УДК 519.22.35(075.8)
ББК 22.172я73

Рецензент
д-р фіз.-мат. наук, проф. Козаченко Ю.В.

Майборода Р.Є., Сугакова О.В.

Статистичний аналіз даних за допомогою пакета STATISTICA : Навчальний посібник

Описано статистичний аналіз даних за допомогою пакета STATISTICA. Містить розбір численних прикладів.

Для студентів кваліфікаційного рівня «бакалавр» та «магістр» всіх спеціальностей, які прослухали курс математичної статистики і спрямовані на оволодіння прикладними (комп'ютерними) аспектами статистичного аналізу даних.

Затверджено вченою радою радіофізичного факультету
(протокол №8 від 13 лютого 2012 року)

УДК 519.22.35(075.8)
ББК 22.172я73

Вступ

Сучасну статистику неможливо уявити без комп'ютерної обробки великих масивів даних із застосуванням спеціального програмного забезпечення. Даний посібник являє собою рекомендації до вивчення пакету «STATISTICA6». Головну увагу приділено висвітленню практичного аспекту застосування пакету, містить розбір численних прикладів. Файли з даними, які використовуються в прикладах, або є стандартними і знаходяться в папці STATISTICA6\Examples\Datasets, або пропонуються авторами у вигляді додаткових матеріалів до розробки і знаходяться на сайті http://matphys.rpd.univ.kiev.ua/ukr/courses/mmatstat.html. Відмітимо, що на даний час існують більш нові версії пакету STATISTICA, однак статистичний аналіз даних у розглядуваних нами областях за допомогою цих версій практично не відрізняється від роботи у шостій версії.

Посібник написаний на основі курсів лекцій і лабораторних занять, які автори впродовж багатьох років вели на радіофізичному і механіко-математичному факультетах Київського національного університету імені Тараса Шевченка. Призначений для магістрів першого року навчання всіх спеціальностей або бакалаврів, які прослухали курс математичної статистики, тому не містить розширеного викладення теоретичних засад статистичних методів.

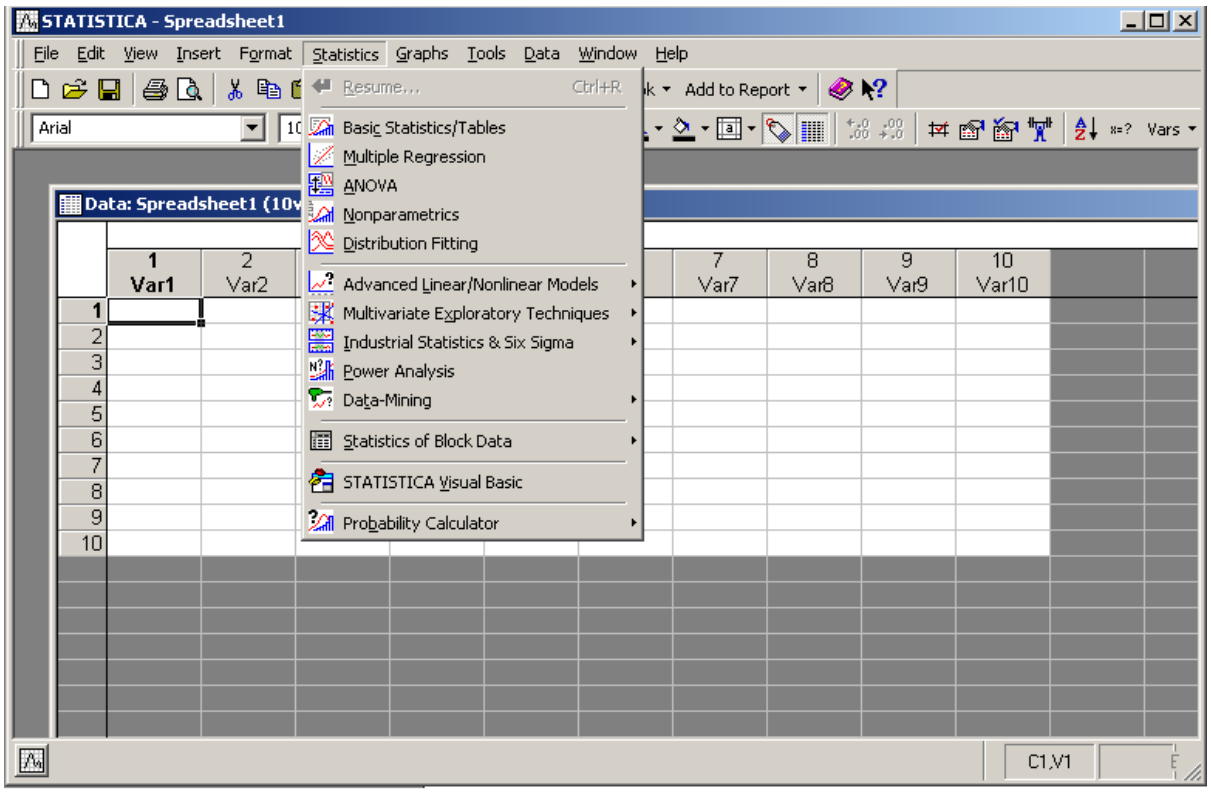
1. Робота з даними

Для того, щоб мати змогу проводити статистичний аналіз даних у пакеті STATISTICA, потрібно або завантажити у пакет який-небудь вже існуючий файл з даними, або створити новий.

Для того, щоб відкрити вже існуючий файл, можна вибрати у головному меню: File->Open. Учбові дані пакету STATISTICA 6 знаходяться у папці STATISTICA6\Examples\Datasets, яка зазвичай буває розміщена у папці Program Files на жорсткому диску комп'ютера.

Для створення нового файлу можна вибрати File->New.

Після того, як файл завантажений або створений, можна почати статистичний аналіз. Для цього виберемо на верхній панелі меню Statistics. Нам пропонують обрати один з модулів для подальшої роботи.



Опис модулів

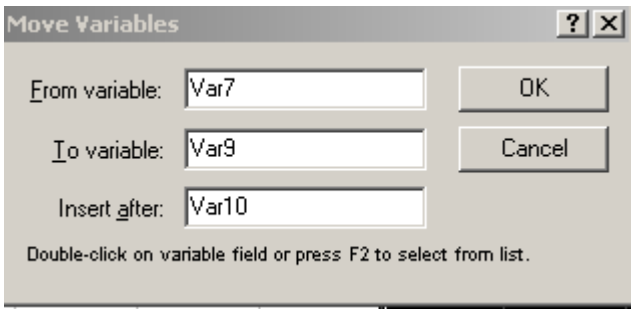
Basic Statistics/Tables	Описова статистика
Multiply Regression	Множинна регресія
ANOVA	Дисперсійний аналіз
Nonparametrics	Непараметричні тести
Distribution Fitting	Підгонка експериментальних кривих
Advanced Linear/Nonlinear Models	Аналіз часових рядів, нестандартні типи регресій і т.д.
Multivariate Exploratory Techniques	Дискримінантний аналіз. Кластерний аналіз. Факторний аналіз.

Робота зі змінними. Стандартна таблиця даних, що створюється за умовчанням, має розмір 10x10, де стовпчики відповідають різним змінним (Var1,…,Var10), рядки – різним спостережуваним об’єктам або випадкам (Cases).

Додавання нової змінної: Vars->Add (або Data->Vars->Add). У вікні, що з’явилося, вказуємо: скільки змінних хочемо додати, після якої змінної, її ім’я і тип даних.

Переміщення змінної: Vars->Move (або Data->Vars->Move).

Вказуємо, з якої по яку змінну хочемо перемістити і після якої змінної їх вставити.



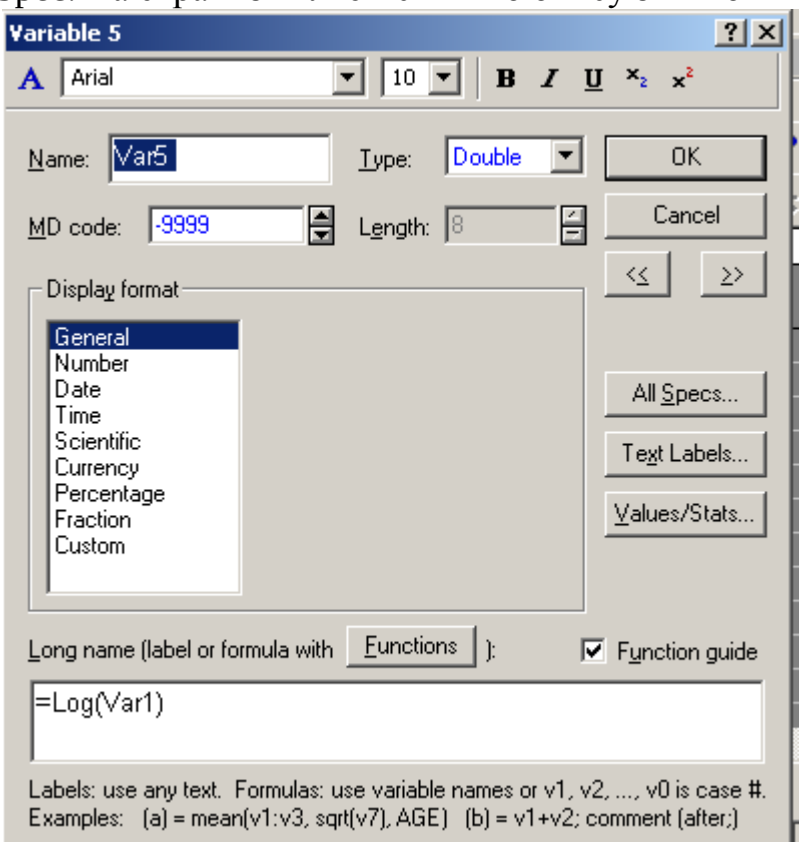
Аналогічно виконуються операції *копіювання і знищення змінної.* (Data->Vars->Copy, Data->Vars->Delete).

Робота з рядками таблиці (Cases)

Над рядками таблиці можна виконувати аналогічні дії за допомогою Cases або Data->Cases->…

Специфікація змінної

Якщо ми хочемо довідатися все про певну наявну змінну, то підводимо курсор до її імені, натискаємо па праву кнопку та обираємо в меню Variable Spec. На екрані з'являється вікно опису змінної



Отже, змінна має такі характеристики:

A	Тип шрифта
Name	Ім'я
Type	Тип даних – число подвійної точності, байт, ціле число, текст
MD code	(missing data code) – значення, яке за умовчанням присвоюється змінній, якщо її справжнє значення відсутнє з якихось причин.

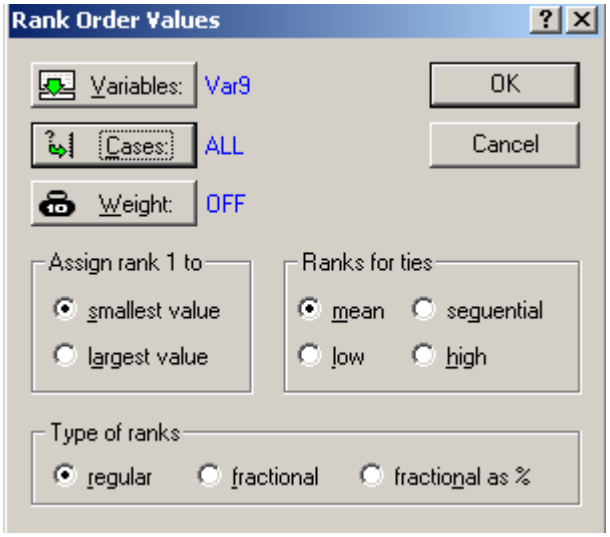
Length	Ширина колонки даної змінної.
Display format	Формат відображення числа.
Long name	Поле, в якому можна задавати формулу для обчислення значення даної змінної.

Кнопки <<,>> здійснюють перехід до специфікацій сусідніх змінних.
All specs відкриває таблицю з усіма специфікаціями змінних.

Values/Stats відображає «швидку статистику» - середнє, стандартне відхилення та ін.

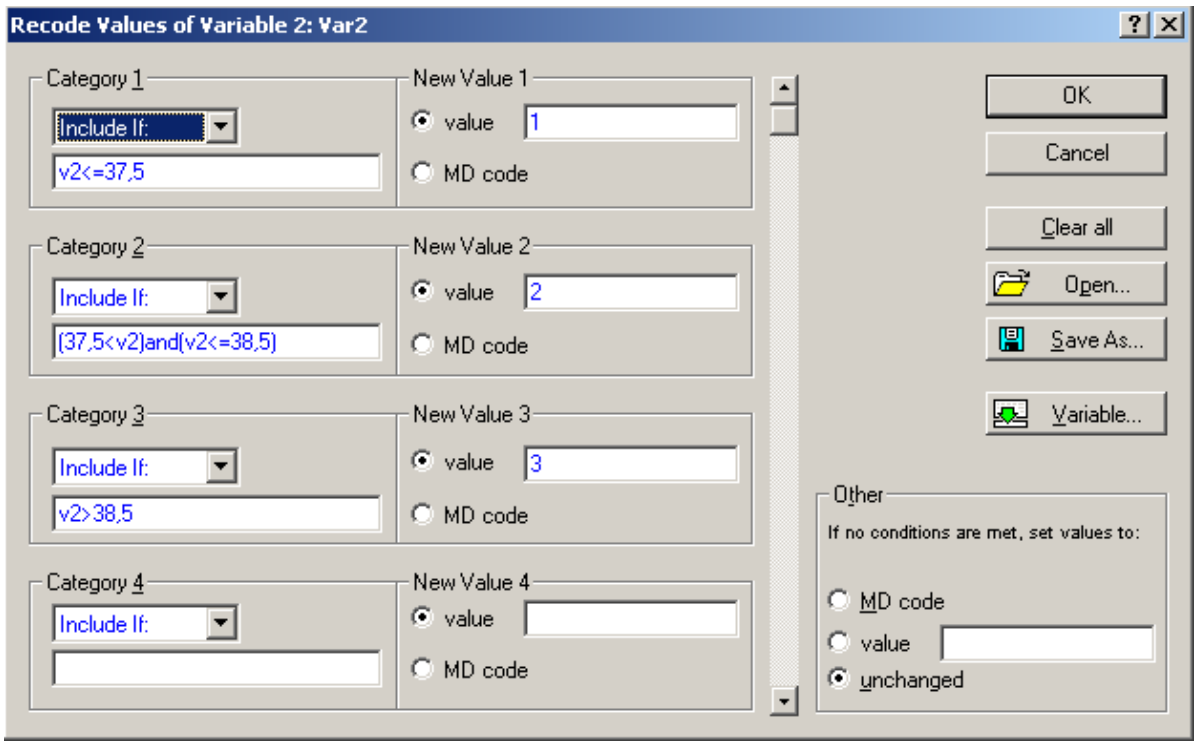
Якщо ви змінили дані і треба перерахувати якусь змінну, треба натиснути Vars->Recalculate.

