Мартин Пъшев Иванов

РЪКОВОДСТВО ЗА ИЗСЛЕДВАНЕ НА КОЛИЧЕСТВЕНИ ЗАВИСИМОСТИ

София
2008

Мартин Пъшев Иванов

РЪКОВОДСТВО ЗА ИЗСЛЕДВАНЕ НА КОЛИЧЕСТВЕНИ ЗАВИСИМОСТИ

София
2009

Настоящото ръководство е предназначено за научно-изследователска работа и обучение в сферата на приложение на количествени методи за обработка на мениджмънт статистическа информация. Използвано е за подготовка на магистри и докторанти на Лятна школа'2009 по Електронно управление, провеждана със съдействието на Фонд научни изследвания. Тестовият пример и данните са посготвени от доц. Ана Розева по идея на доц. Б. Колчагова. Авторът изказва искрена благодарност и за консултативната помощ на проф. Р.Цанкова.

Авторът е преподавател в департамент „Информатика" на Нов Български Университет.

Редактор проф.д.т.н.Р.Цанкова

Оценка на наличието на количествена връзка между независима и зависима величини чрез регресионен анализ

1. Същност на задачата.

Поставената задача изисква да се оцени количествената връзка между независима и зависима величини, като тази връзка се представи в математически вид на просто уравнение. За целите на обучението се разглежда пример, в който се приема, че независимата величина в уравнението е "Потребности общо" от специалисти с висше образование и, че изменението на нейните стойности влияе върху стойностите на величината "Брой завършващи висше образование". Данните за стойностите на двете величини са получени чрез наблюбение по статистически път (т.нар.статистическа извадка), а математическата форма на зависимостта между величините се установява чрез прилагане на статистическия метод на простата линейна регресия. Те са представени по специалности, местоположение времеви период.

Методът на едномерната (проста) регресия е предназначен да оцени наличието и вида на статистическата връзка между две количествени величини, зададени чрез извадка от техни наблюдавани стойности. Приема се, че поведението на едната величина (наречна „зависима") е статистически свързано с изменението на другата величина (наречна „независима"). Задачите на регресията са:

- да определи вида на статистическата връзка между независимата и зависима статистическа величина под формата на уравнение на изравняваща права във вида:

$$Y = B_0 + B_1 * X,$$

където Y представя изравнените стойности на зависимата величина, а X – стойностите на независимата променлива.

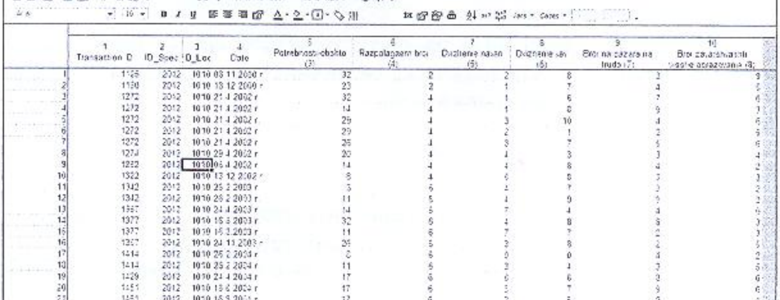
- да провери статистическата значимост на влиянието на независимата величина върху поведенивто на зависимата;
- да установи статистическата значимост на стойностите на параметрите B_0 и B_1 в регресионното уравнение.

В разглеждания пример се установява наличието и формата на регресионната връзка между величините „Брой завършващи висше образование" (зависима) и „Потребности общо" (независима), представени чрез наблюдаваните количествени стойности в извадката.

2.Подготовка на данните и съдържание на извадката.

Данните се подготвят във файл, имащ формат .xls, в съответствие с поставеното задание, след което се импортират в специализирания програмен продукт Statistica. След импортирането те се разполагат в работното пространство на продукта, което е организирано във вид на таблица (подобно на това в MS Excel). Колоните на работното пространство (таблицата) отговарят на статистическите променливи, включени в изследването, а редовете – на наблюдаваните стойности (случаи) в извадката. Таблицата в работното пространство на Statistica няма възможностите за извършване на електронни изчисления както в Excel, затова те трябва да бъдат извършени предварително в подготвителната фаза. Продуктът предлага обаче средства за форматиране на данните според техния тип, средства за задаване на имена и за спесифициране на статистическите променливи, за свързване на стойностите в колоните и други удобства.

Извадката от данните, използвани за решаването на поставената задача, се организира по специалности, местоположение и времеви период и е представена в продукта Statistica в следния вид:


STATISTICA - [Data: Copy of Supply_Demand_1_* (10v by 22c)]

	1 Transaction D	2 ID_Spec	3 ID_Loc	4 Date	5 Potrebnost-obshto (3)	6 Razpolagaem broi (4)	7 Dvizhenie navan (5)	8 Dvizhenie van (6)	9 Broi na pazara na trudo (7)	10 Broi zavarshvashti visshe obrazovanie (8)
1	1125	2012	1010	08.11.2000 г	32	2	2	8	3	9
2	1130	2012	1010	13.12.2000 г	23	2	1	7	4	5
3	1272	2012	1010	21.4.2002 г	32	4	1	6	7	6
4	1272	2012	1010	21.4.2002 г	14	4	1	8	9	7
5	1272	2012	1010	21.4.2002 г	29	4	3	10	4	6
6	1272	2012	1010	21.4.2002 г	29	4	2	1	2	5
7	1272	2012	1010	21.4.2002 г	25	4	3	7	5	6
8	1274	2012	1010	29.4.2002 г	20	4	4	3	3	4
9	1282	2012	1010	06.4.2002 г	14	4	4	8	4	2
10	1322	2012	1010	13.12.2002 г	8	4	6	8	3	3
11	1342	2012	1010	25.2.2003 г	5	6	2	7	3	2
12	1342	2012	1010	26.2.2003 г	11	5	4	9	5	2
13	1357	2012	1010	24.4.2003 г	14	5	7	4	4	5
14	1377	2012	1010	15.5.2003 г	32	6	4	8	6	3
15	1377	2012	1010	15.5.2003 г	11	6	7	7	2	3
16	1387	2012	1010	24.11.2003 г	25	5	3	8	2	5
17	1414	2012	1010	25.2.2004 г	8	6	0	0	4	2
18	1414	2012	1010	25.2.2004 г	11	5	3	4	3	5
19	1429	2012	1010	24.4.2004 г	17	6	6	6	2	6
20	1451	2012	1010	15.6.2004 г	17	6	5	7	9	6
21	1451	2012	1010	15.9.2004 г	17	6	2	5	2	4
22	1464	2012	1010	24.11.2004 г	26	5	3	4	7	9

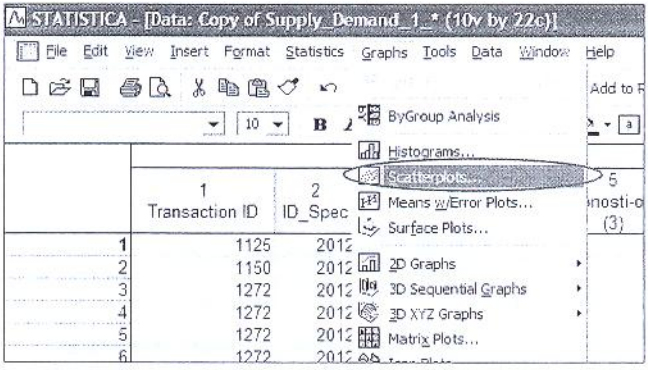
Пълният набор от данни, използвани в примера така, както са представени в продукта Statistica, е показан в таблица в края на този раздел.

3. Визуализиране на данните в двумерна диаграма.

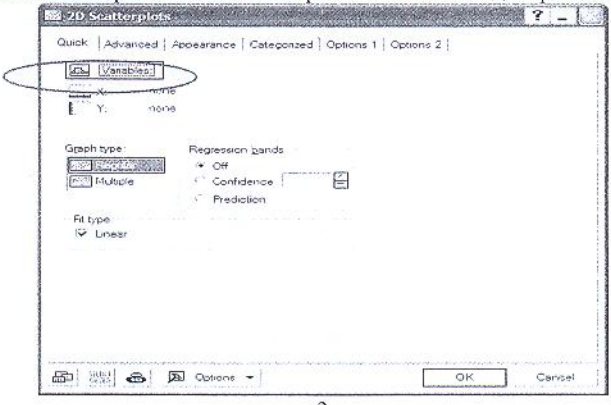
Данните за независимата и зависима променливи от извадката (за 22 случая) могат да бъдат изобразени в двумерна диаграма, наречена „диаграма на разсейването" (англ.scatterplot). Всяка двойка от стойности на независимата и зависима променлива се представя като точка в двумерна координатна система. Диаграмата дава нагледна представа за взаимното разположение на точките и евентуално за вида на връзката между тях.

