics for Data Analysis Using Microsoft Excel

Chief Assist. Prof. Dr. Svetlana Todorova
University of Economics – Varna, Varna, Bulgaria
svetlana.todorova@ue-varna.bg

Abstract

Today business research is based on the manipulation of large amounts of data from the Internet, company sources, public sources, and others. Business analytics must extract useful information from a big data set and to make a decision. Whether you are analyzing a client's data or company's data to make decisions, your tools have to be able to handle the tasks you perform with that information. The goal of this paper is to illustrate the use of Excel's Analysis ToolPak for developing complex statistical analysis. MS Excel is widespread, user-friendly, easy to learn, and powerful. Therefore, the purpose of this paper is to show how free Analysis ToolPak add-in works and what it can do. In one hand Analysis ToolPak is powerful, but in the other hand it has some weaknesses that we should be aware of.

Keywords: Statistics, Data Analysis, Excel, Analysis ToolPak

JEL Code:C880; doi:10.36997/IJUSV-ESS/2019.8.2.68

Въведение

Съвременните бизнес изследвания се основават на извличане на големи масиви от данни от Интернет, фирмени източници, публични източници, статистически справочници и др. Статистическият анализ на данни, които включват много на брой променливи и наблюдения, изисква от изследователя задълбочени знания за статистическите методи. В повечето случаи обработката на такива данни се реализира с помощта на специализиран софтуерен продукт. В широк смисъл статистически е всеки софтуер, който е в състояние да обработва съвкупностна (статистическа) информация и притежава възможности за изчисляване на обобщаващи характеристики. В тесен смисъл статистически е всеки софтуер, който притежава възможности за изчисляване на обобщаващи характеристики от основните дялове на теоретичната статистика – анализ на емпирични разпределения, статистически заключения, анализ на корелационни зависимости и др. (Хаджиев, 2009, p. 9-10).

MS Excel е широко използван софтуер за статистически анализ на данни, представени във вид на електронни таблици. Намира приложение и при обработка на големи масиви от данни. Той предоставя възможности за графично изобразяване на данните и дава възможност за реализиране както на основни статистически функции, така и на комплексен статистически анализ с модула Data Analysis. MS Excel е добър избор, защото той е леснодостъпен, спестява време, осигурява лесен обмен на данни с други приложения, визуализира много добре графично и таблично крайните резултати и др.

Всяка следваща версия на MS Excel добавя нови възможности за графично представяне и интерпретация на данни: Box & Whisker, Histogram, Scatter (X, Y), Trendline и др. В MS Excel съществуват две възможности за приложение на статистически методи: основни статистически функции и модула Data Analysis. Според предназначението си функциите са класифицирани в 13 категории. Всяка функция е достъпна от раздел Formulas, група Statistical Function (Сълов и др., 2017 p. 218-222). Обработката на данните по този начин се извършва чрез вградени функции, които участват във формули за определяне на съдържанието на дадена клетка. Модулът Data Analysis съдържа 19 инструмента за различни статистически анализи и тестове, които позволяват извършването на един по-комплексен статистически анализ.

Наред с новите диаграми, Microsoft постоянно обновява и статистическите функции на MS Excel. Те стават все по-точни и все повече задоволяват нарастващото потребителско търсене. От друга страна функционалността и потребителският интерфейс на Analysis

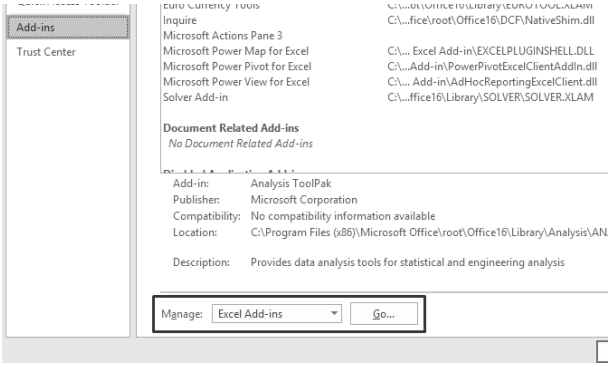
ToolPak са същите от десетилетия. Любопитно е, че предвид все по-голямото внимание на Microsoft към анализа на данни и съсредоточаването му върху подобряване на функциите и добавянето на нови диаграми за анализ на данни, защо запазва Analysis ToolPak такъв, какъвто е.