Ранжування змінної: (Це впорядкування за зростанням чи спаданням. Воно виражається в присвоєнні кожному значенню певного *рангу* – порядкового номера в списку впорядкованих значень.) Натискаємо Var->Rank.

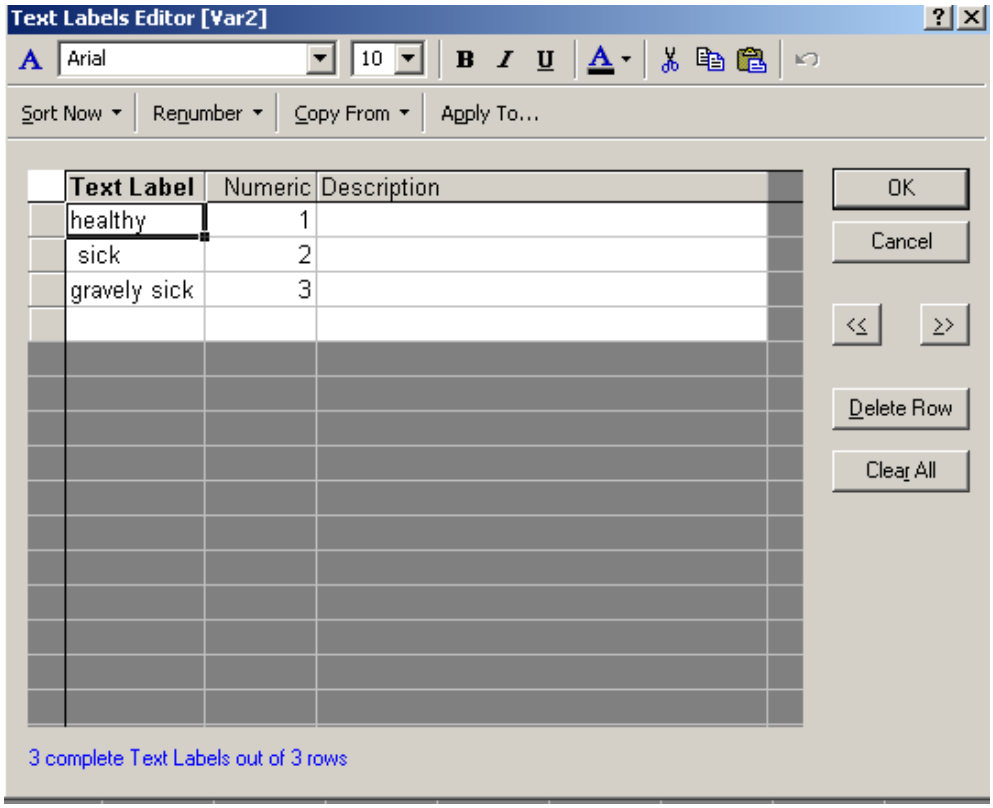


Обираємо: ранжування за зростанням чи спаданням за допомогою Assign rank 1 to, Rank for ties: sequantial – якщо ми хочемо, щоб однакові значення мали послідовні значення рангу; mean – якщо ми хочемо, щоб однакові значення мали однаковий усереднений ранг (тобто середнє арифметичне від того, що отримано у випадку sequantial).

Інколи потрібно значення змінної розбити на групи (наприклад, в залежності від температури віднести людину до групи здорових, хворих чи тяжкохворих). Значення температури знаходиться в Var1. Копіюємо в Var2, натискаємо Vars->Recode. Заповнюємо поля наступним чином



Бачимо, що всі значення, які були не більші за 37,5, замінилися на 1 (здорові), ті, що були в межах від 37,5 до 38,5, замінилися на 2(хворі), ті, що були більші за 38,5 – стали 3 (тяжкохворі). А тепер ми хочемо перейменувати групи, щоб було зручніше з ними працювати. Переходимо Data->Text Labels Editor і вводимо назви груп для відповідних номерів. Замість 1,2,3 з'являться назви груп.



Заповнення випадковими числами: якщо виділити стовпчик зі змінною, натиснути на праву кнопку миші, перейти Fill/Standartize Block->Fill Random Values, то змінна заповниться значеннями випадкової величини, рівномірно розподіленої на [0,1].

Заповнення однаковим значенням: набираємо значення, яким ми хочемо заповнити змінну, в першій клітинці. Виділяємо стовпчик, клацаємо на праву кнопку миші, Fill/Standartize Block->Fill/Copy Down. З рядками можна оперувати аналогічно, тільки виділити рядок, далі Fill/Standartize Block->Fill/Copy Right.

Заповнення арифметичною прогресією: в перших двох клітинках вводимо два перших члени арифметичної прогресії, виділяємо ці клітинки, переміщуємо курсор у правий нижній кут нижньої клітинки – він перетворюється на хрестик, і тягнемо вниз з натиснутою лівою кнопкою миші до кінця стовпчика.

Зсув всіх даних вниз на кілька позицій: виділити стовпчик, Vars->Shift(Lag).

Стандартизація змінної (перетворення $\frac{x_i - \bar{x}}{\sigma}$, де $\bar{x}$ - середнє, σ - середньоквадра-тичне відхилення): Vars->Standartize.

Генерація даних, розподілених за певним законом. Якщо ми хочемо якусь змінну заповнити, наприклад, вибіркою, нормально розподіленою з параметрами $a = 0; \sigma = 1$, то ми повинні в специфікації змінної в полі Long Name написати =VNormal(rnd(1);0;1). Якщо заповнення не відбудеться відразу, то треба перерахувати змінну за допомогою Recalculate.

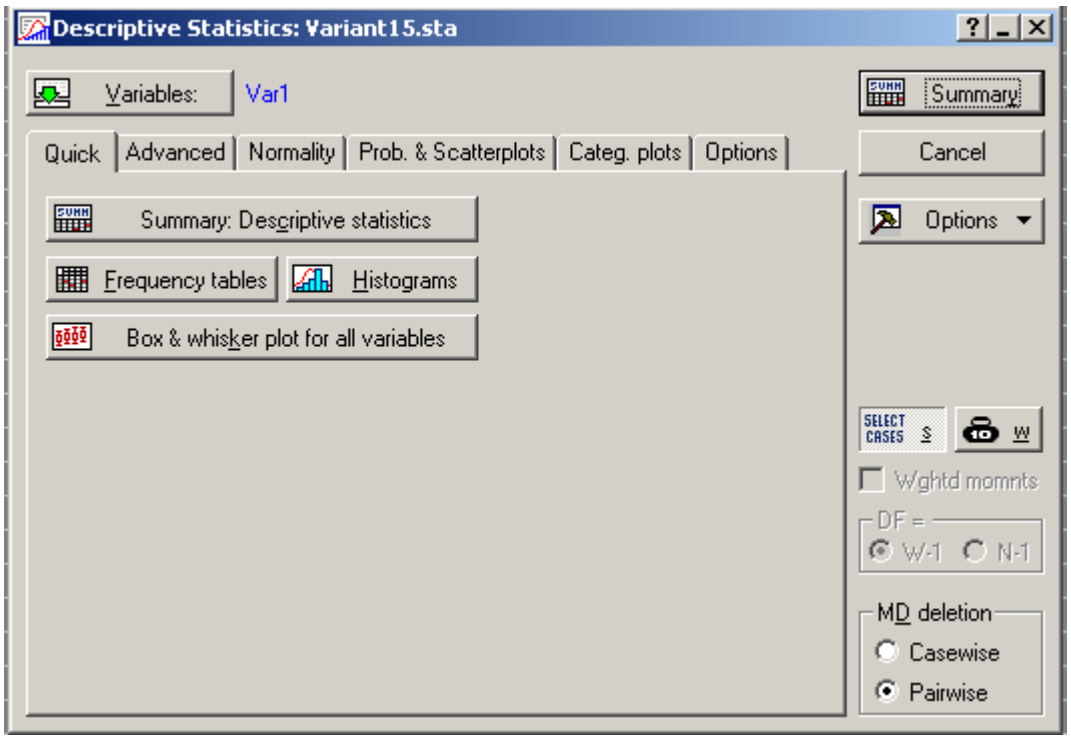
Інші закони розподілу задаються аналогічно, наприклад:

=rnd(5)	Рівномірний розподіл на $[0,5]$
=VNormal(rnd(1);2;3)	Нормальний з параметрами $a = 2; \sigma = 3$
=VExpon(rnd(1);1/2)	Експоненціальний розподіл з $\lambda = 1/2$
=VLognormal(rnd(1);0,5;0,5)	Логнормальний з параметрами 0,5;0,5
=VCauchy(rnd(1);0;1)	Розподіл Коші з параметрами 0;1.
=VChi2(rnd(1);8)	Розподіл χ^2 з 8 степенями свободи

Для створення звіту - файлу, в якому будуть записані всі результати, натискаємо File-> Output Manager, відмічаємо Also send to Report Window, Single Report. Файли з даними у вигляді таблиць мають розширення .sta, а файли робочих книг та звітів .stw.

Вилучення викидів

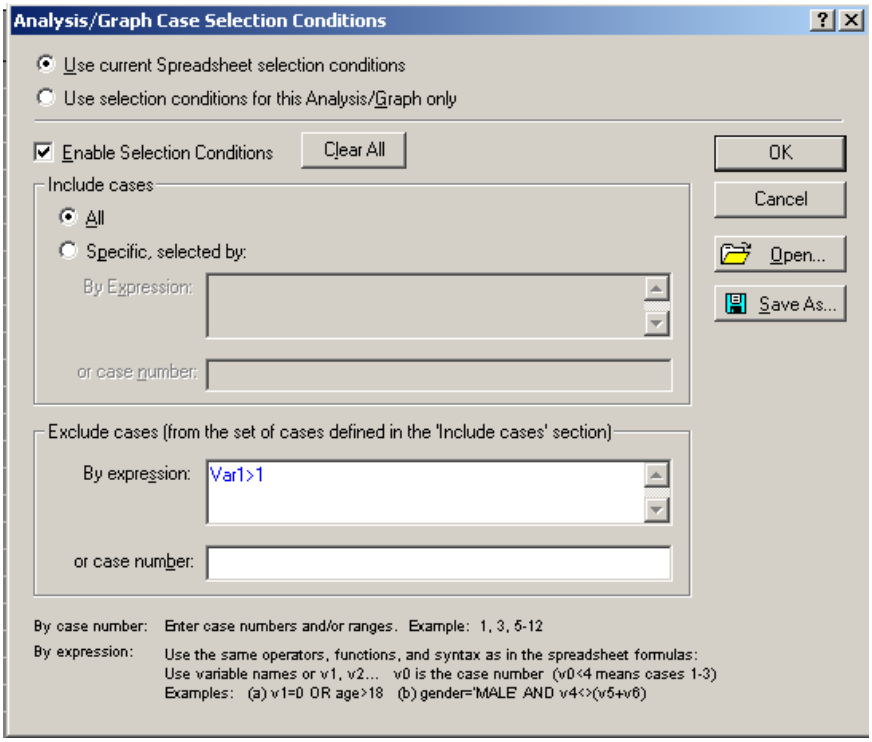
Викиди можна побачити на P-P та Q-Q діаграмах, вусатих коробочках. Якщо ми хочемо вилучити з розгляду викид, наприклад, дуже велике значення, не прибираючи його з даних, то можна перейти Statistics->Basic Statistics/Tables->Descriptive statistics. Відмічаємо змінну.



На закладці Quick натиснути на кнопку Summary:Descriptive statistics.

Variable	Descriptive Statistics (Variant15.sta)				
	Valid N	Mean	Minimum	Maximum	Std.Dev.
Var1	100	8,060020	0,000293	800,0000	79,99399

Одразу бачимо величину викиду – 800 – як максимальне значення з елементів вибірки. Вилучимо з нашого подальшого розгляду всі значення, які більше за 1. Для цього натиснемо на кнопку CASES поряд з гирею.



Відмітимо галочкою Enable Selection Conditions. У віконці Exclude cases By expression поставимо умову Var1>1. Натискаємо ОК. Тепер елементи вибірки, >1, з розгляду вилучаються. Тепер значення максимуму, як ми бачимо, =0,800001.