Изобразяването на диаграмата на разсейването в продукта Statistica става в следната последователност:

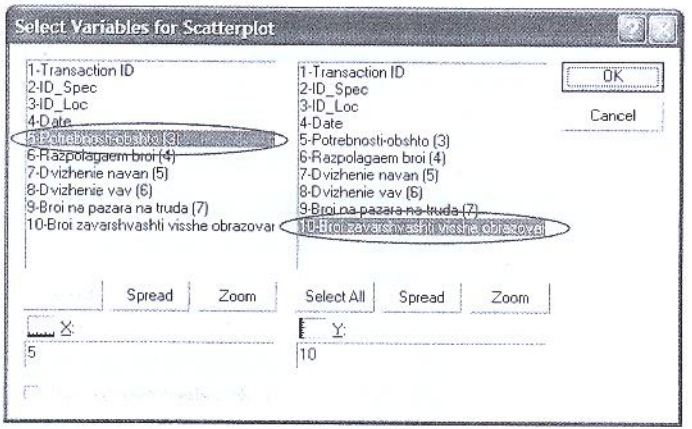
1) От главното меню се избира "Graphs"с подменю "Scatterplots":



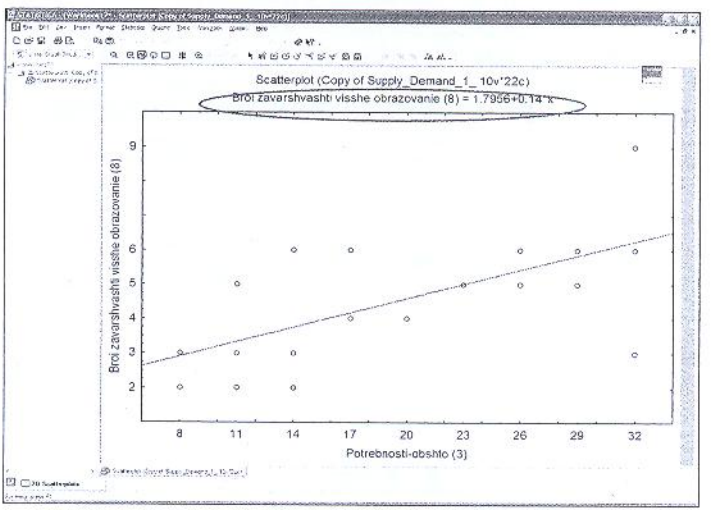
2) От отворения прозорец "2D Sctterplots" се избират променливите, изобразявани по хоризонталната и вертикалната ос на диаграмата.



Изборът на променливите по хоризонталната (X) и по вертикалната ос (Y) се прави след натискането на бутона "Variables" в отворения прозорец "Select Variables for Sctterplot":



3) След избора на променливите по двете оси се натиска бутон "OK", след което се изобразява прозорец, съдържащ диаграмата на разсейването.



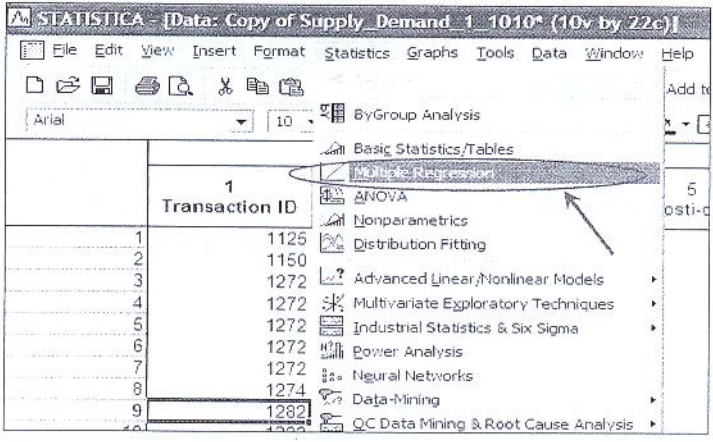
В заглавното поле на показаната диаграма се съдържа запис на регресионната зависимост между променливите. Това в общия случай е същата регресионна зависимост, която би се получила от предстоящия регресионен анализ, но без да съдържа съществена информация, отнасяща се до оценка на значимостта на изчислените стойности на параметрите в регресията.Тази информация е непълна и може да бъде използвана само като предварителна оценка на зависимостта между променливите.

Линията, изравняваща данните в диаграмата и отговаряща на посочената по-горе регресионна връзка е изобразена в червено.

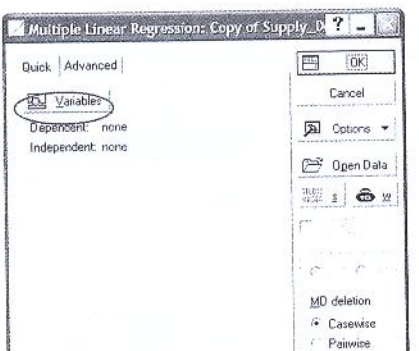
4. Изследване на връзката между величините „Брой завършващи висше образование" и „Потребности общо" с модул Multiple linear Regression на продукта Statistica.

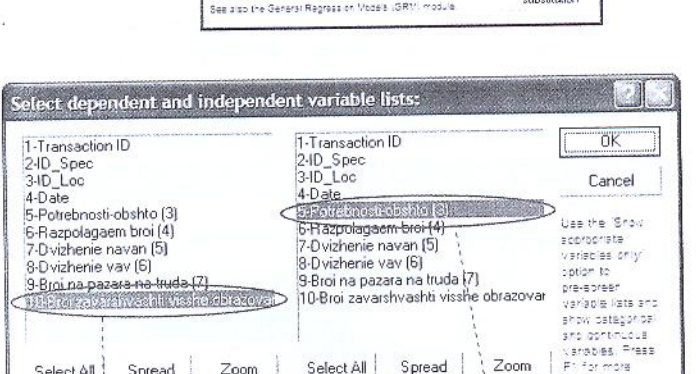
Установяването на наличие на регресионна връзка между величините „Брой завършващи висше образование" и „Потребности общо", стойностите на коефициентите в регресионното уравнение и тяхната значимост чрез модула Multiple Regression става в следната последователност:

1) От основното меню Statistics се избира позиция Multiple Regression

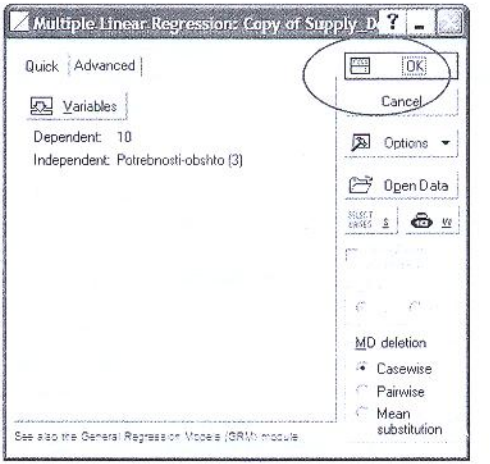


2) След избирането на модул Multiple Regression се посочват независимата (Independent) и зависимата (Dependent) променлива в регресионното уравнение. Това става след натискането на бутон "Variables" и отварянето на диалоговия прозорец "Select depend and independent variables":

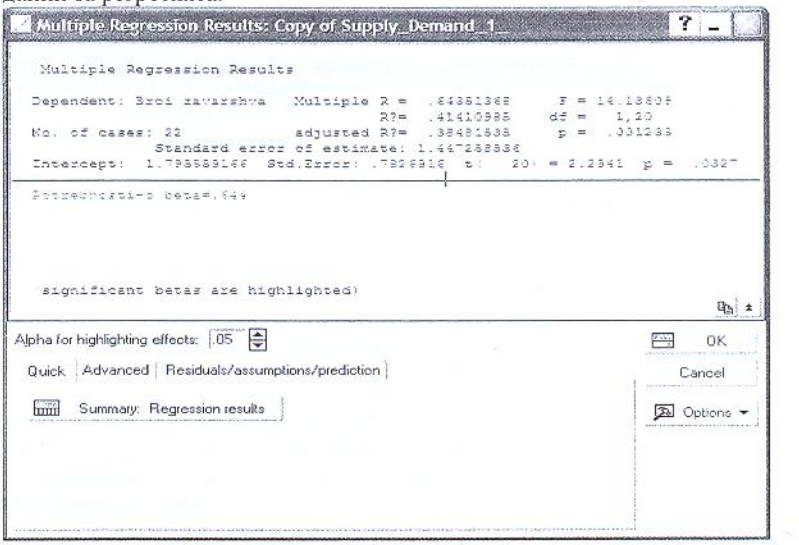




Зависимата променлива (“Broj zavarshvashvashti visshe obrazovanie”) се избира от левия списък, а независимата променлива (“Potrebnosti-obshto”) – от десния. В случая на множествена регресия се избират няколко независими променливи.