Целта на този доклад е, въпреки някои от слабите стрни на Analysis ToolPak, да се илюстрира лекотата и достъпността на Analysis ToolPak за разработване на комплексен статистически анализ.

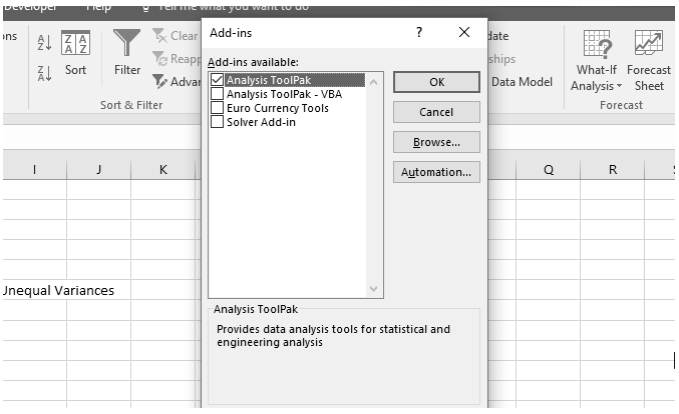
1. Инсталиране на модула Data Analysis ToolPak на MS Excel

В настоящата статия се ограничаваме до описание на възможностите на специализирания модул Data Analysis в MS Excel. Той допълва наличната функционалност на електронната таблица, като добавя към възможностите й да представя и трансформира динамични редове и мощни възможности за извършване на статистически анализ (Славева и др., 2016, р. 522-523). Основни статистически анализи и тестове в Data Analysis са: описателна статистика, дисперсионен анализ, регресионен анализ, корелационен анализ, метод на плъзгащите се средни, проверка на статистически хипотези и други.

Преди да се започне с използването на Analysis ToolPak, той трябва да бъде зареден, чрез add-ins: За целта се избира от „Опциите” на Excel (в Excel 2010 или по-нови версии):

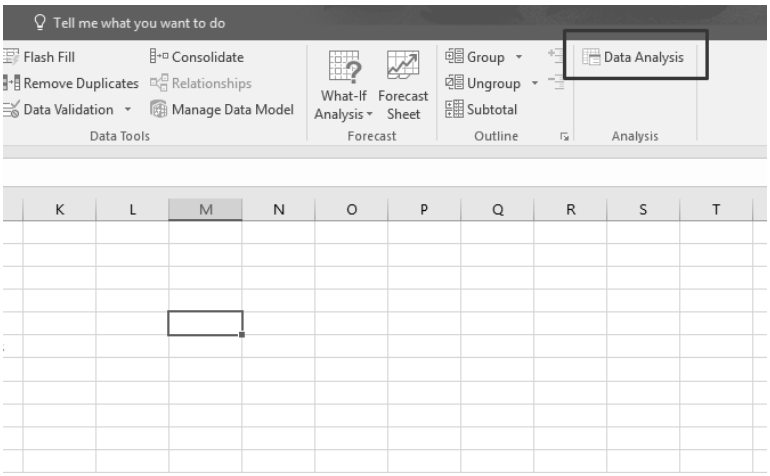


Add-ins, Go и след това Add-ins ToolPak:

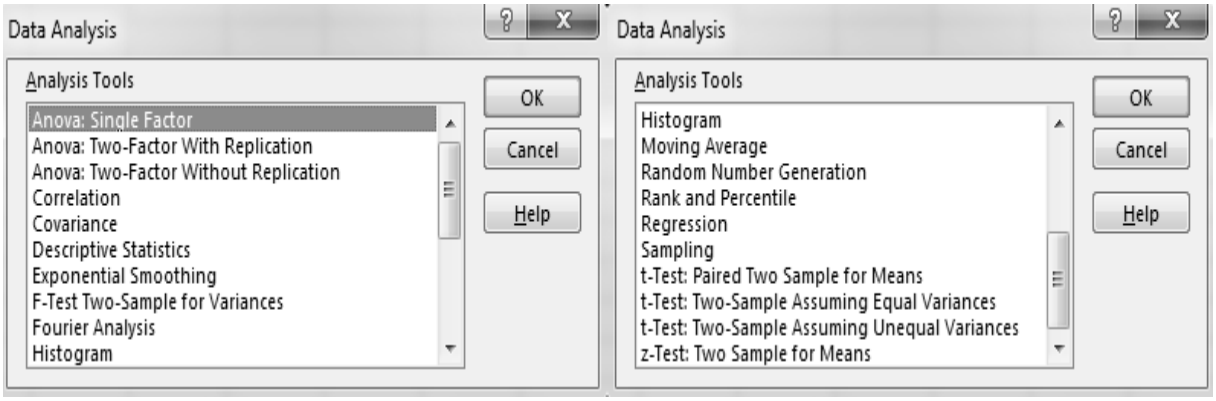


Трябва да се обърне внимание, че има и Analysis ToolPak - VBA. Версията VBA включва Visual Basic за разработване на приложения на основата на създаване на макроси. Ако не планирате да автоматизирате Analysis ToolPak с VBA, той няма нужда да се инсталира.

При правилно добавяне на Analysis ToolPak в Excel, на лентата с команди в раздела Data се появява бутон Data Analysis:

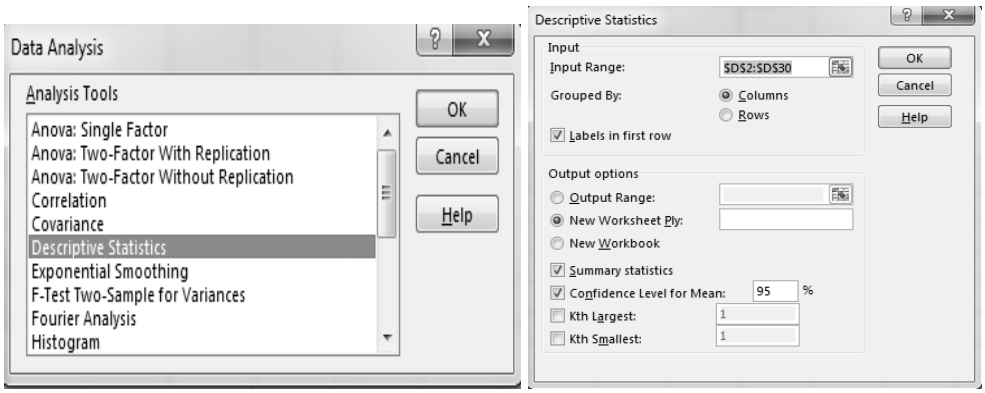


След като модулът за анализ на данни е добавен се появява Data Analysis в крайната дясна част на лентата. При избор на Data Analysis ще се отвори диалоговия прозорец, който предлага достъп до различни статистически анализи и тестове:



2. Описателна статистика в MS Excel

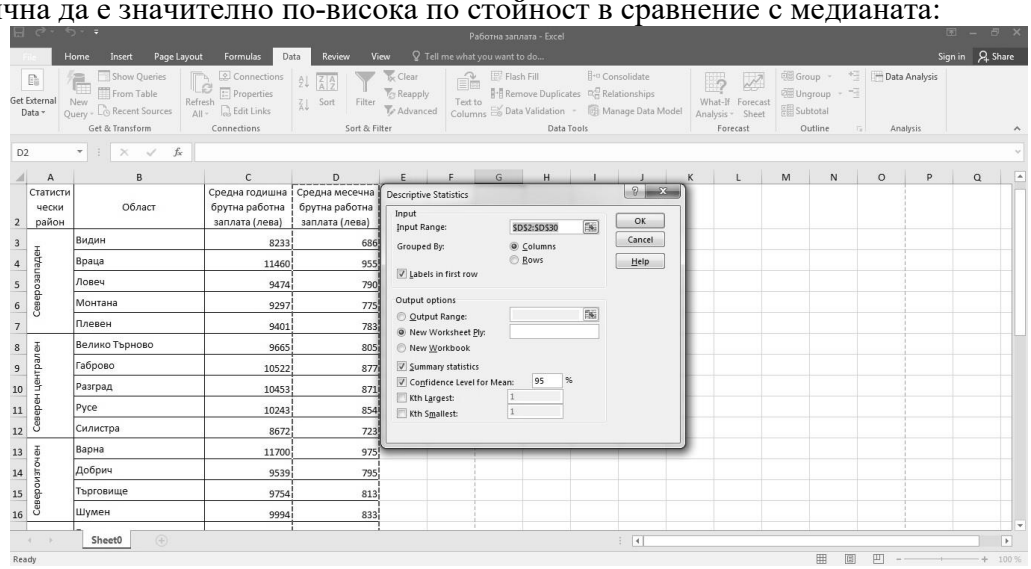
Независимо какъв статистически анализ или тест ще се провежда, почти винаги първо е необходимо да се получат обобщаващи числови характеристики, графики и таблици (описателната статистика). Това дава информация за средната, медианата, модата, стандартното отклонение, дисперсията, асиметрията и ексцеца на емпиричното разпределение на данните, с които се разполага. Изпълнението на описателната статистика в MS Excel е лесно и бързо. Избира се Data Analysis в раздела Data и след това Descriptive Statistics като данните се въвеждат чрез маркиране. Може да се окаже на Excel дали данни имат етикети, дали резултатите от анализа да са на нов лист или на същия и други опции:



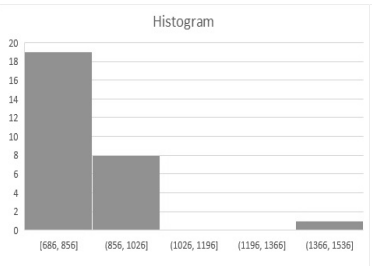
Струва си да се спомене, че ако се избере Confidence Level (равнище на значимост), при проследяване на резултатите от анализа на последния ред се появява число, чиято стойност трябва да се извади и прибави към средната, за да се получи 95% доверителен интервал (по подразбиране е заложено 95% равнище на значимост). Следователно, за да се получи доверителен интервал, се нужни допълнителни изчисления.

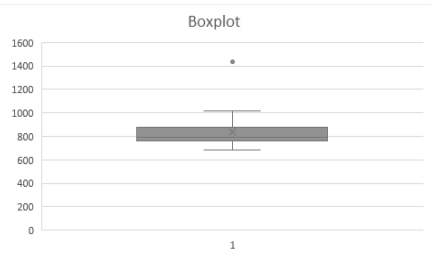
Опцията Histogram в Analysis ToolPak позволява да се създаде честотна таблица и съпътстваща диаграма - хистограма. В диалоговия прозорец на Histogram се изисква не само диапазона на данни, но също така и диапазона Bins, но за него няма стойности по подразбиране. Потребителят избира Bins като го поставя някъде в работния лист. Това всъщност е горната граница на интервалите при групирани данни. По подразбиране резултатите се появяват на нов работен лист и включват таблицата с честотните разпределения и съответната диаграма. За да се направи така, че стълбчетата на хистограмата да са точно долепени едно до друго, се избира форматиране на диаграмата и Gap Width = 0. Трябва да се обърне внимание, че слабостите на хистограмата от Analysis ToolPak могат да се преудолеят ако се използва новата вградена диаграма, която за първи път се въвежда в Excel 2016. За да създадете хистограма в Excel 2016, трябва само да се маркират данни, да се избере от раздела Вмъкване на групата Графики и да се намери Хистограма. Получава се хистограмата без да е необходимо дефиниране на долна или горна граница на интервалите и не се налага редактирането и. Добрата новина е, че Box & Whisker plot е нов вграден тип диаграма, представена също в Excel 2016. За да създадете Box & Whisker plot, се следват същите стъпки като се избира Box & Whisker. Освен това всяка диаграма в MS Excel може да се персонализира чрез Дизайн / Формат. Статистическите диаграмите позволяват по-добро възприемане на зависимостите, емпиричното разпределение или тенденциите в развитието.