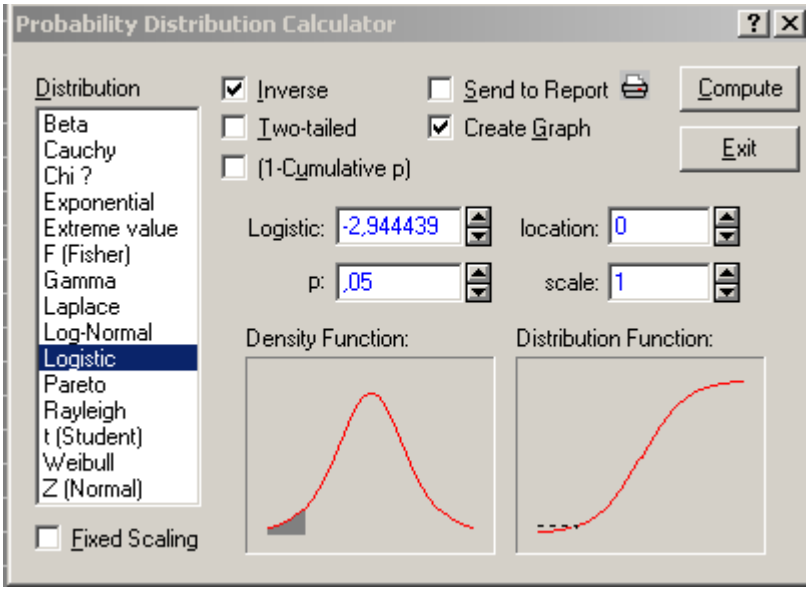
Descriptive Statistics (Variant15.sta)

Variable	Valid N	Mean	Minimum	Maximum	Std.Dev.
Var1	99	0,060626	0,000293	0,800001	0,089067

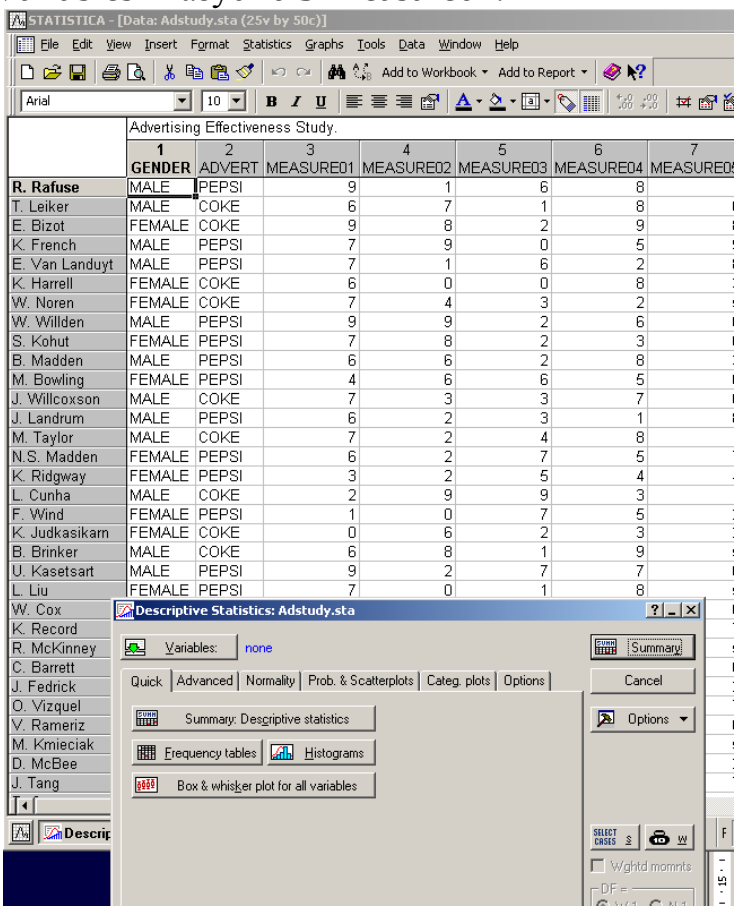
2. Описова статистика

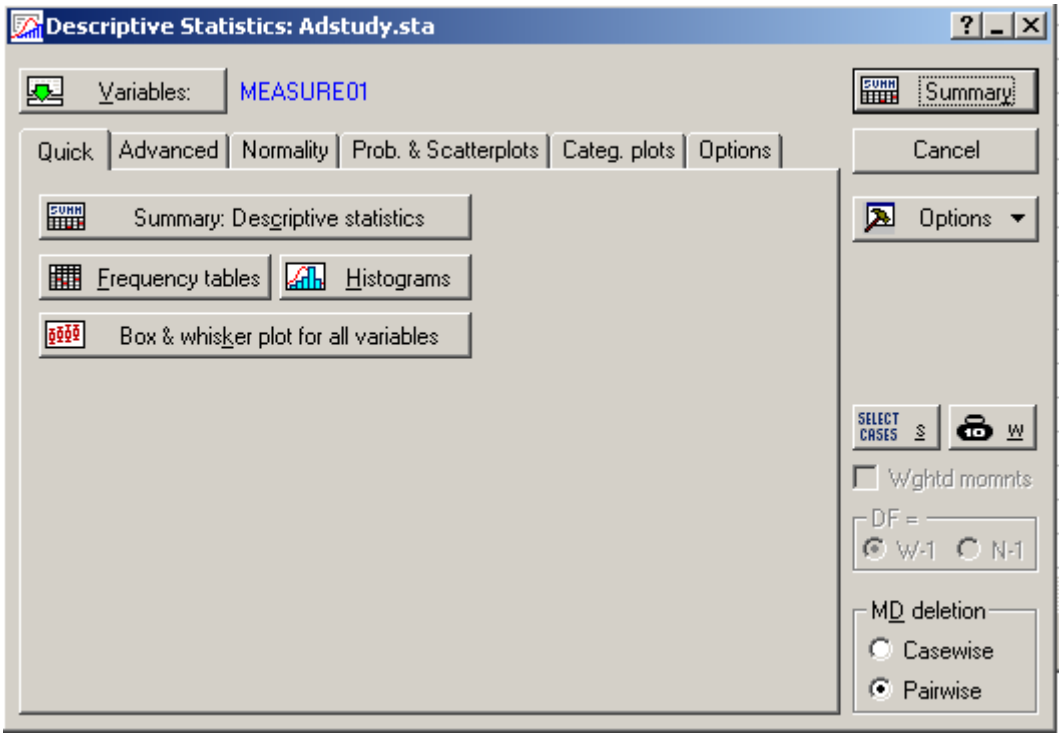
Робота з калькулятором : якщо треба знайти квантиль певного розподілу, подивитись на графік теоретичної функції розподілу або обчислити її значення в певній точці, використовують ймовірнісний калькулятор Statistics->Probability Calculator->Distribution

Панель Distribution дозволяє подивитись на графік щільності та функції розподілу, параметри розподілу задаємо в останньому стовпчику справа. Для обчислення значення функції розподілу задаємо значення, натискаємо Compute, отримуємо p. Для квантилю рівня p аналогічно: набираємо p, натискаємо Compute. Також можна отримати двосторонню квантиль (two-tailed). Якщо відмітити Create Graph і натиснути Compute, отримаємо графік.

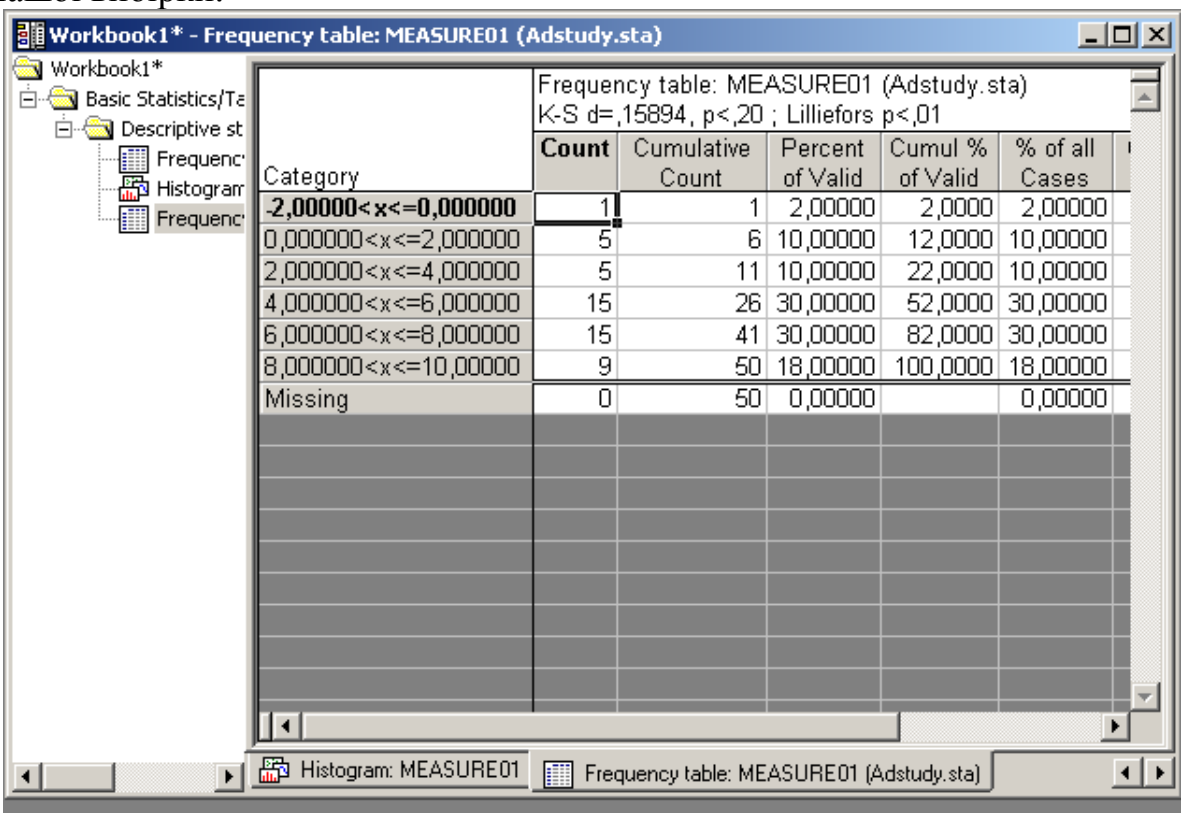


Приклад 2.1 Відкриємо файл Adstudy.sta, в якому зібрані дані про оцінки чоловіками і жінками реклами напоїв Pepsi та Coke. Кожний опитуваний оцінював рекламу по різних показниках, виставляючи оцінку від 1 до 9. Натиснемо Statistics-> Basic Statistics/Tables->Descriptive Statistics. В полі Variables вказуємо 3-Measure01.





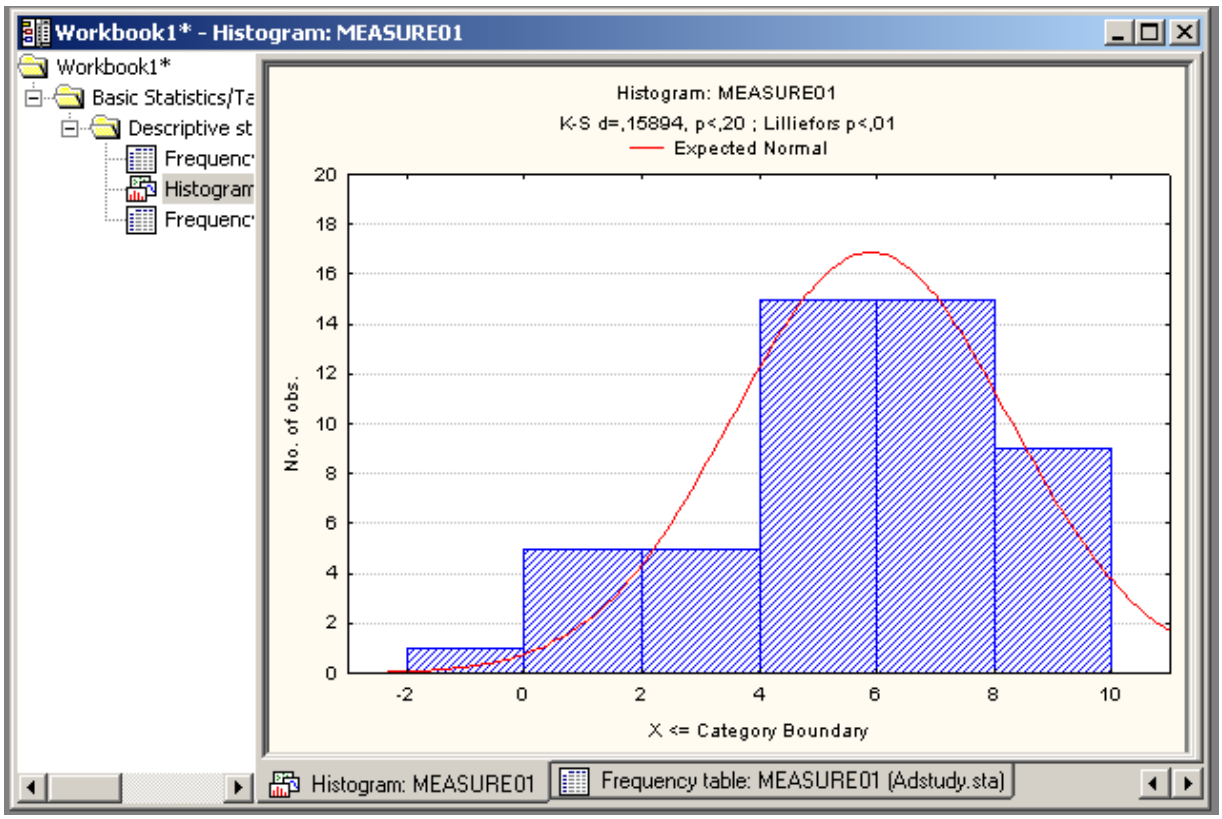
Найбільш вживані описові статистики знаходяться на вкладці Quick. Натиснувши кнопку Frequency Tables, отримаємо таблицю частот для нашої вибірки.