3) Обработка на данните става след натискане на бутон "OK": Обобщените резултати от анализа се изобразяват в прозореца "Multiple regression results:", който съдържа най-общи числови характеристики и данни за регресията.



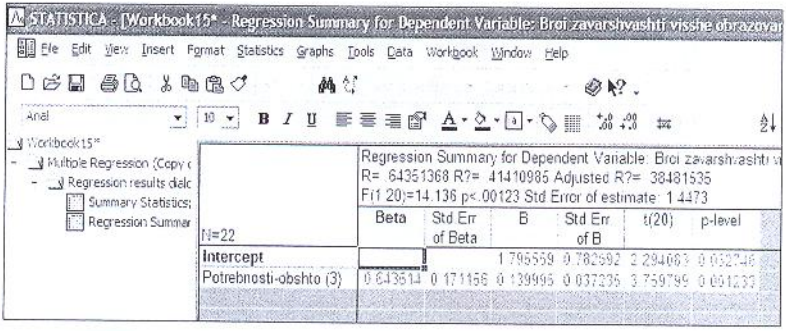
Основните параметри, показани в прозореца са:

Коефициент на множествена корелация (Multiple R): 0.643513368. Този коефициент оценява степента на корелация между участвщите в анализа променливи (зависима променлива и една или повече независими променливи). Стойностите на коефициента са между -1 и 1. Близките до единица (по абсолютна стойност) стойности на коефициента говорят за значителна корелация между променливите в регресионния модел.

Коефициент на детерминация (R^2): 0.41410985. Това е стойността на коефициента на множествена корелация, повдигната на втора степен. Коефициентът на детерминация оценява каква част от изменението на зависимата величина спрямо нейната средна стойност се обяснява с влиянието на участващите в регресионния модел независми променливи (в разглеждания случай такава е само една – „Potrebnosti - obshto").

Следващите параметри в прозореца (F, dF, p и др.) са свързани с оценяването на статистическата значимост на оценките на Multiple R и R^2.

4) След натискане на бутон "Summary: Regression results" в табличен вид се получават най-важните статистически оценки за регресионната връзка между независимата и зависимата статистически величини:



Резултатите са показани в следните редове и колони на таблицата:

- В колона „B" - стойностите на свободния член (B_0) и на коефициента (B_1) пред независимата променлива в регресионното уравнение (съответно в редове Intercept и "Broj zavarshvashti wisdhe obrazowanie"). Тези стойности са съответно 1.795559 и 0.139995.

- В колона "Std.Err. of B" - стандартна грешка на оценката (съответно за свободния член и за коефициента) ;
- В колона "t(20)" - статистики на Стюдент, необходими за изчисляване на значимостта за оценките на B_0 и B_1;
- В колона "p – level" – оценка на статистическата значимост на изчислените стойности на B_0 и B_1. Величините в тази колона оценяват т.нар. „вероятностт за грешка от първи род", т.е въз основа на получените резултати да бъде погрешно отхвърлено предположението (хипотезата) за липса на регресионна връзка между независимата и зависимата величина, при условие, че такава връзка наистина отсъства. Стойностите на тези оценки за свободния член B_0 и за коефициента B_1 съответно са: 0.032746 и 0,001233. Тези стойности се сравняват с априорно приета допустима вероятност за грешка, която за разглеждания пример е 0.05 (5%). И в двата случая стойностите на оценките е под праговата стойност, което показва, че получените оценки са статистически значими.

5. Заключение.

След завършване на анализа и изчисляване стойностите на параметрите B_0 и B_1 регресионното уравнение, свързващо независимата променлива „Потребности общо" със зависимата „Брой завършващи висше образование" получава вида:

$$Y = 1.795559 + 0.139995*X$$

където X е независимата, а Y – зависимата статистическа величина.

Въз основа на горното регресионно уравнение може да се пресметне очакваната средна стойност на величината „Брой завършващи висше образование" при условие, че е известна стойността на независимата променлива „Потребности общо". Важно е да се отбележи, че в зависимост от пълнотата и достоверността на началните данни и прецизността на изпълнение на анализа, крайната оценка на параметрите в регресионното уравнение може да бъде повече или по-малко неточна. За да се оценят възможните граници на отклонението на изчислените стойности, (т.нар.доверителни граници) следва да се направи допълнителен анализ, който тук няма да бъде показан поради своята специфика и теоретични особености.

Таблица с началните данни за извършване на анализа

	Transaction ID	ID_Spec	ID_Loc	Date	Potrebnosti-obshto (3)	Razpolagaem broi (4)	Dvizhenie navan (5)	Dvizhenie vav (6)	Broi na pazara na truda (7)	Broi zavarshvashti visshe obrazovanie (8)
1	1002	2011	1001	8/11/00	32	2	2	8	3	9
2	1045	2027	1001	12/18/00	23	2	1	7	4	5
3	1105	2008	1001	4/21/02	32	4	1	6	7	6
4	1359	2009	1001	4/21/02	14	4	1	8	9	3
5	1437	2026	1001	4/21/02	29	4	3	10	4	6
6	1448	2022	1001	4/21/02	29	4	2	1	2	5
7	1468	2009	1001	4/21/02	26	4	8	7	5	6
8	1006	2011	1002	4/29/02	20	4	4	3	3	4
9	1081	2008	1002	6/4/02	14	4	4	8	4	2
10	1091	2027	1002	12/13/02	8	4	8	8	3	3
11	1287	2011	1002	2/26/03	8	6	4	7	3	2
12	1305	2011	1002	2/26/03	11	6	4	9	9	2
13	1365	2026	1002	4/24/03	14	5	7	4	4	6
14	1052	2011	1003	8/16/03	32	6	4	8	8	3
15	1112	2027	1003	8/16/03	11	6	7	7	2	3
16	1303	2022	1003	11/24/03	26	5	3	8	2	5
17	1303	2022	1003	2/26/04	8	6	0	0	4	2
18	1395	2026	1003	2/26/04	11	6	3	4	3	5
19	1092	2008	1004	4/24/04	17	6	8	6	8	6
20	1258	2011	1004	8/16/04	17	6	5	7	9	6
21	1300	2011	1004	8/16/04	17	6	2	5	2	4
22	1374	2022	1004	11/24/04	26	5	9	4	7	6

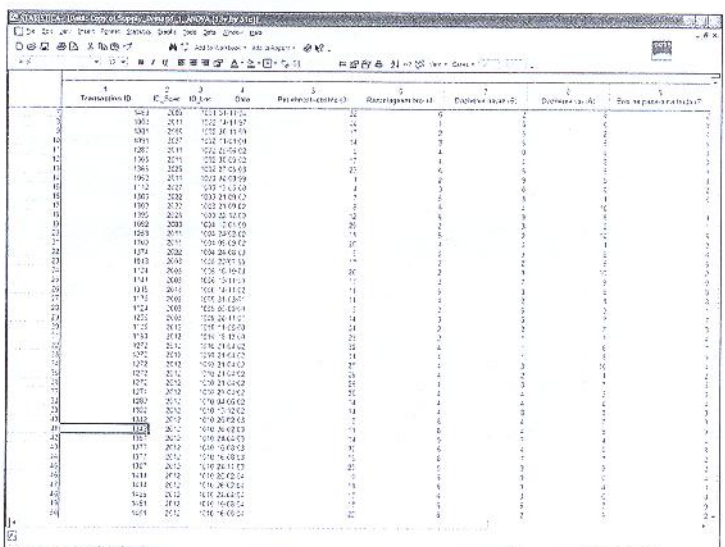
Оценка на наличието на връзка между качествен фактор и количествена статистическа величина със средствата на дисперсионния анализ (ANOVA)

1. Същност на задачата.