За представяне на описателната статистика използваме реални данни за средната месечна брутна работна заплата по области за 2017 г. Резултататите от анализа могат да са на нов работен лист (по подразбиране) или на същия работен лист ако се зададе поле, в което да се поместят. Средната месечна работна заплата според средната аритметична е 840 лв., като най-ниска е в област Видин – 686 лв. и най-висока в София – град – 1433 лв. Епиричното разпределение е несиметрично, с ясна изразена положителна асиметрия (Histogram & Boxplot). Това показва, че за определяне центъра на разпределение е по-подходящо да се използва медианата, която е 795 лв. Това означава, че 50% от населението е с месена заплата по-малка от 795 лв. и 50% от населението е със заплата по-висока от 795 лв. Средната месечна работна заплата в София-град е екстремална стойност (Boxplot) по отношение на средните работни заплати в другите области и не показва общото и типичното за съвкупността. Това е и причината, изчислената средна работна заплата, чрез средната аритметична да е значително по-висока по стойност в сравнение с медианата:



Средна месечна брутна работна заплата (лева)	
Mean	840.4047619
Standard Error	27.74880954
Median	794.5416667
Mode	#N/A
Standard Deviation	146.8328984
Sample Variance	21559.90006
Kurtosis	9.364579
Skewness	2.626338851
Range	747.1666667
Minimum	686.0833333
Maximum	1433.25
Sum	23531.33333
Count	28
Confidence Level(95.0%)	56.93585355





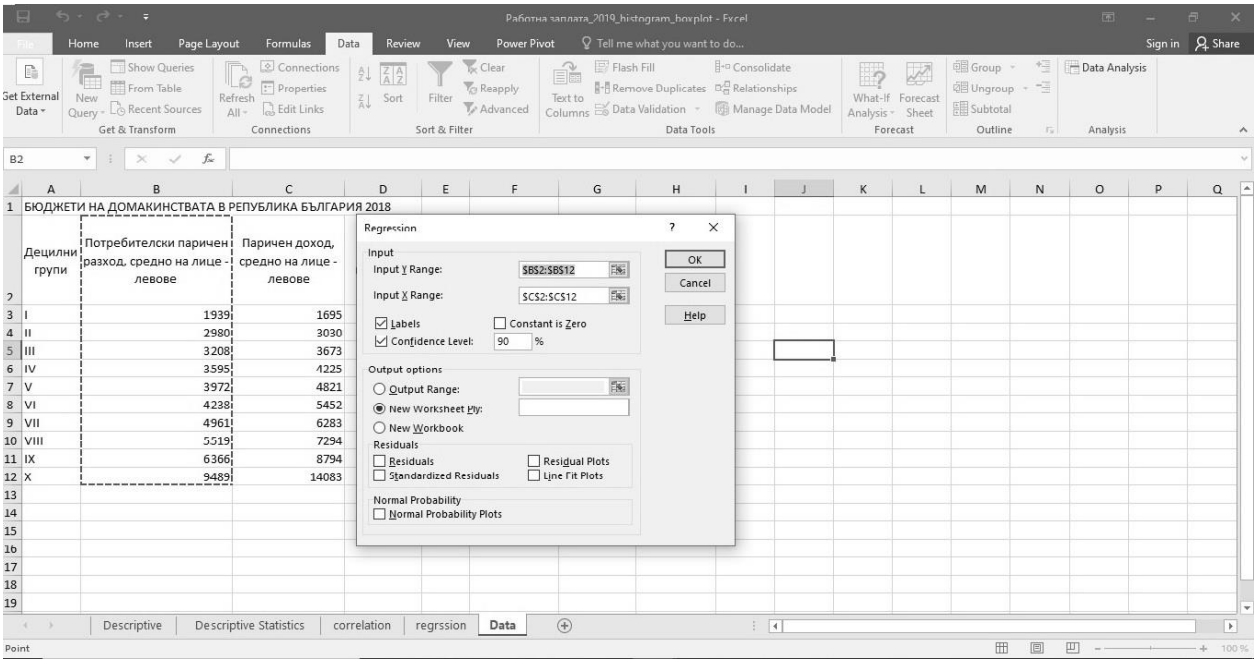
3. Корелационен и регресионен анализ в MS Excel