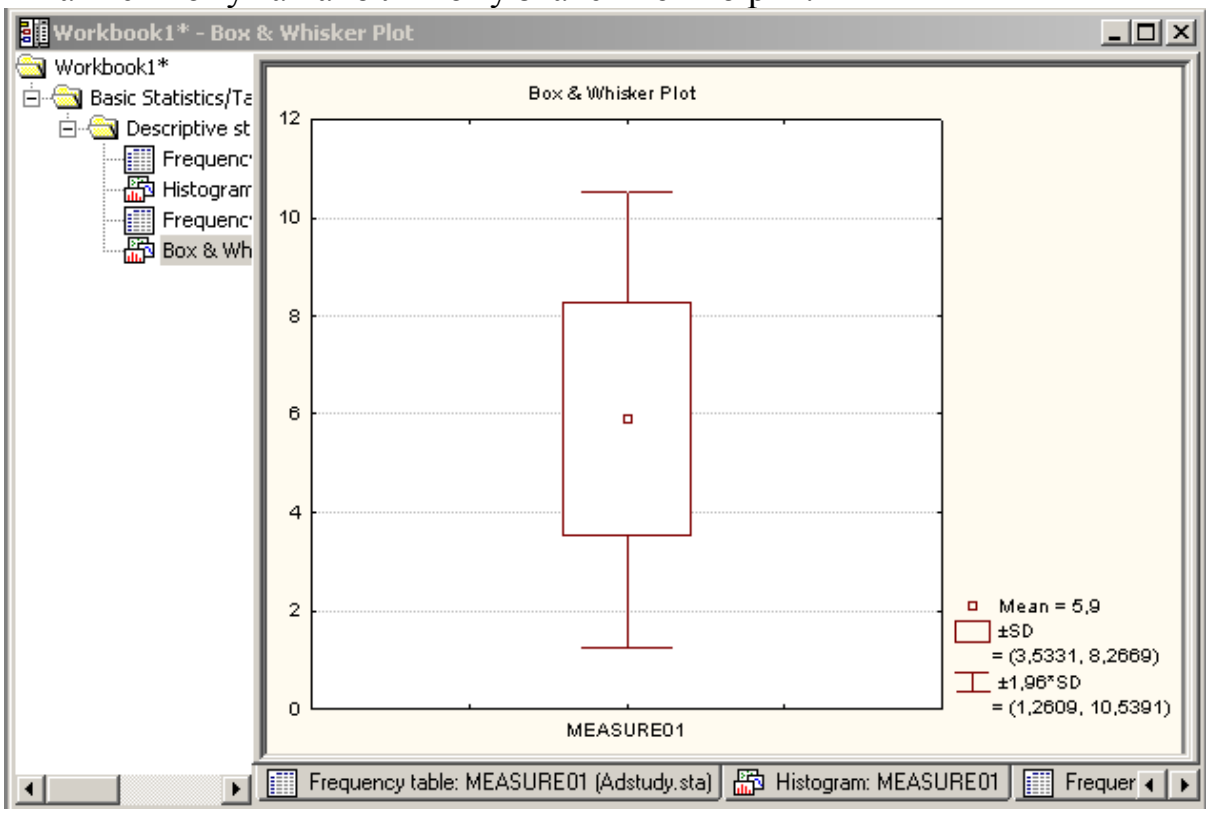
Frequency table: MEASURE01 (Adstudy.sta)
K-S d=,15894, p<,20 ; Lilliefors p<,01

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases
-2,00000<x<=0,000000	1	1	2,00000	2,0000	2,00000
0,000000<x<=2,000000	5	6	10,00000	12,0000	10,00000
2,000000<x<=4,000000	5	11	10,00000	22,0000	10,00000
4,000000<x<=6,000000	15	26	30,00000	52,0000	30,00000
6,000000<x<=8,000000	15	41	30,00000	82,0000	30,00000
8,000000<x<=10,00000	9	50	18,00000	100,0000	18,00000
Missing	0	50	0,00000		0,00000

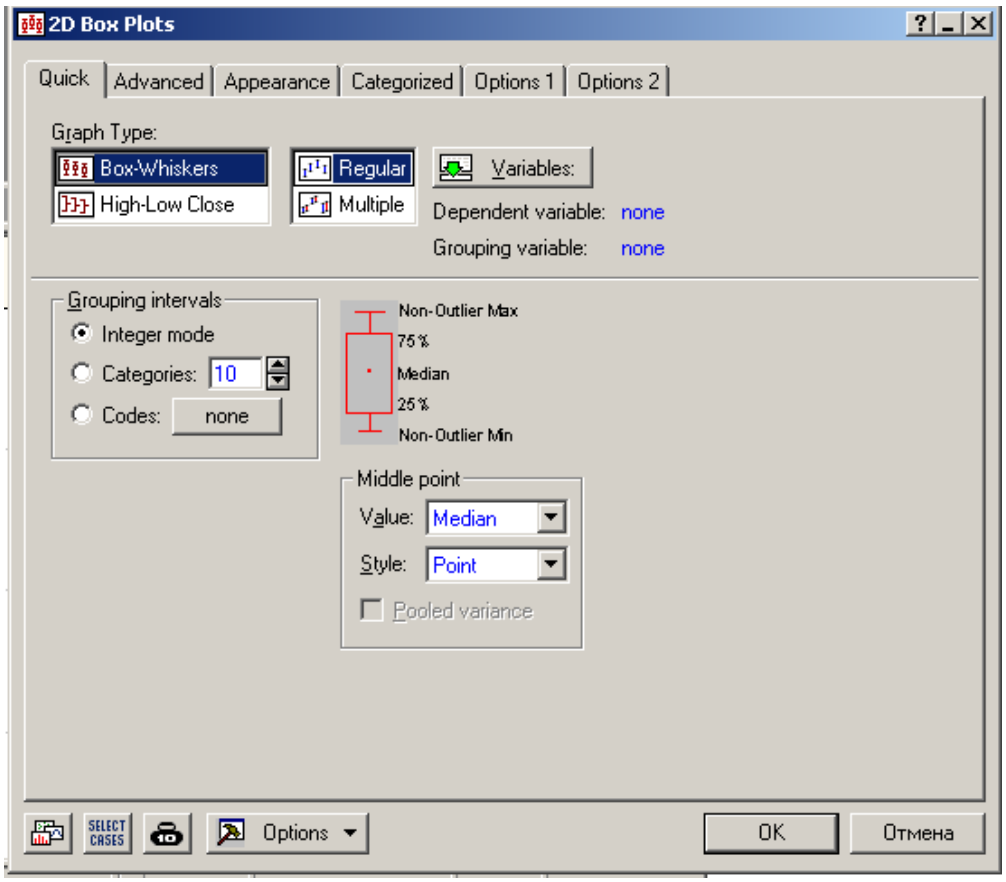
Натиснувши кнопку Histograms, отримаємо гістограму, на якій знаходиться підігнана крива нормального розподілу.



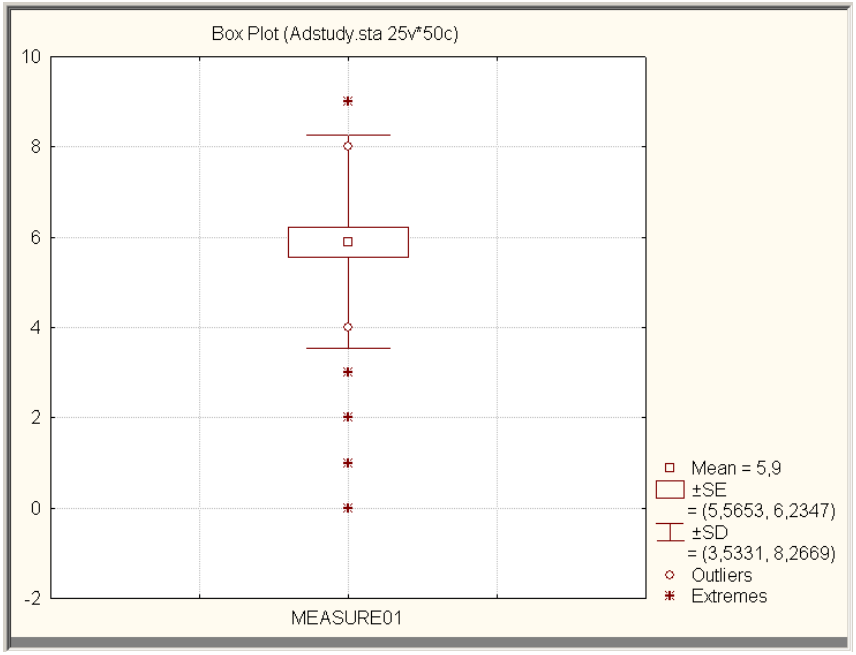
Натиснемо Box&whisker plot for all variables – з'явиться вікно з рисунком коробки з вусами, в якій маленький прямокутник відповідає значенню медіани, великий прямокутник – верхній та нижній квартилі, а вуса – найменшому та найбільшому значенню вибірки.



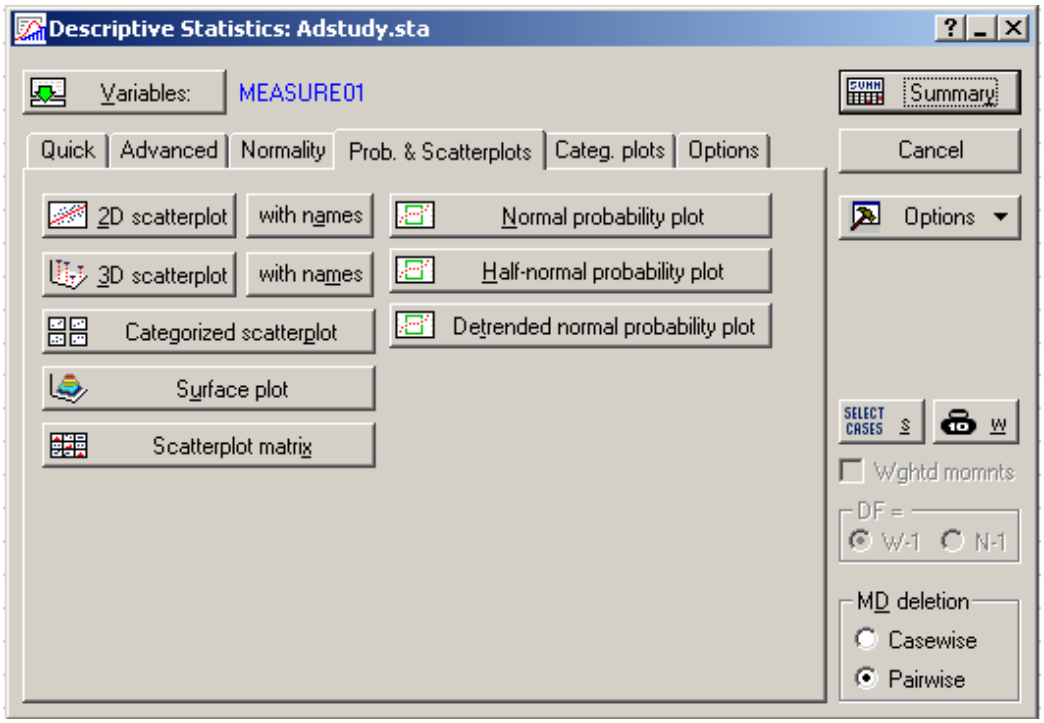
Зазначимо, що вусаті коробочки можна будувати в інший спосіб. Для цього перейдемо Graphs->2DGraphs->Box Plots.



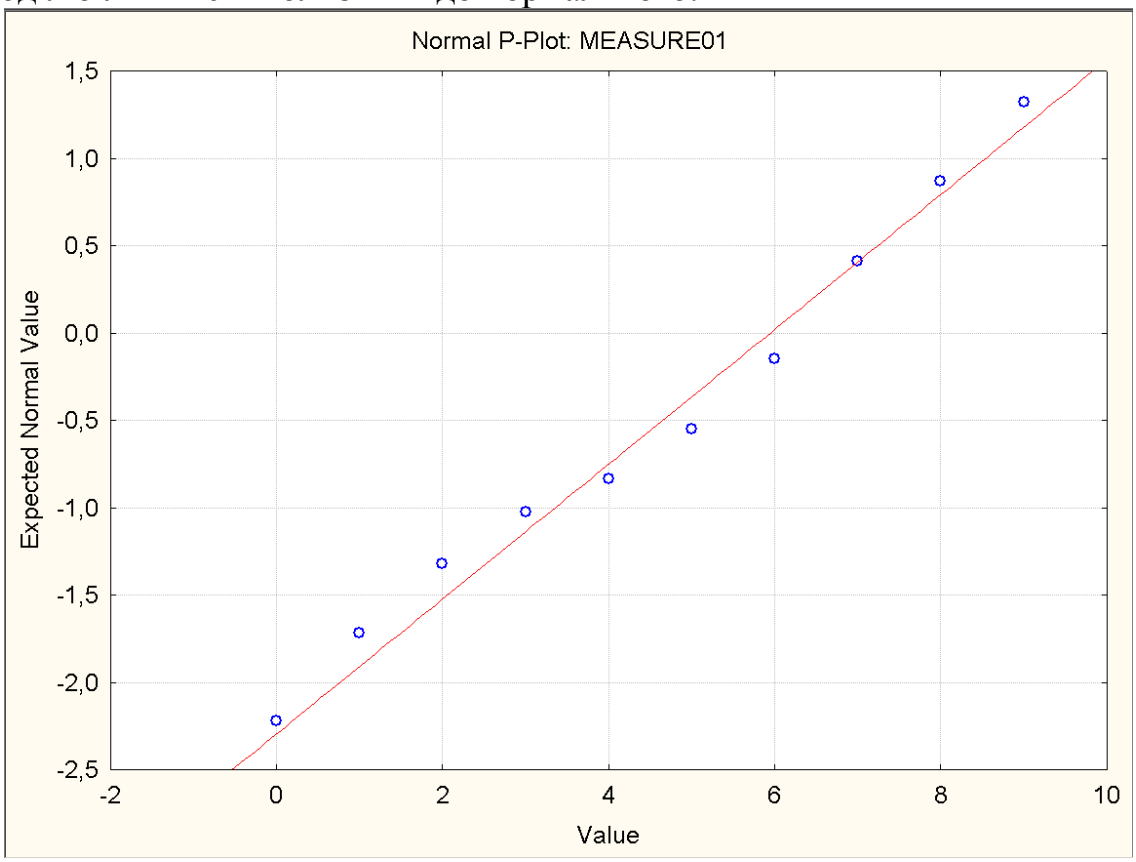
Як ми бачимо, нам пропонують «вусату коробочку» такого самого вигляду, як ми будували. А ми, наприклад, хочемо, щоб центральна точка показувала не медіану, а середнє, мат. сподівання, прямокутник відповідав би – Mean-SE та Mean+SE, а вуса відповідали б Mean-SD та Mean+SD, де SE-standard error (дорівнює $\sqrt{D_B}/\sqrt{n}$, n – об'єм вибірки, D_B - вибіркова дисперсія); SD-standard deviation (дорівнює $\sqrt{D_B}$). Для цього обираємо Middle point Value: Mean. Потім ОК. Вусата коробочка набуде вигляду, як на малюнку.