Статистическата оценка за наличие на връзка между две величини, едната от които се изразява в качествени категории (т.нар.нива), а другата е количествена се извършва със средствата на диперсионния анализ. Такъв е случаят с величините „Местоположение (регион)" и „Потребности общо", първата от които има смисъл на фактор (изразен качествено), влияещ върху стойностите на втората (приета за зависима от фактора). Крайната цел на анализа е да се състави статистически обоснован извод за наличието или за отсъствието на такава връзка между двете величини.

2. Съдържание на данните.

Данните са представени в следната таблица, импортирана в работното пространство на продукта Statistica.

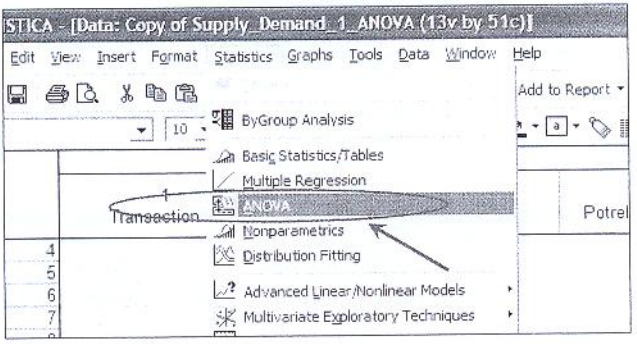


Пълният набор от данни, използвани в примера така, както са представени в продукта Statistica, е показан в таблица в края на този раздел.

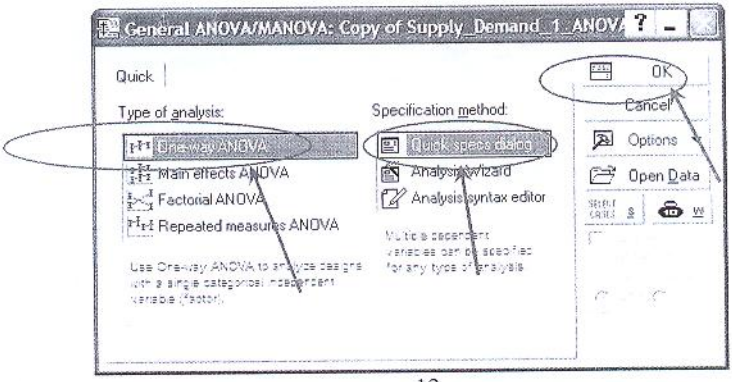
3. Изследване на значимостта на влиянието на качествения фактора „Местоположение (регион)" върху количествения показател „Потребности общо"

Изследването на връзката между качествения фактор „Регион (местоположение)" и количествената зависима величина „Потребности общо" става със средствата на едномерния дисперсионнен анализ (ANOVA Analysis Of Variance). Техниката на това изследване с продукта се състои от няколко последователни стъпки.

1) От основното системно меню на продукта се избира позиция Statistics, след което от падащото меню се избира модулът ANOVA:

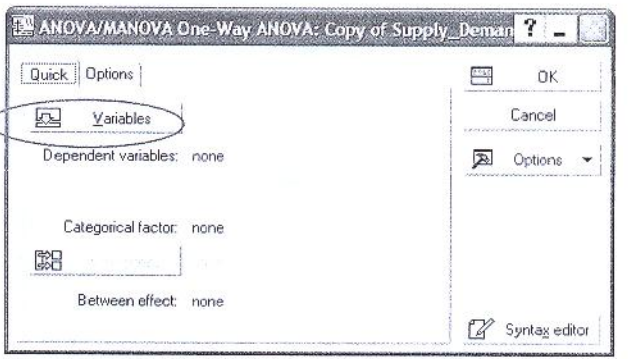


2) След избирането на модул ANOVA се посочва типът на анализа.

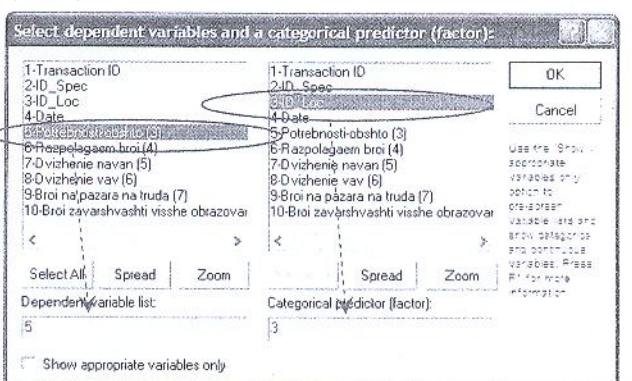


В случая се извършва опростен едномерен анализ на връзката между фактора на въздействието и зависимата величина, поради което като тип на обработката се посочва "One-way ANOVA". След този избор в прозореца "Specification method" се посочва "Quick specs dialog" и се натиска бутонът "OK".

3) В следващия прозорец трябва да се изберат променливите (изследван фактор и зависима променлива), които участват в анализа.

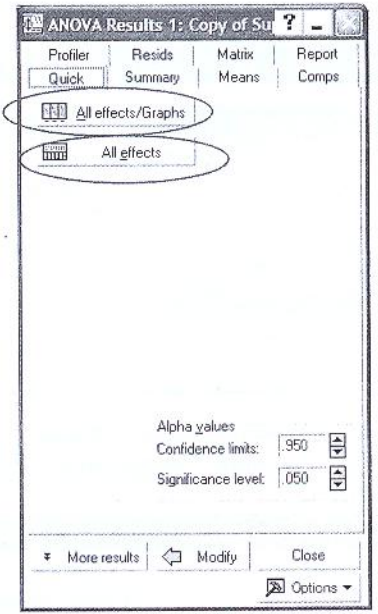


Това става след натискането на бутон "Variables", при което се отваря нов диалогов прозорец "Select dependent variables and a categorical predictor (factor):"



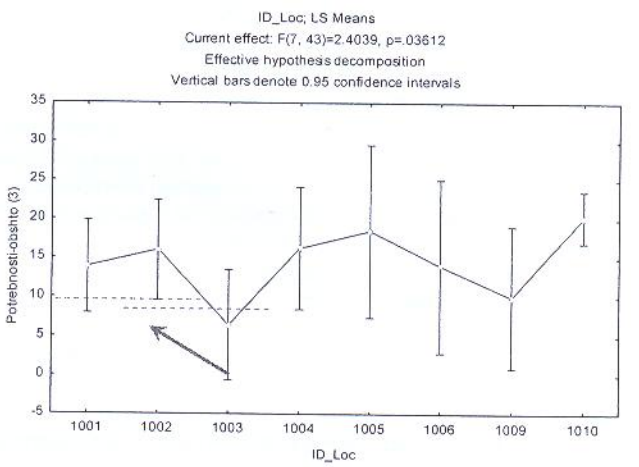
Зависимата променлива се избира от левия списък (Dependent variable list), а независимата (факторът) – от десния (Categorical predictor - factor), след което се натиска бутонът "OK" и се активира същинската обработка.

4) Възможностите за получаване на разнообразни резултати от анализа са представени в прозореца "ANOVA Results":



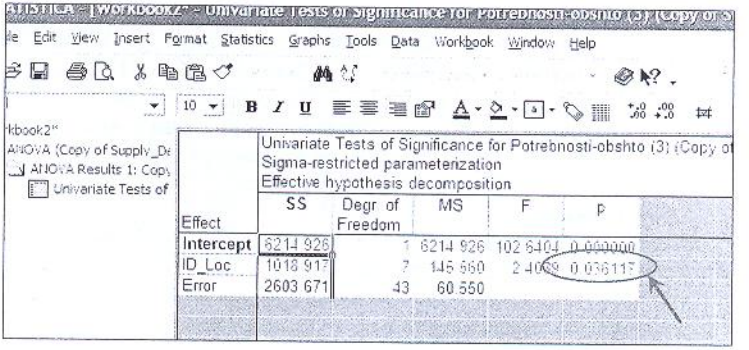
За решаването на конкретната задача е достатъчно ползването на страницата „Quick" от която най-напред с избира бутон "All effects/Graphs", а след това "All effects".

Задействането на бутона "All effects/Graphs" представя резултатите от анализа в графичен вид.



Графиката представя по вертикалната си ос средните стойности и 95% -те доверителни интервали на зависимата количествена величина по групи (нива на фактора). От графиката се вижда, че средните стойности за някои от групите попадат извън 95%-те доверителни граници за други групи, което е предварителен признак за значимо влияние на фактора върху зависимата величина.

Чрез бутона "All effects" резултатите се представят в табличен вид.