Изчисляването на корелационни коефициенти в Excel е бързо и удобно. Трябва само да изберете Correlation и да маркирате променливите, за които ще се изчислят корелационните коефициенти. Предимство на инструмента за корелационен анализ в Analysis ToolPak е, че вместо функцията, която се използва само за две променливи, тук може да маркарират повече променливи и като резултат се получава корелационна матрица с изчислените коефициенти за всички двойки променливи. Това дава обща картина между кои променливи съществува положителна и между кои отрицателна зависимост. Трябва да се отбележи, обаче, че Excel изчислява само корелационния коефициент на Браве-Пирсън, който се прилага при линейни зависимости. Този коефициент на корелация се променя в границите -1 и 1. Ако абсолютната му стойност е близка до 1, това означава, че между променливите съществува силна връзка, а ако е близък до 0, то връзката е слаба.

Описаният корелационен анализ се реализира чрез пример за паричния доход на лице от домакинство и потреблението на основни хранителни продукти на лице от домакинство в натурални измерители по децилни групи за 2018 г. Изчисления корелационен коефициент между паричния доход и хляб и тестени изделия е -0,88. Това показва силна и обратна зависимост между променливите. При увеличаване на дохода, потреблението на хляб и тестени изделия като малоценна стока намалява, защото хората започват да купуват по-скъпи и по-полезни хранителни продукти като риба, пресни плодове, зеленчуци и месо. Коефициентите на корелация между рибата, пресните плодове, зеленчуци и месото с паричния доход показват силна положителна зависимост, т.е. с увеличаване на дохода, тяхното потребление расте. За преориентацията в потреблението при увеличаване на дохода говорят и отрицателните корелационни коефициенти на хляба и тестените продукти с всички останали хранителни продукти. При намаляване на потреблението на хляб хората започват да потребяват повече от нормалните стоки:

	Паричен доход, средно на лице - левове	*Хляб и тестени изделия - кг.*	*Месо - кг*	*Риба и рибни продукти - кг*	*Плодове - пресни и замразени - кг*	*Зеленчуци - пресни и замразени - кг*
Паричен доход, средно на лице - левове	1					
Хляб и тестени изделия - кг.	-0.882373341	1				
Месо - кг	0.672479743	-0.61520942	1			
Риба и рибни продукти - кг	0.959197318	-0.83847332	0.80952	1		
Плодове - пресни и замразени - кг	0.876718988	-0.71727976	0.904456	0.924294	1	
Зеленчуци - пресни и замразени - кг	0.803511883	-0.65707538	0.948255	0.886495	0.98537474	1

Регресионният анализ е един от най-често използваните статистически анализи в бизнес изследванията, а изчисления при него са много трудоемки ако не се използва статистически софтуер. В MS Excel обаче, регресионният анализ се реализира бързо и лесно. Първо трябва да се избере Regression и след това внимателно да се попълнят полетата и да се изберат променливите. Необходимо е да се определи коя е зависимата и коя е независимата променлива или независимите променливи, а въвеждането им става чрез маркиране на данните. Резултатът от приложения регресионен анализ по подразбиране е на нова страница и включва таблица с обобщаващи числови характеристики, ANOVA за проверка на адекватността на целия регресионен модел и трета таблица с информация за отделните коефициенти на регресия. Една любопитна характеристика е, че автоматично се получават два доверителни интервала за параметрите на регресионния модел, независимо дали сте поставили отметка в полето за равнище на значимост. Оказва се, че първият доверителен интервал винаги е с 95% равнище на значимост, а вторият е с равнище на значимост по избор на потребителя, например 90%, но само ако поставите отметка в полето за равнище на значиммост:



БЮДЖЕТИ НА ДОМАКИНСТВАТА В РЕПУБЛИКА БЪЛГАРИЯ 2018

Децилни групи	Потребителски паричен разход, средно на лице - левове	Паричен доход, средно на лице - левове
I	1939	1695
II	2980	3030
III	3208	3673
IV	3595	4225
V	3972	4821
VI	4238	5452
VII	4961	6283
VIII	5519	7294
IX	6366	8794
X	9489	14083

Регресионният анализ се реализира, чрез модел на потребителска функция на основата на данни за потребителския паричен разход и паричния доход средно на лице от домакинство за 2018 г. Зависимата променлива е потреблението, а независимата дохода. Оценките на параметрите на модела са поместени в третата таблица. Те са статистически значими и показват, че доходът е статически значим фактор при определяне на потреблението. В същата насока е и коефициента на детерминация (99,8%), който показва, че 99,8% от вариацията в потреблението се определя от вариациата в дохода. Оценката на свободния член в регресионното уравнение (1038,70) се нарича още автономно потребление и е онази част от потреблението, която не зависи от разполагаемния доход и показва размера на потреблението дори при нулев доход. Тази стойност може да е отправна точка при определяне на минималната и средната работни заплати от правителството. Оценката на параметъра пред независимита променлива (0,60) е известна като пределната стойност на потребление и разкрива, че при всеки допълнителен лев доход за потребление се изразходват 60 ст.:

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.999218515
R Square	0.998437642
Adjusted R Square	0.998242347
Standard Error	89.66467459
Observations	10

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	41102950.1	41102950	5112.464	1.63024E-12
Residual	8	64318.031	8039.754		
Total	9	41167268.1			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 90.0%*	*Upper 90.0%*
Intercept	1038.707148	57.637444	18.0214	9.22E-08	905.7949637	1171.619332	931.5275521	1145.886743
Паричен доход, средно на лице	0.60454808	0.00845504	71.50149	1.63E-12	0.58505072	0.624045439	0.588825524	0.620270635

Заключение

Статията показва как с помощта на модула Data Analysis на MS Excel е възможно да се извърши един комплексен статистически анализ, демонстриран с примерите за доходите, разходите и потребителското търсене. Резултати от анализа съвпадат с изискванията на икономическата теория, а получените оценки са в границите на теоретичните очаквания.

Въпреки, че се налага да се пренареждат данни, да се разширяват изходните колони, да се преформатират графики и други, то анализите се реализират сравнително бързо и лесно и са достъпни за всички потребители на MS Excel. С всяка нова версия на MS Excel се добавят по-нови и по-добри инструменти, като нови функции и диаграми за анализ на данни, но подобрения в Analysis ToolPak няма. Загадка е защо Microsoft не изразходва минималното време и средства, необходимо за обновяване и доразвитие на Analysis ToolPak.

References

1. HADZHIEV, V., DIMITROVA, V., LUBENOV, L. (2009) *Statistical and Econometric Software*, Varna: Science and Economics [in Bulgarian].
2. SALOV, V. et.al. (2019) *Informatics*, Varna: Science and Economics [in Bulgarian].
3. SLAVEVA, K., PETKOV, P., IVANOV, L., VARBANOV, T., and GEORGIEVA, N. (2016) Improving Statistics Training Using the Modern Information and Communication Technologies. *Scientific Research Almanac, D. A. Tsenov Academy of Economics, Svishtov, Bulgaria,* issue 23, pages 498-528 [in Bulgarian].
4. WINSTON, L. W. (2016) *Microsoft Excel 2016. Data Analysis and Business Modeling.* Washington: Microsoft Press.