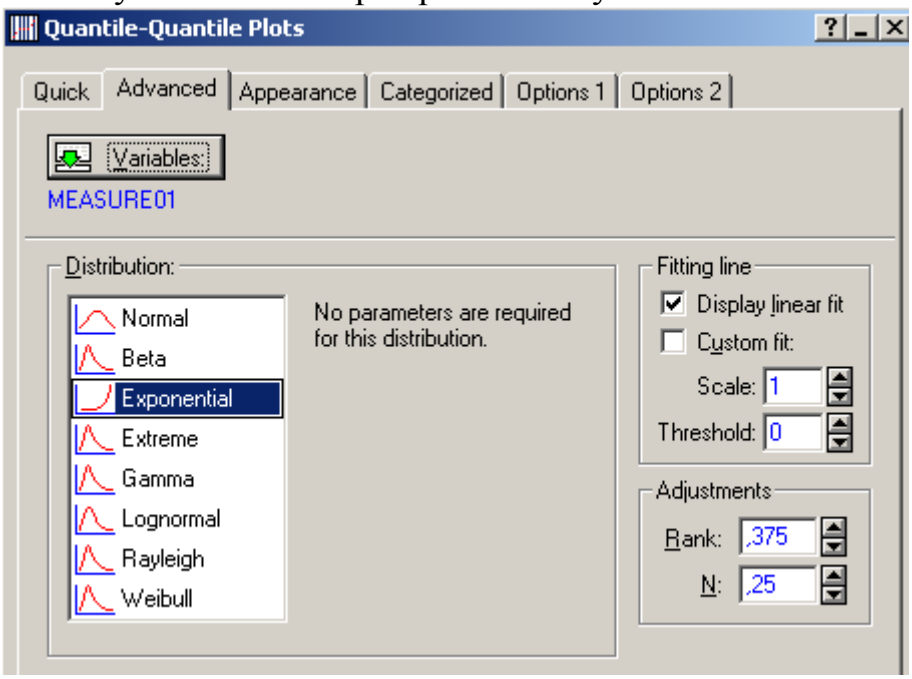
Вусаті коробочки можна будувати на одному графіку за допомогою закладки Categorized .Для побудови P-P діаграми у вікні Descriptive Statistics Prob.& Scatterplots натиснемо Normal Probability Plots.



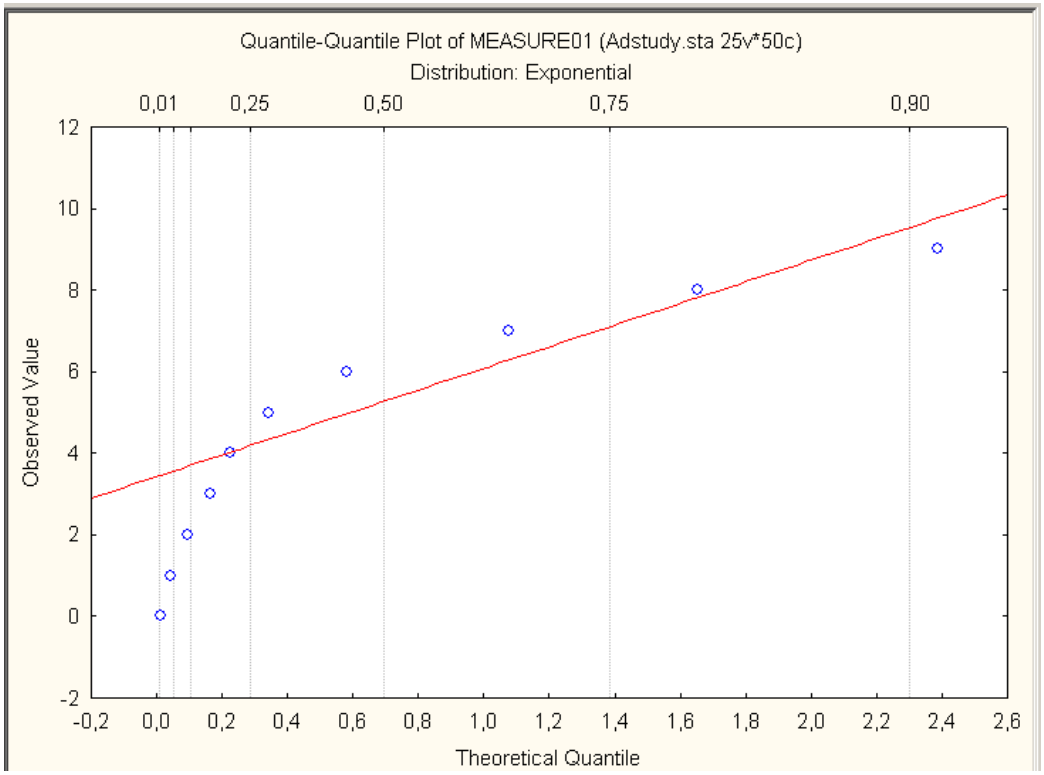
Точки на діаграмі лежать поблизу червоної лінії, отже, бачимо, що розподіл більш-менш близький до нормального.



Для побудови Q-Q діаграми переходимо Graphs->2D Graphs->Quantile-Quantile Plots. Перейдемо на закладку Advanced, за допомогою Variables задаємо змінну MEASURE01. В полі Distribution вибираємо розподіл, на відповідність якому ми хочемо перевірити змінну. ОК.



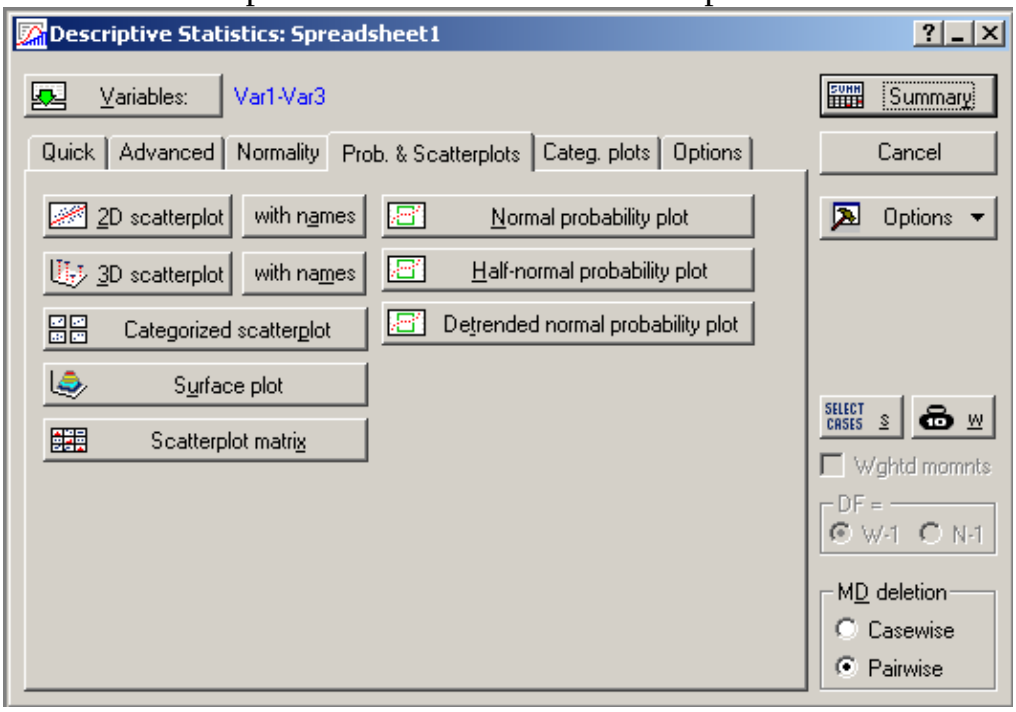
На діаграмі видно, що експоненційний розподіл погано відповідає нашим даним.



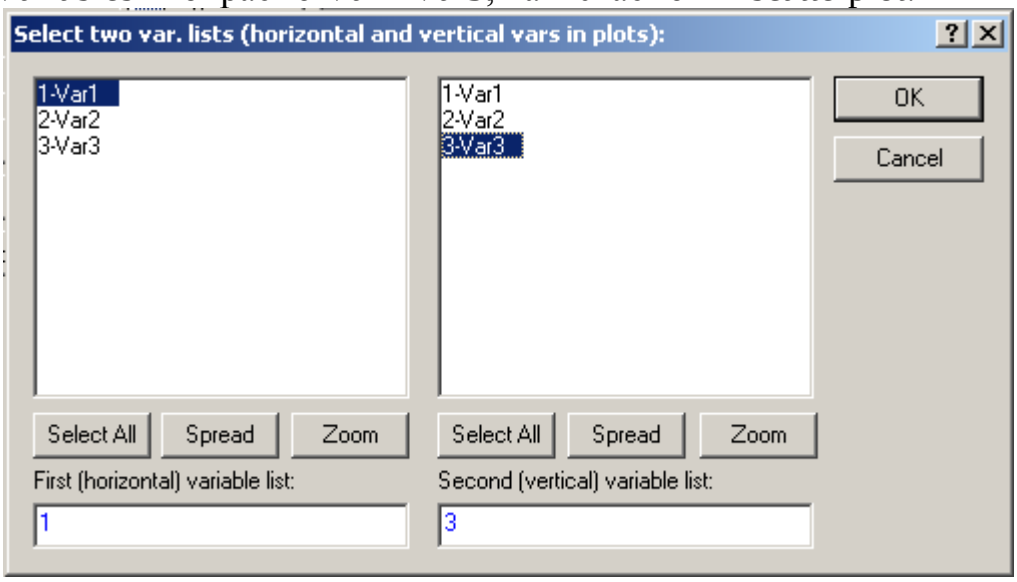
3. Проста лінійна регресія

Приклад 3.1 Створимо новий файл, в якому змінну Var1 заповнимо послідовно значеннями від 0 до 10, змінну Var2 – випадковими значеннями від 0 до 1, а Var3 задамо як Var1+Var2.

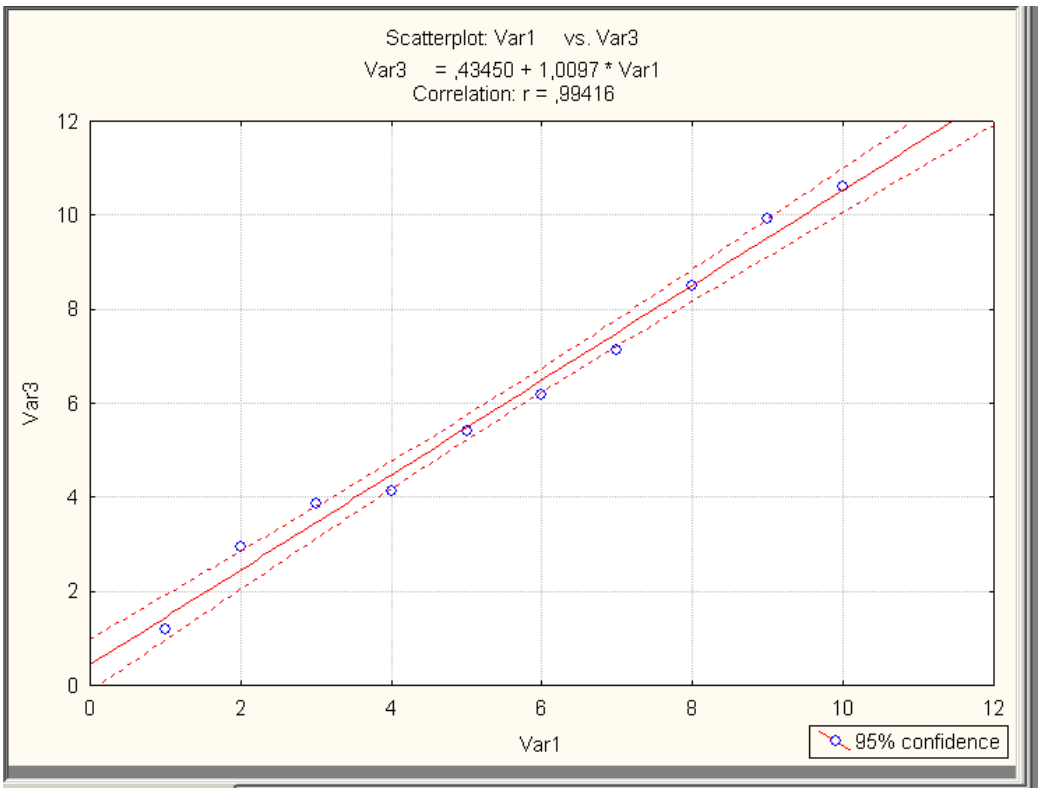
Передусім ми хочемо побудувати діаграму розсіювання даних, а також графік регресійної прямої Var3 на Var1. Перейдемо Statistics->Basic statistics/Tables-> Descriptive statistics->Prob.&Scatterplots.