Таблицата показва обобщени крайни резултати от изпълнението на дисперсионния анализ. Тя позволява да се прецени каква част от измененията на зависимата количествена величина се дължат на влиянието на оценявания фактор и каква – на чисто случайни и външни за анализа причини. По този начин може да се провери хипотезата за наличието на статистически значима връзка между фактора " н Местоположение (регион)" и величината „Потребности общо".

Характеристиките, представени в таблицата са организирани по следния начин:

- редовете на таблицата съдържат:
  - **Intercept** – този ред няма отношение към решаването на конкретната задача в опростения и вид;
  - **ID_Loc** – името на променливата, съдържаща нивата на изследвания независим фактор „Местоположение (регион)" и данни за свързания с нея статистически анализ;
  - **Error** - ред, съдържащ статистическите оценки, отнасящи се до случайна грешка и участието на външни за анализа фактори.
- колоните съдържат:
  - **SS** (Sum of Squares) – сума от квадратите на отклоненията на наблюдаваните стойности на зависимата статистическа величина спрямо средната и стойност, причинено от влиянието на ... , на изследвания фактор или на външни случайни причини.
  - **Degr.of Freedom**- степени на свобода – статистическа характеристика, свързана с обема на извадката и броя на наблюдаваните нива на фактора "".
  - **MS** – колона SS, разделена на Degr.of Freedom.
  - **F** – статистика на Фишер – статистическа оценка, необходима за проверка на хипотезата за отсъствие (респ. за наличие) на влияние на фактора "Местоположение (регион)" върху наблюдаваната величина "Потребности общо".
  - **p-level** – оценка на вероятността за грешка, ако бъде отхвърлена хипотезата за отсъствие на влияние на фактора върху зависимата величина (т.нар нулева хипотеза).

4. Заключение.

Съществен за съставянето на статистически извод относно наличието на връзката между фактора и зависимата статистическа велияина е параметърът **p-level**. По същество той изразява вероятността да бъде направена т.нар.грешка от първи род (т.е.да бъде отхвърлена нулевата хипотеза при условие, че е вярна). Високата стойност на показателя се свързва с висока вероятност неправилно да бъде отхвърлено предположението за отсъствие

на статистическа връзка между двете величини. Ниската стойност на p-level – обратно – говори за по-малка вероятност да бъде съставен грешен статистически извод. Обикновено за гранична стойност за такава грешка се приема 0.05 (5%) и ако стойността на p-level е под тази граница, се прави статистически обоснован извод за наличие на значимо влияние на фактора „Местоположение (регион)" върху изменението на стойностите на величината „Потребности общо" (т.е.за наличие на статистическа връзка между качествения фактор и зависимата количествена величина).

В нашия пример стойността на показателя p-level е 0.036117, по-малка от граничната стойност 0.05, което дава основание да се отхвърли хипотезата за отсъствие на значима връзка между фактора на въздействието и зависимата променлива, т.е. следва да се приеме алтернативната хипотеза за наличието на такава връзка между двете величини. По отношение на силата на тази връзка дисперсионният анализ не дава конкретен отговор.

Таблица с изходните данни на примера

	Transaction ID	ID_Spec	ID_Loc	Date	Potrebnosti-obshto (3)	Razpolagaem broi (4)	Dvizhenie navan (5)	Dvizhenie vav (6)	Broi na pazara na truda (7)	Broi zavarshvashti visshe obrazovanie (8)
1	1002	2011	1001	26/05/97	17	1	5	7	8	5
2	1045	2027	1001	12/07/99	8	1	6	9	6	9
3	1105	2008	1001	30/03/00	8	2	4	9	6	2
4	1359	2009	1001	02/05/03	8	6	6	2	5	6
5	1437	2026	1001	27/05/04	29	6	9	9	1	7
6	1448	2022	1001	24/08/04	5	5	1	5	2	8
7	1468	2009	1001	01/11/04	22	6	2	8	2	8
8	1006	2011	1002	14/11/97	20	1	6	6	9	7
9	1081	2008	1002	30/11/99	17	2	5	2	4	2
10	1091	2027	1002	11/01/00	14	3	5	5	5	4
11	1287	2011	1002	25/06/02	5	4	8	5	8	1
12	1305	2011	1002	30/09/02	17	4	4	8	3	8
13	1365	2026	1002	27/05/03	23	6	5	5	9	5
14	1052	2011	1003	08/08/99	1	2	9	9	8	1
15	1112	2027	1003	15/05/00	4	3	6	8	2	7
16	1303	2022	1003	21/09/02	7	5	3	1	6	4
17	1303	2022	1003	21/09/02	8	5	4	10	1	7
18	1395	2026	1003	22/12/03	12	6	9	6	9	4
19	1092	2008	1004	15/01/00	29	2	3	2	1	7
20	1258	2011	1004	24/02/02	11	5	2	10	5	8

21	1300	2011	1004	09/09/02	20	4	4	4	2	1
22	1374	2022	1004	24/08/03	5	5	9	8	8	8
23	1048	2008	1005	22/07/99	17	2	2	5	5	7
24	1134	2003	1005	18/10/00	20	2	9	10	2	9
25	1141	2003	1006	13/11/00	17	2	7	9	9	7
26	1315	2018	1006	14/11/02	11	5	8	8	8	6
27	1175	2003	1009	31/03/01	11	4	2	1	8	7
28	1124	2008	1009	05/08/00	5	2	5	3	1	8
29	1235	2003	1009	28/11/01	14	3	6	7	7	7
30	1125	2012	1010	11/08/00	34	2	2	8	3	3
31	1150	2012	1010	18/12/00	23	2	1	7	4	3
32	1272	2012	1010	21/04/02	32	4	1	6	7	4
33	1272	2012	1010	21/04/02	14	4	1	8	9	3
34	1272	2012	1010	21/04/02	29	4	3	10	4	1
35	1272	2012	1010	21/04/02	29	4	2	1	2	4
36	1272	2012	1010	21/04/02	26	4	8	7	5	6
37	1274	2012	1010	29/04/02	20	4	4	3	3	9
38	1282	2012	1010	04/06/02	14	4	4	8	4	2
39	1322	2012	1010	13/12/02	14	4	8	8	3	4
40	1342	2012	1010	26/02/03	8	6	4	7	3	3
41	1342	2012	1010	26/02/03	11	6	4	9	9	1
42	1357	2012	1010	24/04/03	14	5	7	4	4	7
43	1377	2012	1010	16/08/03	32	6	4	8	8	3
44	1377	2012	1010	16/08/03	15	6	7	7	2	1
45	1387	2012	1010	24/11/03	29	5	3	8	2	1
46	1414	2012	1010	26/02/04	8	6	0	0	4	6
47	1414	2012	1010	26/02/04	11	6	3	4	3	8
48	1429	2012	1010	24/04/04	17	6	8	6	8	8
49	1451	2012	1010	16/08/04	17	6	5	7	9	7
50	1451	2012	1010	16/08/04	22	6	2	5	2	3
51	1464	2012	1010	24/11/04	27	5	9	4	7	6

Оценка на характеристиките на динамичен ред от наблюдения

1. Същност на задачата.

При анализа на данните за количествения показател „Потребности общо" е съществен въпросът за изучаване на изменението (динамиката) на неговите стойности във времето. За целта се обработва съвкупност от стойности на този показател за изминал период от време (няколко години), отстоящи на равни (или поне на приблизително равни) времеви интервали (например месец). Тези дани образуват т.нар."динамичен ред" и могат да бъдат обработени със специфични статистически методи – методи за изравняване на динамични редове (наричаи понякога и времеви редове). Резултатът от такава статистическа обработка може да бъде използван за изясняване на някои закономерности в изменението на стойностите на изследвания показател „Потребности общо" във времето, а също така и за прогнозиране на очакваните негови стойности за предстоящ период (или за няколко периода). Широко разпространен метод за изравняване на данни от динамични редове е така нареченото „експоненциално изравняване" или „експоненциално изглаждане", който се основава на следната връзка между стойностите на динамичния ред и прогнозните стойности за неговото очаквано поведение:

$$S_t = \alpha * X_t + (1-\alpha) * S_{t-1} ,$$

където:

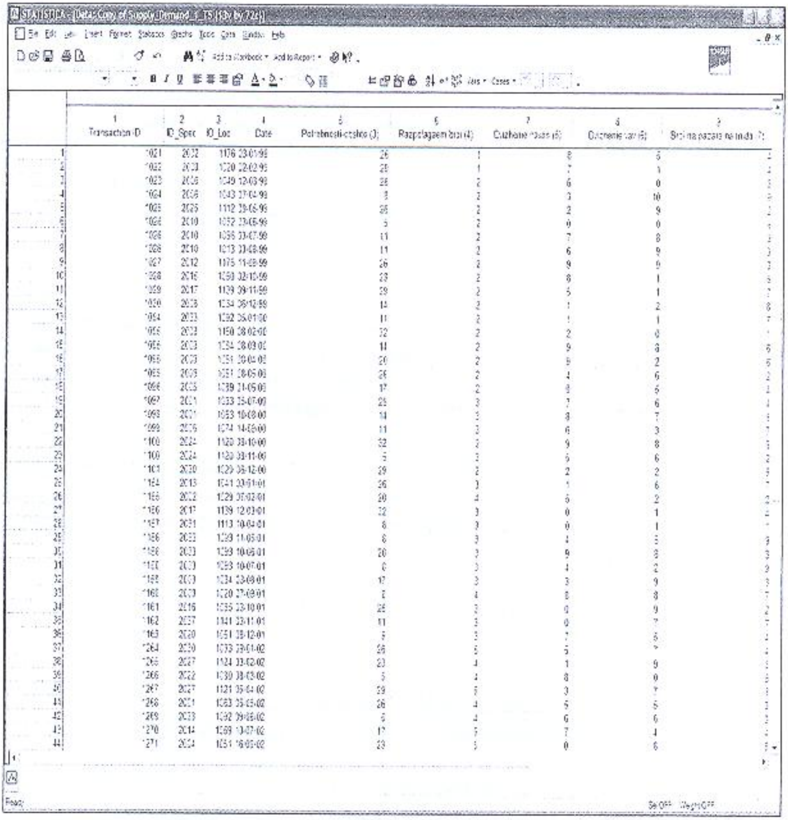
S_t – е прогнозната стойност за момент t, t = 1, 2, 3, ...;

X_t – е наблюдаваната стойност на динамичния ред в момент t, t = 1, 2, 3, …;

α - е параметър на изравняването – съпоставя наблюдаваните и прогнозните (изравнените) стойности в реда – централен параметър за прилагане на метода. Параметърът има стойности в интервала [0,1].

2. Съдържание на данните.

Данните се подготвят в табличен вид във файлов формат .xls, след което се импортират в продукта Statistica. Импортираната в работното пространство на продукта таблица има следния вид:



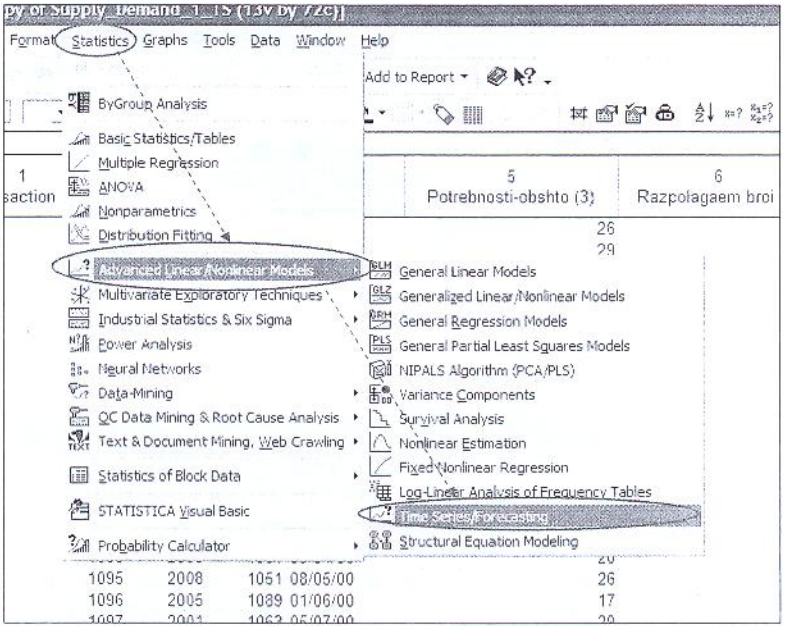
	1 Transaction ID	2 ID_Spec	3 ID_Loc	4 Date	5 Potrebnosti-obshto (3)	6 Razpolagaem broi (4)	7 Cuzhane [illegible] (5)	8 [illegible] (6)	9 Broi na [illegible] (7)
1	1021	2032	1176	23-01-99	26	1	8	5	2
2	1022	2033	1020	22-02-99	29	1	7	1	2
3	1023	2036	1049	12-03-99	28	2	6	0	[illegible]
4	1024	2036	1043	27-04-99	8	2	3	10	[illegible]
5	1025	2025	1112	29-05-99	35	2	2	9	2
6	1026	2010	1052	23-06-99	5	2	9	0	[illegible]
7	1026	2010	1056	30-07-99	11	2	7	8	[illegible]
8	1026	2010	1013	30-08-99	11	2	6	9	3
9	1027	2012	1175	11-09-99	26	2	9	9	3
10	1028	2016	1050	02-10-99	23	2	8	1	5
11	1029	2017	1139	09-11-99	39	2	5	1	7
12	1030	2028	1034	09-12-99	14	2	1	2	8
13	1054	2033	1092	25-01-00	11	2	1	1	7
14	1055	2033	1150	28-02-00	32	2	2	0	[illegible]
15	1056	2033	1034	28-03-00	14	2	9	8	6
16	1056	2033	1051	30-04-00	20	2	9	2	5
17	1055	2009	1051	28-05-00	26	2	4	6	2
18	1096	2035	1039	21-06-00	17	2	8	5	1
19	1097	2031	1033	25-07-00	25	3	7	6	4
20	1098	2031	1053	10-08-00	14	3	8	7	[illegible]
21	1099	2035	1074	14-09-00	11	3	6	3	7
22	1100	2024	1120	30-10-00	32	2	9	8	3
23	1100	2024	1120	30-11-00	5	3	5	6	2
24	1101	2030	1029	06-12-00	29	2	2	2	3
25	1164	2013	1041	20-01-01	26	3	1	6	[illegible]
26	1165	2022	1029	07-02-01	20	4	5	2	2
27	1156	2017	1139	12-03-01	22	3	0	1	2
28	1157	2031	1113	10-04-01	8	3	0	1	[illegible]
29	1156	2033	1139	11-05-01	8	3	4	5	9
30	1158	2033	1139	10-06-01	20	3	9	8	3
31	1150	2033	1093	10-07-01	0	3	4	2	9
32	1159	2033	1034	23-08-01	17	3	3	9	3
33	1160	2033	1020	27-09-01	2	4	8	8	7
34	1161	2016	1035	23-10-01	25	3	0	9	2
35	1162	2037	1141	23-11-01	11	3	0	7	7
36	1163	2020	1051	28-12-01	5	3	7	5	[illegible]
37	1264	2030	1073	29-01-02	26	5	5	[illegible]	[illegible]
38	1265	2027	1124	23-02-02	23	4	1	9	[illegible]
39	1266	2022	1030	28-03-02	5	4	8	0	[illegible]
40	1267	2027	1121	25-04-02	29	5	3	7	[illegible]
41	1268	2031	1083	25-05-02	26	4	5	5	[illegible]
42	1269	2033	1092	29-06-02	5	4	6	6	[illegible]
43	1270	2014	1069	10-07-02	17	5	7	4	[illegible]
44	1271	2034	1051	16-08-02	23	5	0	6	[illegible]

Данните са ежемесечни и обхващат период от пет последователни години. Таблица с даните, използвани в разглеждания пример, е представена на края на раздела.

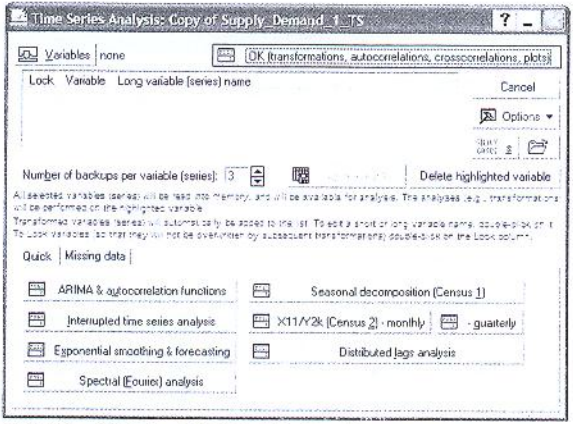
3. Обработка на данните от динамичния ред.

Обработката на данните от динамичния ред става с модула "Time series/Forecasting " на продукта Statistica и протича в следния ред:

1) От основното меню на продукта се избира Statistics, от падащото меню се избира позиция Advanced Linear/Nonlinear Models, а от следващото отварящо се падащо меня се избира модулът Time Series/Forecasting.

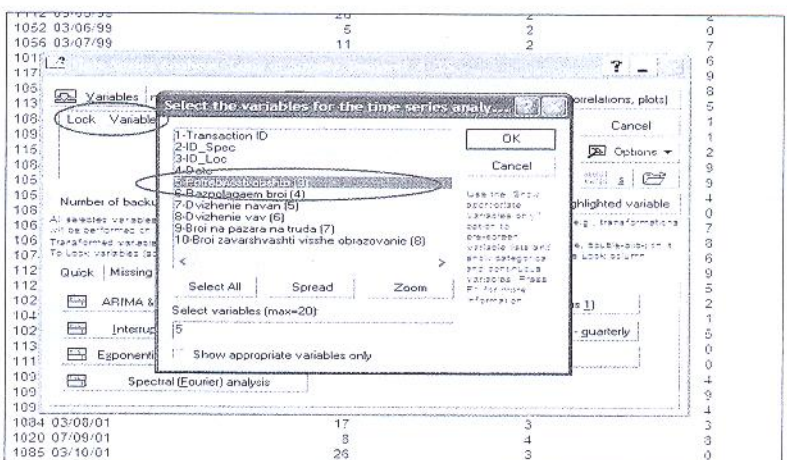


2) След избора на модула „Time Series/Forecasting" се отваря диалогов прозорец, чрез който се указват основните параметри на обработката.



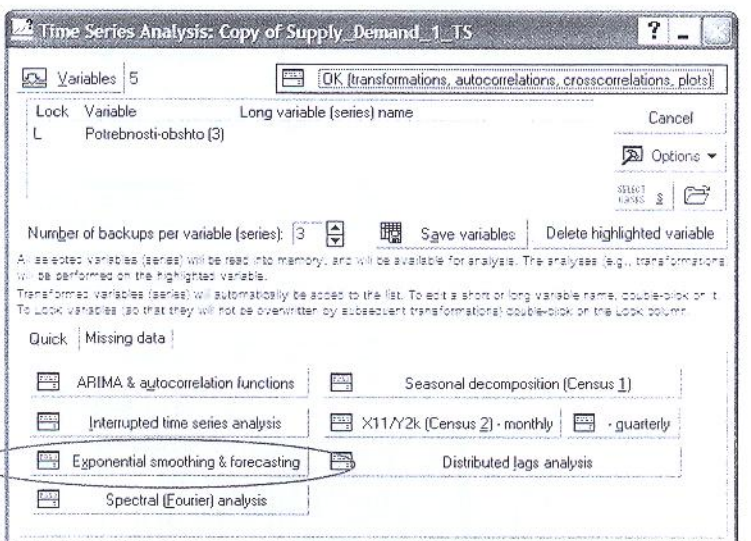
Първата стъпка, която следва да се изпълни след отварянето на прозореца е да се посочи името на променливата (колоната), за която се извършва анализа. В нашия случай това е променливата от колона 5 "Potrebnosti – obshto". По принцип продуктът позволява да се изберат няколко променливи едновременно, които съвместно да се обработват.

Изборът на променливата, чийто стойности ще се обработват със средствата на модул Time Series/Forecasting става след натискането на бутон Variables, след което се отваря следния диалогов прозорец.



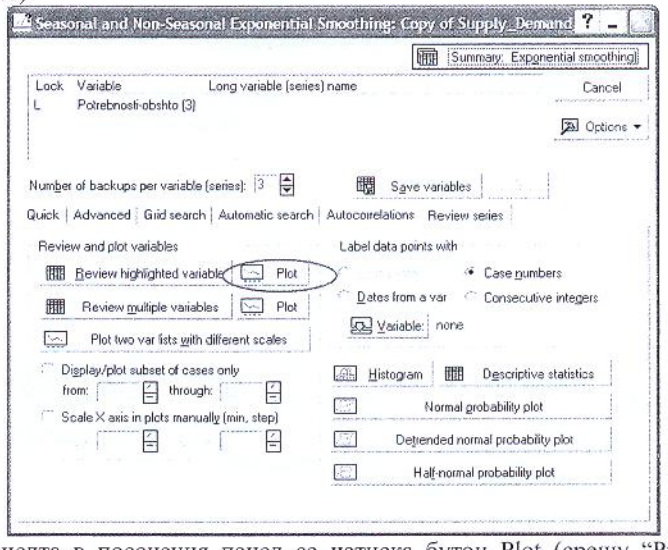
След посочване на името на променливата (в нашия случай "Potrebnosti – obshto") се натиска бутон "OK" и работата продължава в основния прозорец "Time Series Analysis:..".

3) Изборът на вида на обработката за посочената променлива се прави в основния диалогов прозорец "Time Series Analysis:..". Като цяло модулът "Time Series/Forecasting" предлага богати и разнообрани възможности за обработка в съответствие с теоретичните постижения в областта на обработката на динамични редове и статистическото прогнозиране. В нашия случай задачата е сравнително проста и за изпълнение на обработката съгласно посочения по-горе метод се избира „Exponential smoothing & forecasting" чрез натискане на съответния бутон.

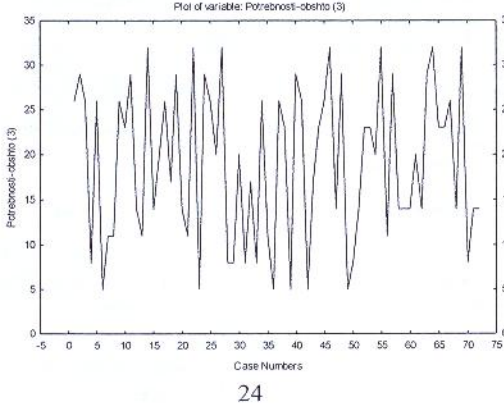


4) След натискането на бутона „Exponential smoothing & forecasting" се появява показания по-долу диалогов прозорец, съдържащ няколко панела. Избира се панел "Preview series", в който се съдържат средства за графично визуализиране на динамичния ред. Визуалната преценка се отнася до най-обща предварителна проверка на свойствата на

динамичния ред и преди всичко на неговата стационарност (способността на реда да запазва основните характеристики на поведението си във времето).

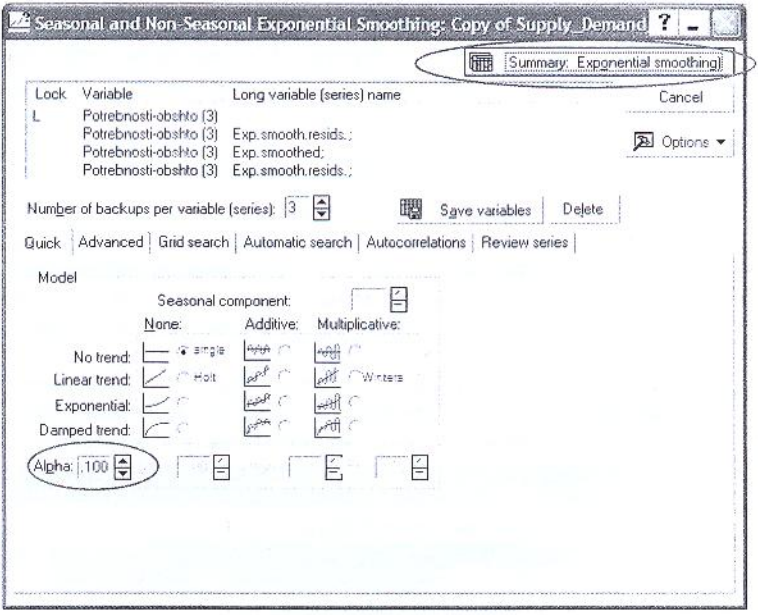


За целта в посочения панел се натиска бутон Plot (срещу "Preview Highlighted variables"). След натискането на бутона в графичен прозорец се получава графика за поведението на динамичния ред във времето. Тази графика е преди всичко илюстративна и е предпоставка за най-общ и начален анализ на данните.