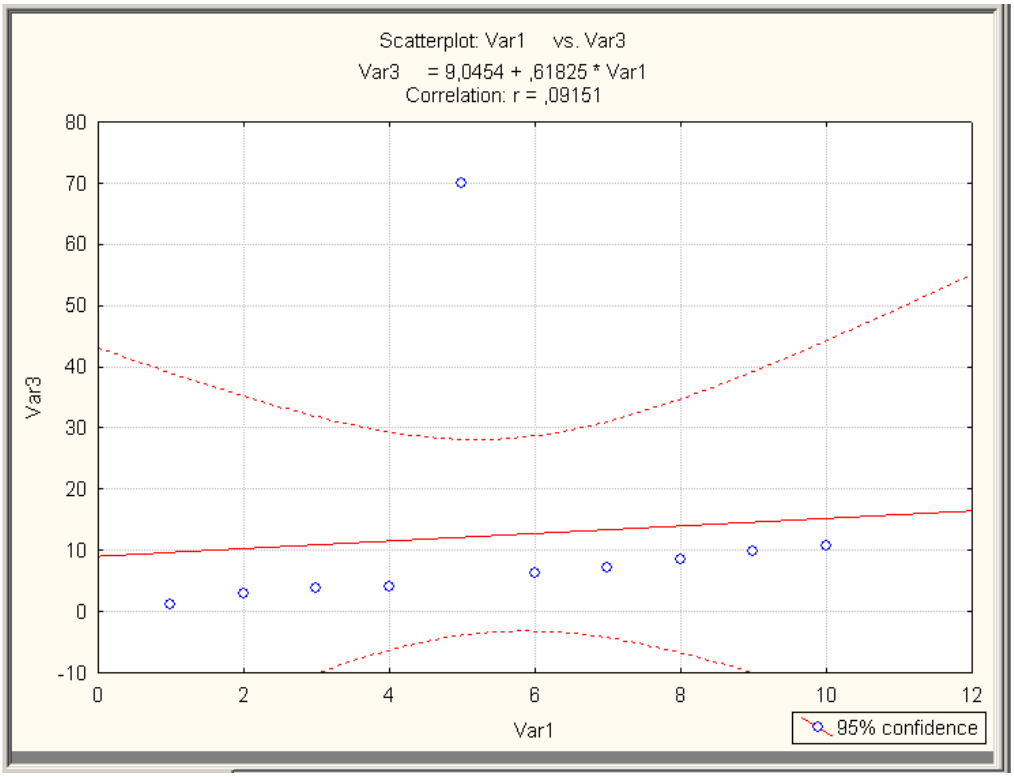
Як Variables вибираємо Var1-Var3, натискаємо 2D scatterplot.



В першому списку змінних вказуємо Var1, в другому – Var3, потім ОК. З'являється графік діаграми розсіювання і регресійної прямої, а також рівняння відповідної регресії.

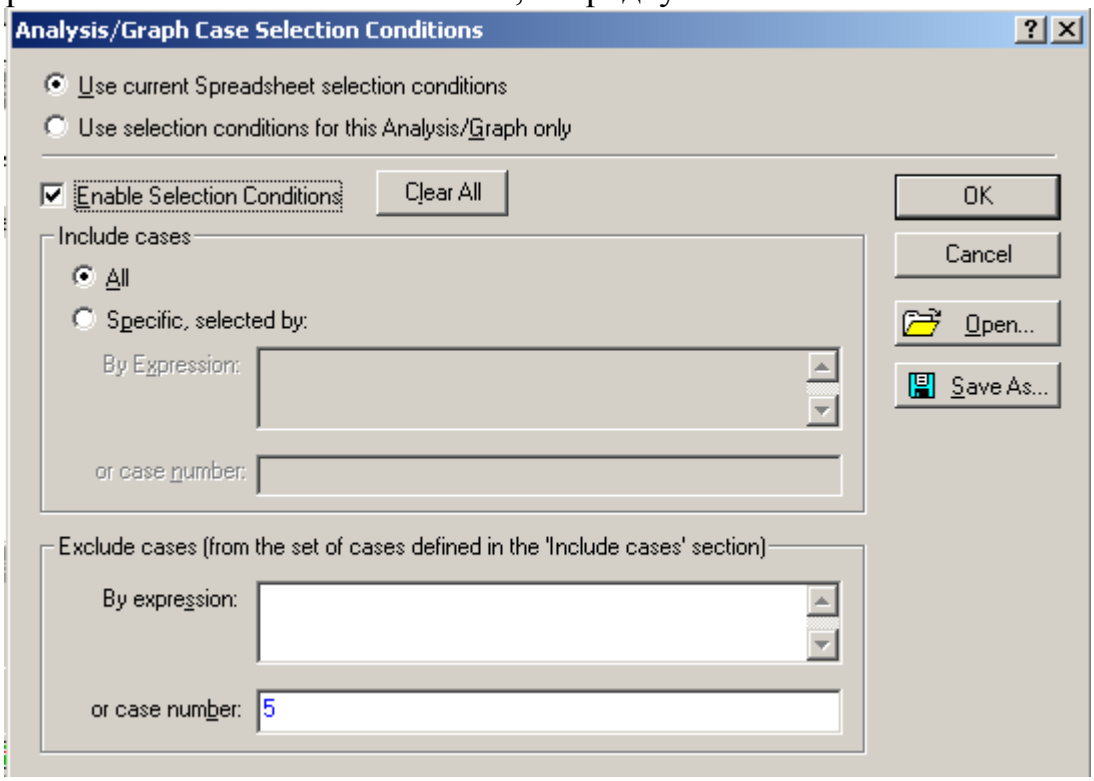


Якщо у змінній Var3 замінити одне зі значень, наприклад, на 70 і побудувати графік знову, то він буде геть іншим.



Очевидно, що 70 є викидом. Спробуємо вилучити викид з даних: побудуємо регресійну пряму без врахування 5-го спостереження, яке

відповідає значенню 70. Для цього натиснемо кнопку SELECT CASES, яка знаходиться поряд з гирею у вікні Descriptive statistics. Поставимо галочку навпроти Enable Selection Conditions, а в рядку or case number поставимо 5.



Натискаємо ОК, а потім знов будуємо діаграму розсіювання – 2D scatterplot. Діаграма знов буде схожа на ту, яка була перед введенням викиду.

4. Множинна регресія

Приклад 4.1 Нижче наведені дані про капітальні витрати на будівництво атомних електростанцій з реактором водяного охолодження. Дані зібрані для 32 станцій США.

№	Price	D	T1	T2	S	PR	NE	CT	BW	N	PT
1	460.05	68.58	14	46	687	0	1	0	0	14	0
2	452.99	67.33	10	73	1065	0	0	1	0	1	0
3	443.22	67.33	10	85	1065	1	0	1	0	1	0
4	652.32	68.00	11	67	1065	0	1	1	0	12	0
5	642.23	68.00	11	78	1065	1	1	1	0	12	0
6	345.39	67.92	13	51	514	0	1	1	0	3	0
7	272.37	68.17	12	50	822	0	0	0	0	5	0
8	317.21	68.42	14	59	457	0	0	0	0	1	0
9	457.12	68.42	15	55	822	1	0	0	0	5	0
10	690.19	68.33	12	71	792	0	1	1	1	2	0
11	350.63	68.58	12	64	560	0	0	0	0	3	0
12	402.59	68.75	13	47	790	0	1	0	0	6	0

13	412.18	68.42	15	62	530	0	0	1	0	2	0
14	495.58	68.92	17	52	1050	0	0	0	0	7	0
15	394.36	68.92	13	65	850	0	0	0	1	16	0
16	423.32	68.42	11	67	778	0	0	0	0	3	0
17	712.27	69.50	18	60	845	0	1	0	0	17	0
18	289.66	68.42	15	76	530	1	0	1	0	2	0
19	881.24	69.17	15	67	1090	0	0	0	0	1	0
20	490.88	68.92	16	59	1050	1	0	0	0	8	0
21	567.79	68.75	11	70	913	0	0	1	1	15	0
22	665.99	70.92	22	57	828	1	1	0	0	20	0
23	621.45	69.67	16	59	786	0	0	1	0	18	0
24	608.80	70.08	19	58	821	1	0	0	0	3	0
25	473.64	70.42	19	44	538	0	0	1	0	19	0
26	697.14	71.08	20	57	1130	0	0	1	0	21	0
27	207.51	67.25	13	63	745	0	0	0	0	8	1
28	288.48	67.17	9	48	821	0	0	1	0	7	1
29	284.88	67.83	12	63	886	0	0	0	1	11	1
30	280.36	67.83	12	71	886	1	0	0	1	11	1
31	217.38	67.25	13	72	745	1	0	0	0	8	1
32	270.71	67.83	7	80	886	1	0	0	1	11	1

Price — Вартість будівництва в млн. $ у цінах 1976 року. **D** — час, за планом відведений на будівництво. **T1** — час між звертанням за дозволом і отриманням дозволу на будівництво. **T2** — час між отриманням ліцензії та дозволом на будівництво. **S** — номінальна потужність електростанції, Мвт. **PR** — наявність в тій же місцевості раніше побудованої електростанції (якщо значення = 1, то маємо вже побудовану станцію). **NE** — характеристика (в геологічному відношенні) району, в якому планується побудувати станцію. **CT** — використання нагрівальної башти (якщо = 1, то використовується, якщо 0 — ні). **BW** — використання силової установки виробництва фірми Babcock-Wilcox (якщо значення = 1, то використовується установка цієї фірми, 0 — ні). **N** — сумарна кількість електростанцій, які побудував архітектор-інженер станції. **PT** — електростанції, які будуються під частковим наглядом (PT=1, якщо нагляд є, PT=0**,** якщо нагляду нема).

Необхідно: 1) Побудувати лінійну регресійну модель залежності ціни Price від змінних *D, T1, T2, S, PR, NE, CT, BW, N, PT* .

2) Залучити до моделі тільки найбільш значущі змінні, проаналізувати залежність саме від них.

2) Проаналізувати залишки моделі.

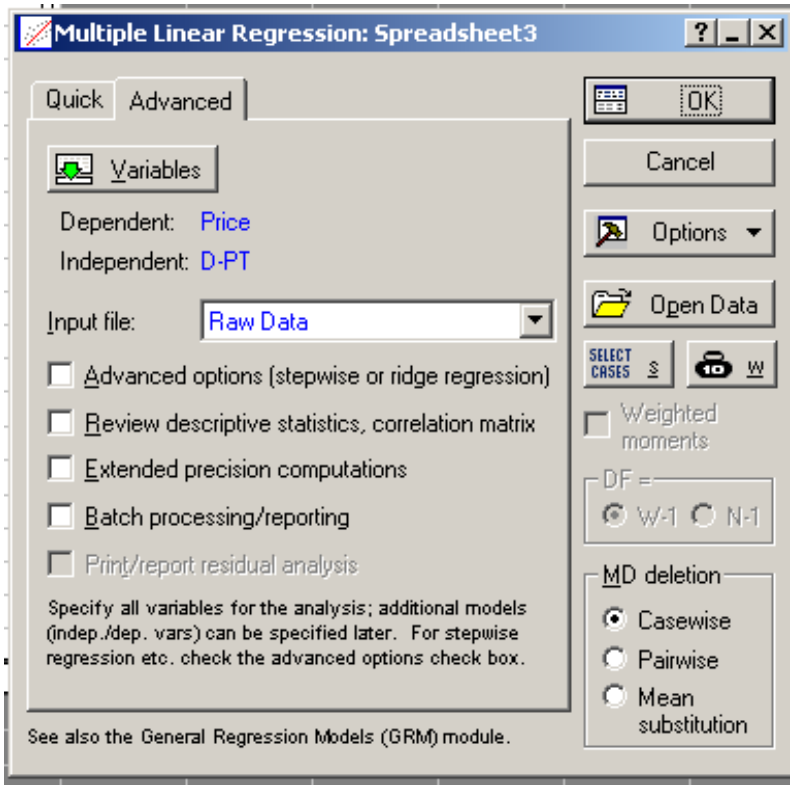
3) Проаналізувати викиди даних.