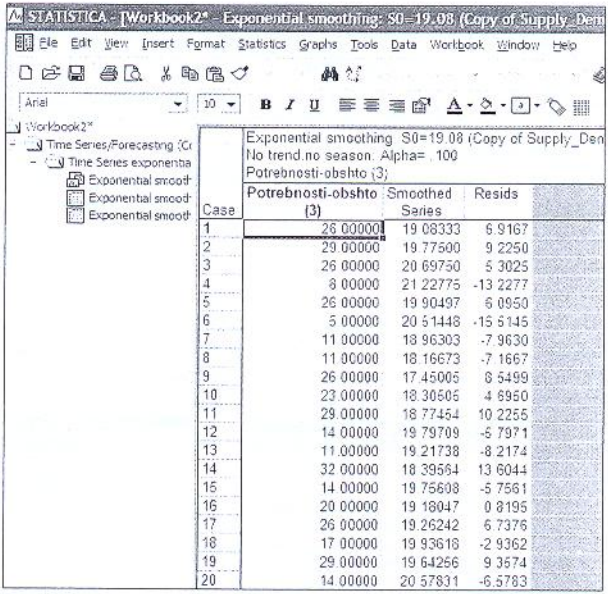


Графиката показва стационарен динамичен ред за изследваната променлива с ограничени колебания в амплитудата по нейните стойности.

5) Натискането на бутон "Summary: Exponential smoothing" дава числените данни от изравняването: стойности на изходния ред (Potrebnosti - obschto), изравнените стойности (Smoothed series) и остатъците (Resids) – т.е. разликите между наблюдаванатите и изравнените стойности. Желателно е изравняването да стане при различни стойности на изглаждащия параметър alpha (специфичен за метод Exponential smoothing). Стойността на този параметър (между 0.0 и 1.0) се избира в панел "Quick" (това е полето Alpha с начална стойност 0.100, близо до долния ляв ъгъл на прозореца):

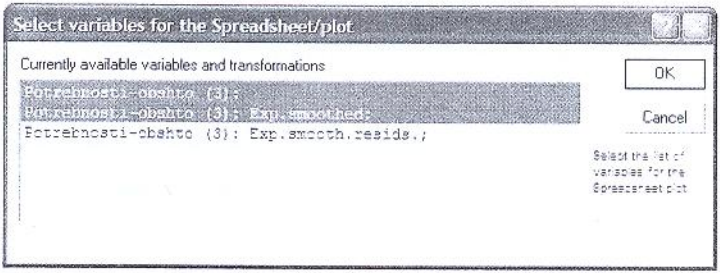


Резултатите от изравняването на динамичния ред за променливата "Potrebnosti - obshto" се представят в следната таблица:

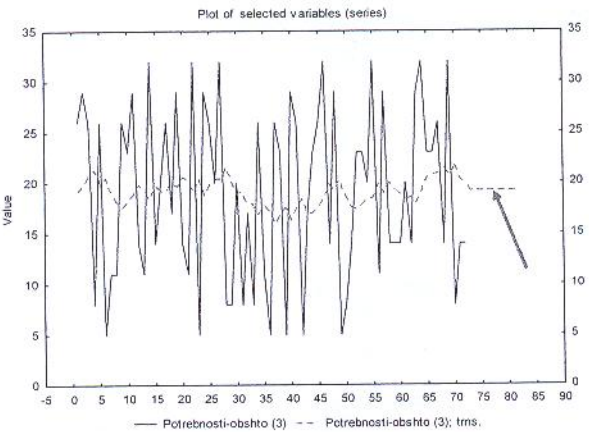


В таблицата колоната "Potrebnosti-obshto" съдържа данните на началния динамичен ред за изследваната променлива, колоната "Smoothed Series" съдържа изравнените данни, а колоната "Resids" – остатъците (отклоненията на изравнените стойности от началните).

6) Данните от таблицата могат да бъдат изобразени в съвместна графика чрез натискането на бутон "Preview multiple variables" в панела "Preview series", при което се отваря следния диалогов прозорец.



В него се избират за съвместно изобразяване двата реда – изходният и изравненият (чрез задържане на клавиш Control при посочването им), а именно: "Potrebnosti-obshto (3)" и "Potrebnosti-obshto (3): Exp.smoothed". След натискане на бутон Plot се изобразява следният графичен резултат:



В графиката основният динамичен ред е показан в синьо, а данните от изравняването – в червено. Забелязва се как данните от изравняването следват основния динамичен ред. В най-дясната част на графиката стойностите от изравняването екстраполират данните от наблюденията (червената хоризонтална линия след стойност 70 за оста x). Екстраполираните изравняващи стойности имат смисъл на прогнозна оценка за няколко периода (в случая - 10) напред.

Екстраполираните стойности могат да се видят и в таблицата с основните резултати от изравняването. Те са посочени в редове от 73 до 82 – всичко екстраполация (прогноза) с десет стъпки напред. В случая екстраполираната стойност е 19.21973.

4. Заключение.

Коректното интерпретиране на получените прогнозни резултати изисква добро познаване на метода „експоненциално изравняване", както и на особеностите на изследвания динамичен ред (неговата физическа или икономическа същност). Механичното прилагане на прогнозната оценка от екстраполацията може да доведе до сериозни недоразумения и грешки и да компрометира извършения анализ.

	Transaction ID	ID_Spec	ID_Loc	Date	Potrebnosti-obshto (3)
1	1021	2002	1176	03/01/99	26
2	1022	2008	1020	02/02/99	29
3	1023	2006	1049	12/03/99	26
4	1024	2006	1043	07/04/99	8
5	1025	2025	1112	09/05/99	26
6	1026	2010	1052	03/06/99	5
7	1026	2010	1056	03/07/99	11
8	1026	2010	1013	03/08/99	11
9	1027	2012	1175	11/09/99	26
10	1028	2015	1050	02/10/99	23
11	1029	2017	1139	09/11/99	29
12	1030	2008	1084	06/12/99	14
13	1094	2033	1092	05/01/00	11
14	1095	2008	1150	08/02/00	32
15	1095	2008	1084	08/03/00	14
16	1095	2008	1051	08/04/00	20
17	1095	2008	1051	08/05/00	26
18	1096	2005	1089	01/06/00	17
19	1097	2001	1063	05/07/00	29
20	1098	2001	1063	10/08/00	14
21	1099	2006	1074	14/09/00	11
22	1100	2024	1120	08/10/00	32
23	1100	2024	1120	08/11/00	5
24	1101	2030	1029	06/12/00	29
25	1154	2013	1041	03/01/01	26
26	1155	2002	1029	07/02/01	20
27	1156	2017	1139	12/03/01	32
28	1157	2031	1113	10/04/01	8
29	1158	2033	1093	11/05/01	8
30	1158	2033	1093	10/06/01	20
31	1158	2033	1093	10/07/01	8
32	1159	2003	1084	03/08/01	17
33	1160	2003	1020	07/09/01	8
34	1161	2016	1085	03/10/01	26

	Transaction ID	ID_Spec	ID_Loc	Date	Potrebnosti-obshto (3)
35	1162	2037	1141	03/11/01	11
36	1163	2020	1051	08/12/01	5
37	1264	2030	1033	09/01/02	26
38	1265	2027	1124	03/02/02	23
39	1266	2022	1080	08/03/02	5
40	1267	2027	1121	05/04/02	29
41	1268	2001	1063	05/05/02	26
42	1269	2033	1092	09/06/02	5
43	1270	2014	1069	13/07/02	17
44	1271	2004	1051	16/08/02	23
45	1272	2012	1175	01/09/02	26
46	1272	2012	1010	12/10/02	32
47	1272	2012	1010	10/11/02	14
48	1272	2012	1010	11/12/02	29
49	1327	2027	1124	01/01/03	5
50	1328	2027	1119	04/02/03	8
51	1352	2021	1164	03/03/03	14
52	1330	2016	1083	11/04/03	23
53	1331	2006	1047	14/05/03	23
54	1332	2017	1139	17/06/03	20
55	1333	2019	1011	21/07/03	32
56	1333	2019	1066	11/08/03	11
57	1333	2019	1011	11/09/03	29
58	1334	2028	1158	04/10/03	14
59	1352	2021	1164	03/11/03	14
60	1352	2021	1164	03/12/03	14
61	1398	2027	1120	01/01/04	20
62	1399	2027	1123	04/02/04	14
63	1400	2005	1088	07/03/04	29
64	1401	2016	1083	11/04/04	32
65	1402	2006	1043	14/05/04	23
66	1403	2017	1139	12/06/04	23
67	1404	2019	1011	11/07/04	26
68	1404	2019	1015	10/08/04	14
69	1404	2019	1011	10/09/04	32
70	1405	2028	1158	04/10/04	8
71	1406	2016	1083	09/11/04	14
72	1408	2022	1106	01/12/04	14