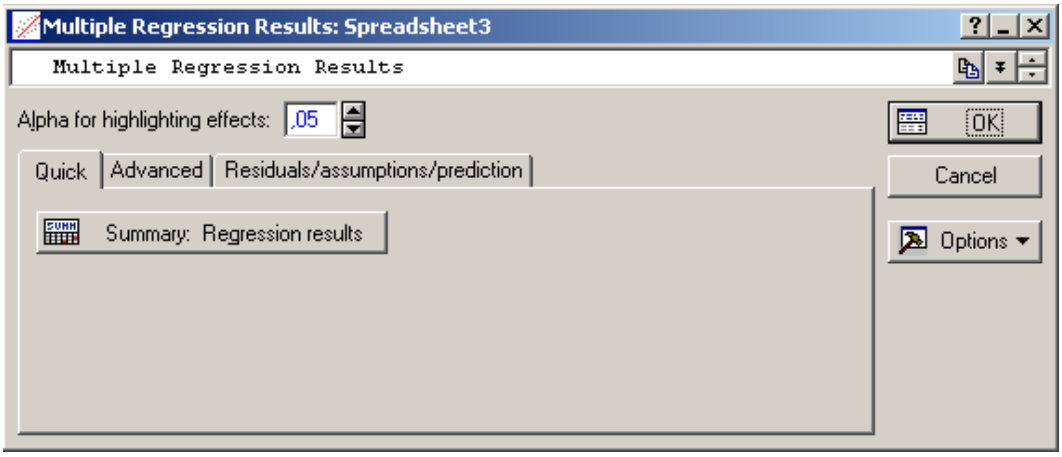
4)Передбачити величину капітальних витрат на будівництво нової станції.

1) Отже, специфікація моделі задається у вигляді

$$\Pr ice = b_0 + b_1 D + b_2 T1 + b_3 T2 + b_4 S + b_5 PR + b_6 NE + b_7 CT + b_8 BW + b_9 N + b_{10} PT + \varepsilon$$

Виконаємо послідовність команд Statistics->Multiple Regression, вкажемо, що Price – залежна змінна, всі інші-незалежні. Натиснемо ОК.

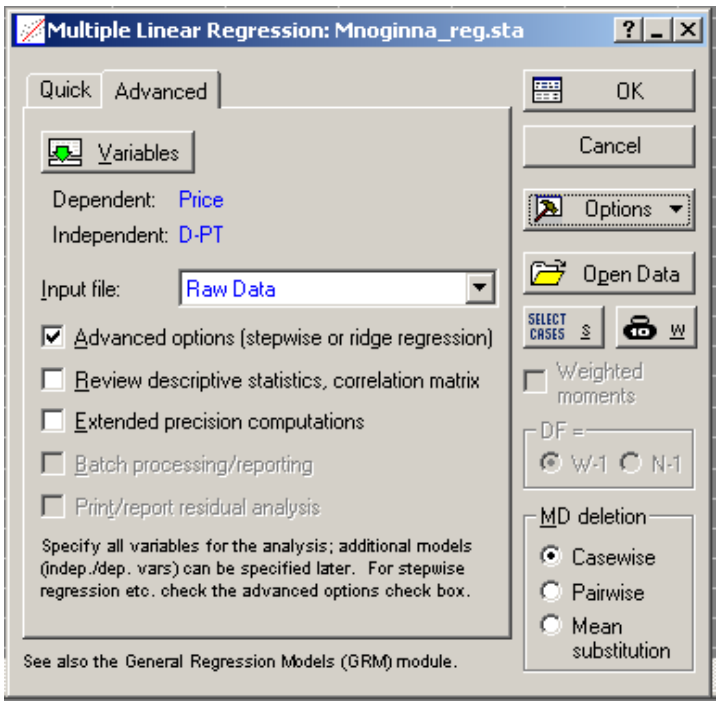




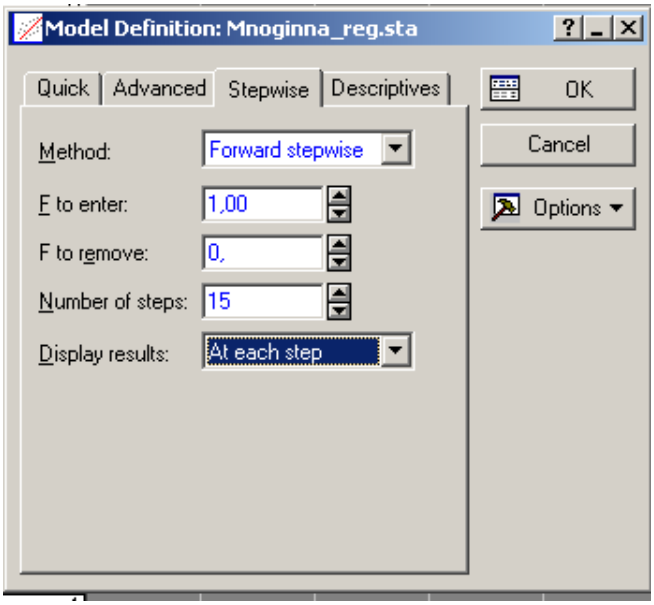
На закладці Quick натиснемо кнопку Summary:Regression results. Отримаємо такі результати.

N=32	Regression Summary for Dependent Variable: Price (Spreadsheet3) R= ,91620052 R?= ,83942339 Adjusted R?= ,76295833 F(10,21)=10,978 p<,00000 Std.Error of estimate: 82,827 **Beta**	Std.Err. of Beta	B	Std.Err. of B	t(21)	p-level
Intercept			-8134,88	2787,794	-2,91804	0,008222
D	0,689194	0,252202	115,48	42,260	2,73271	0,012470
T1	0,117429	0,215645	5,93	10,887	0,54454	0,591803
T2	0,279290	0,137071	4,57	2,243	2,03756	0,054390
S	0,469331	0,098439	0,42	0,088	4,76775	0,000104
PR	-0,224560	0,112857	-81,12	40,769	-1,98978	0,059794
NE	0,355454	0,100056	137,45	38,690	3,55256	0,001883
CT	0,126927	0,100626	43,27	34,307	1,26137	0,221008
BW	-0,019204	0,120943	-8,24	51,884	-0,15879	0,875354
N	-0,260021	0,142196	-6,99	3,822	-1,82861	0,081698
PT	-0,044867	0,148421	-19,25	63,672	-0,30229	0,765401

Як ми бачимо, незначущими є змінні T1, T2, PR, CT, BW, N, PT. Коефіцієнт детермінації $R^2 = 0{,}8394$. Претендентами на пояснювальні змінні є D,S, NE. Але ми будемо включати або виключати змінні за доп. «Регресії вперед» та «Регресії назад». На вкладці Advanced відмічаємо Advanced options (stepwise or ridge regression), OK.



Далі на вкладці Stepwise відмічаємо Method: Forward stepwise; F to enter 1; F to remove 0. Display results: At each step. Це означає, що ми запускаємо процедуру поступового включення змінних: спершу, перебираючи всі регресори, будуємо модель простої регресії і перевіряємо: яке значення F-критерію Фішера при перевірці гіпотези про незначущість змінної є найбільшим. Таку змінну ми залучаємо до множини регресорів.



Натискаємо Summary: Regression results. (Правда, перший раз може з'явитись попередження, що немає жодної змінної – це нормально, клацаємо на кнопку Next і дивимось на Summary Results.)

Regression Summary for Dependent Variable: Price (Mnoginna_reg.sta)
R= ,61045425 R?= ,37265439 Adjusted R?= ,35174287
F(1,30)=17,821 p<,00021 Std.Error of estimate: 136,97

N=32	Beta	Std.Err. of Beta	B	Std.Err. of B	t(30)	p-level
Intercept			-6553,57	1661,963	-3,94327	0,000446
D	0,610454	0,144608	102,29	24,231	4,22144	0,000207

Знов робимо Next і дивимось на Summary Results, до змінної D додалась S.

Regression Summary for Dependent Variable: Price (Mnoginna_reg.sta)
R= ,76428521 R?= ,58413188 Adjusted R?= ,55545132
F(2,29)=20,367 p<,00000 Std.Error of estimate: 113,43

N=32	Beta	Std.Err. of Beta	B	Std.Err. of B	t(29)	p-level
Intercept			-6790,88	1377,668	-4,92926	0,000031
D	0,601426	0,119774	100,78	20,070	5,02134	0,000024
S	0,459955	0,119774	0,41	0,108	3,84019	0,000616

Regression Summary for Dependent Variable: Price (Mnoginna_reg.sta)
R= ,83128928 R?= ,69104187 Adjusted R?= ,65793921
F(3,28)=20,876 p<,00000 Std.Error of estimate: 99,497

N=32	Beta	Std.Err. of Beta	B	Std.Err. of B	t(28)	p-level
Intercept			-4553,31	1406,112	-3,23823	0,003091
D	0,408949	0,121910	68,52	20,428	3,35450	0,002296
S	0,466464	0,105085	0,42	0,094	4,43892	0,000128
PT	-0,379408	0,121890	-162,76	52,290	-3,11271	0,004243

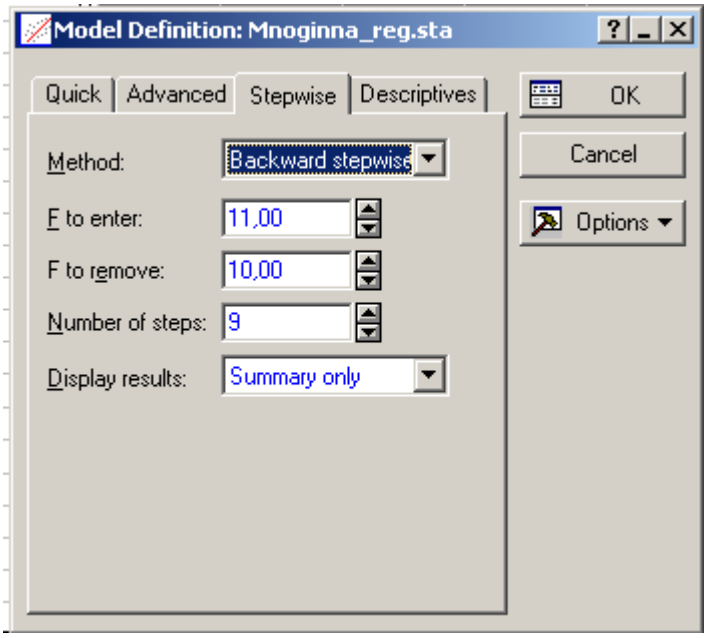
Здійснюємо ще кілька кроків по залученню змінних, отримуємо результат, де в наявності вже є незначуща змінна.

Regression Summary for Dependent Variable: Price (Mnoginna_reg.sta)
R= ,89398530 R?= ,79920972 Adjusted R?= ,75102005
F(6,25)=16,585 p<,00000 Std.Error of estimate: 84,887

N=32	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			-6458,73	1405,975	-4,59378	0,000107
D	0,552783	0,117831	92,63	19,744	4,69133	0,000083
S	0,409879	0,095905	0,37	0,086	4,27378	0,000245
PT	-0,237753	0,111514	-101,99	47,839	-2,13204	0,043012
NE	0,288955	0,094598	111,74	36,580	3,05455	0,005294
T2	0,270656	0,119052	4,43	1,948	2,27343	0,031851
PR	-0,187177	0,104217	-67,62	37,648	-1,79602	0,084587

Отже, з 11 регресорів ми обрали 6. Тепер будемо поступово виключати з цього набору покроково ту змінну, яка має найменше значення F-критерію Фішера.

Задаємо нову регресійну модель. Відмічаємо змінні – регресорами вже будуть змінні D, S, PT, NE, T2, PR. На вкладці Stepwise - відмічаємо Method: Backward stepwise; F to enter 11; F to remove 10. Display results: At each step. OK.



Натискаємо на Next і на Summary: Regression Results; повторюємо стільки разів, скільки треба, слідкуючи за коефіцієнтом детермінації – щоб він залишався достатньо великим. Зупиняємось на такому оптимальному варіанті.

Regression Summary for Dependent Variable: Price (Mnoginna_reg.sta)
R= ,86710337 R?= ,75186825 Adjusted R?= ,71510799
F(4,27)=20,453 p<,00000 Std.Error of estimate: 90,802

N=32	Beta	Std.Err. of Beta	B	Std.Err. of B	t(27)	p-level
Intercept			-4756,21	1285,663	-3,69943	0,000975
D	0,423838	0,111408	71,02	18,668	3,80438	0,000741
S	0,467296	0,095903	0,42	0,086	4,87260	0,000043
NE	0,257051	0,099915	99,40	38,636	2,57269	0,015908
PT	-0,300573	0,115382	-128,94	49,498	-2,60501	0,014761

Будуємо ще одну регресійну модель – стандартну, регресорами якої будуть D, S, NE, PT. На вкладці Stepwise - відмічаємо Method: Standard.

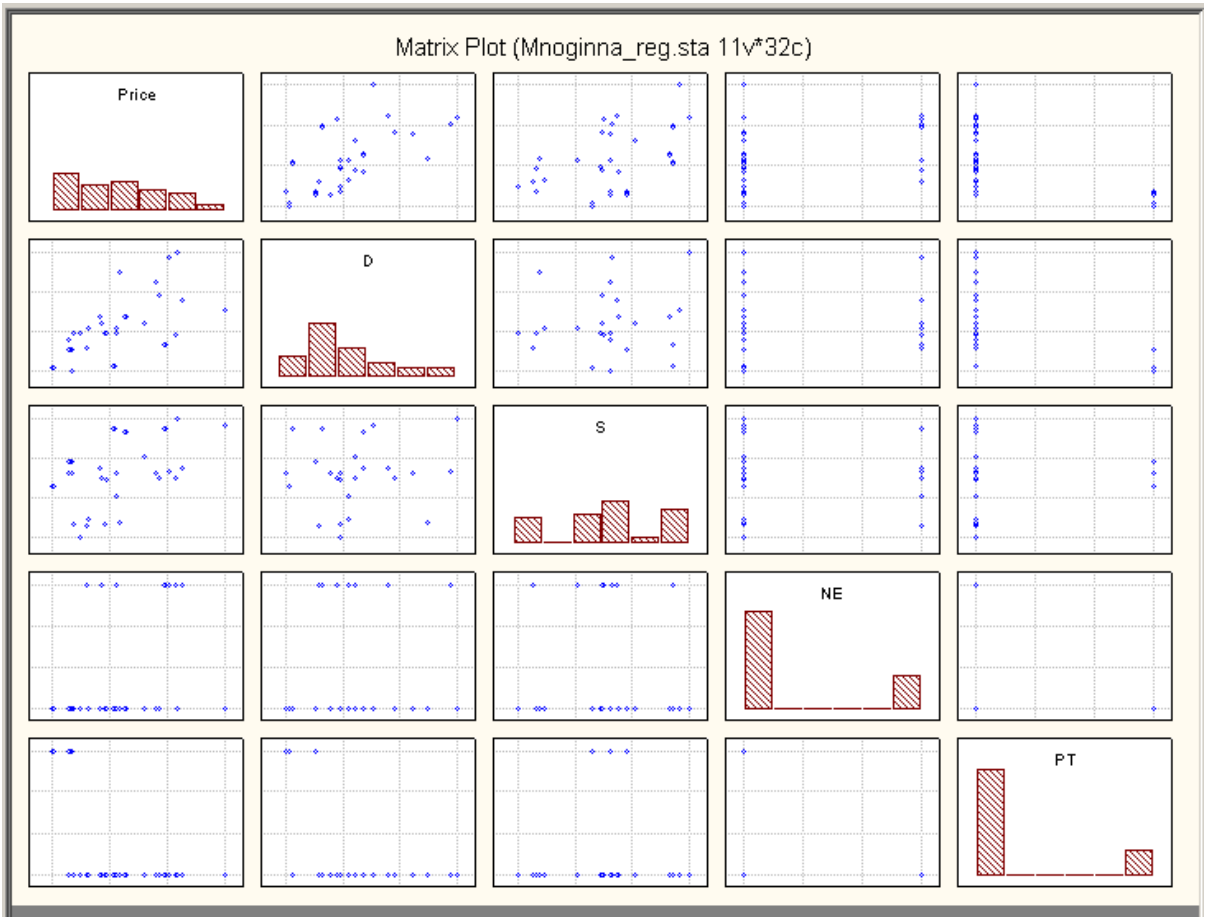
Оцінки будуть такі самі, як і в попередній таблиці. Маємо таку регресійну модель:

$$\Pr ice = -4756{,}21 + 71{,}02D + 0{,}42S + 99{,}4NE - 128{,}94PT + \varepsilon .$$

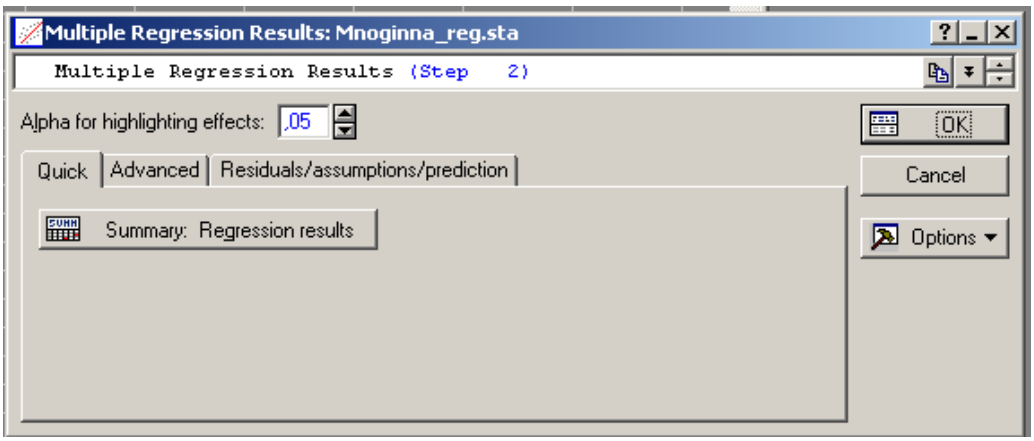
Бачимо, що оцінки коефіцієнтів мають правильні знаки – ціна тим вища, чим більший відрізок часу відведений на будівництво, більша номінальна потужність станції і чим менший нагляд над будівництвом.

Тепер побудуємо матричну діаграму розсіювання даних. Оберемо на верхній панелі Graphs->Matrix Plots. Вибираємо змінні, для яких ми хочемо побудувати діаграму розсіювання, обираємо Graph type: Square Scatter Matrix. Отримуємо попарні діаграми розсіювання, на діагоналі знаходяться гістограми.

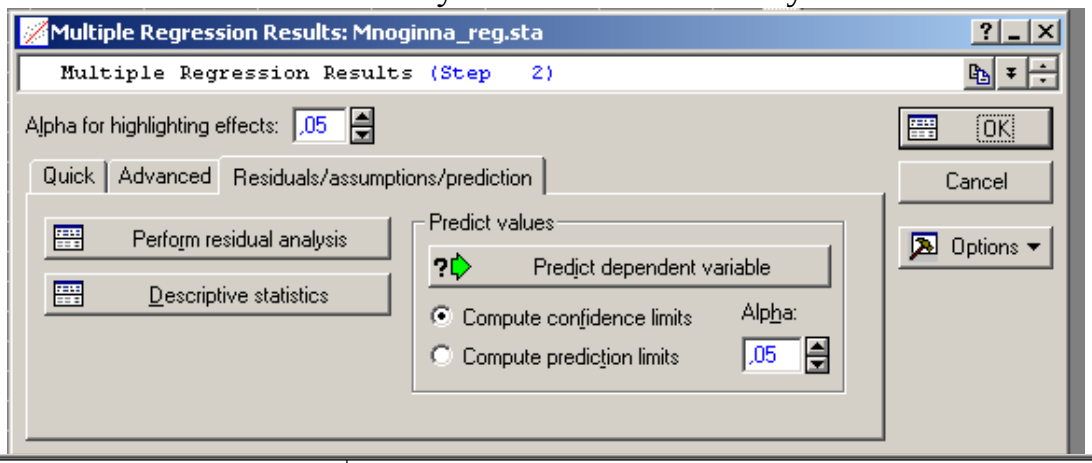
Між регресорами D та S не спостерігається мультиколінеарність, для моделі це добре. Видно, що Price залежить більш-менш лінійно від D та S, в усякому разі не спостерігається яскраво вираженої нелінійної залежності між ними.



2) Тепер проаналізуємо нормальність залишків моделі. В результатах переходимо на закладку Residuals/assumptions/prediction



Потім переходимо на закладку Residuals. Натискаємо на Casewise plot of residuals. Натискаємо на кнопку Perform residual analysis.

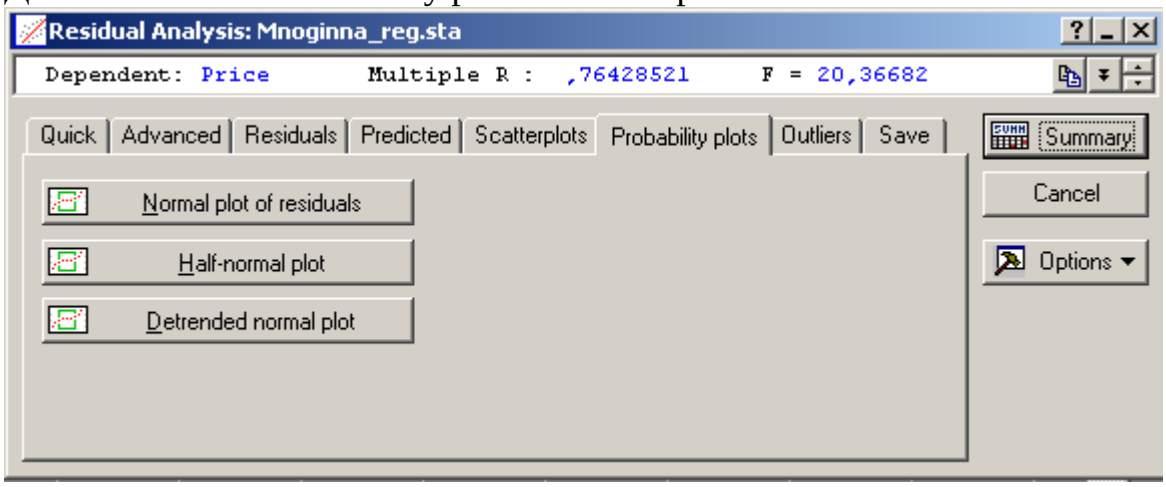


Raw Residual (Mnoginna_reg.sta)
Dependent variable: Price

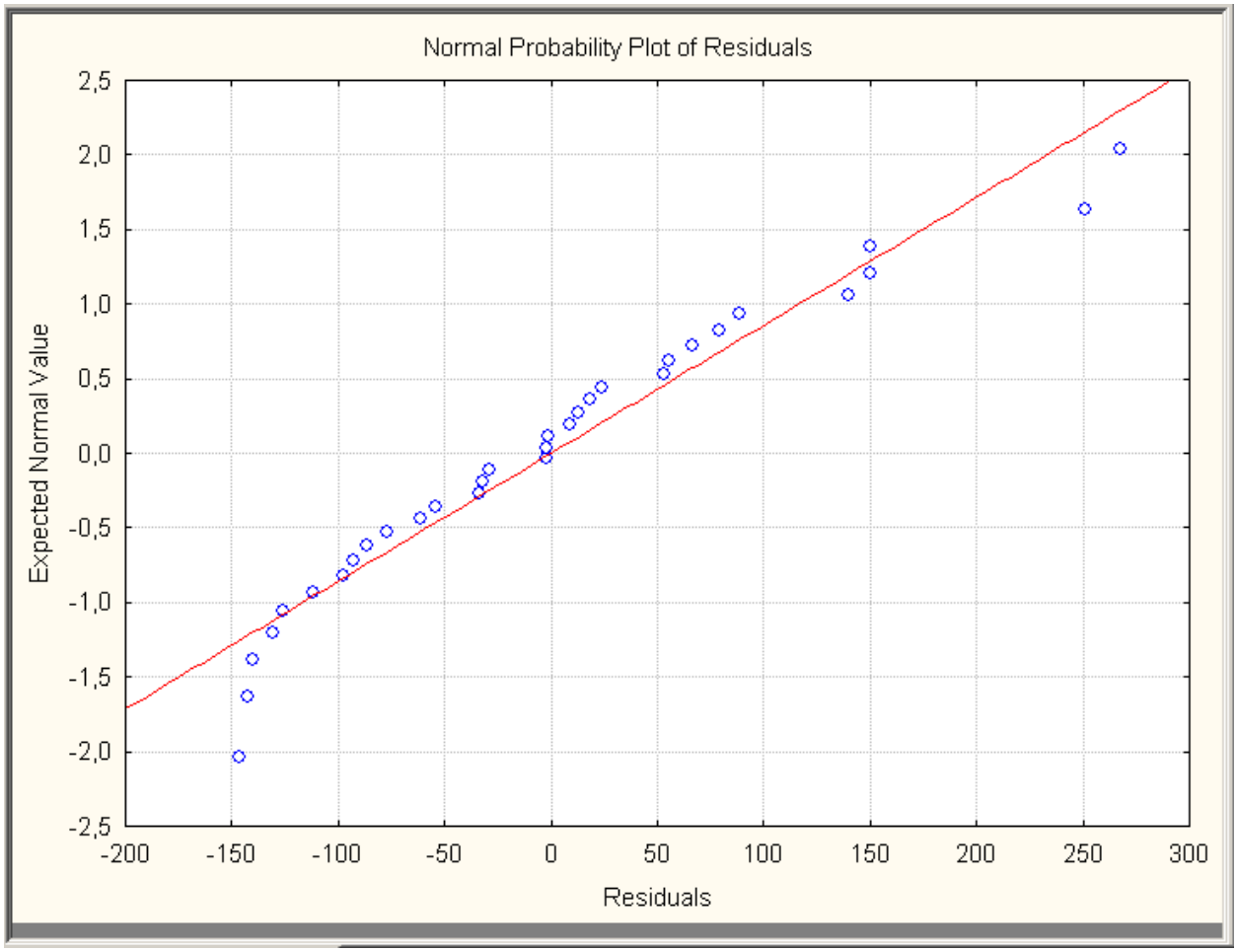
Case	Raw Residuals -3s … 0 … +3s	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1		460,0500	502,1050	-42,055	0,27486	-0,46315	34,30041	3,454747	-49,055	0,008329
2		452,9900	472,6240	-19,634	0,07500	-0,21623	41,58390	5,532811	-24,845	0,003140
3		443,2200	472,6240	-29,404	0,07500	-0,32382	41,58390	5,532811	-37,207	0,007043
4		652,3200	619,6057	32,714	1,07141	0,36028	40,94060	5,333210	41,062	0,008314
5		642,2300	619,6057	22,624	1,07141	0,24916	40,94060	5,333210	28,397	0,003976
6		345,3900	382,6034	-37,213	-0,53526	-0,40983	44,25525	6,394963	-48,807	0,013726
7		272,3700	430,2640	-157,894	-0,21216	-1,73887	24,95579	1,372835	-170,795	0,053448
8		317,2100	294,7845	22,425	-1,13059	0,24697	38,96643	4,740098	27,487	0,003375
9		457,1200	448,0188	9,101	-0,09180	0,10023	22,90860	1,004420	9,720	0,000146
10		690,1900	528,4313	161,759	0,45332	1,78144	33,13948	3,160381	186,616	0,112519
11		350,6300	349,3891	1,241	-0,76042	0,01367	31,64548	2,796473	1,412	0,000006
12		402,5900	557,4197	-154,830	0,64984	-1,70513	32,23106	2,937107	-177,150	0,095912
13		412,1800	325,4314	86,749	-0,92283	0,95536	34,07975	3,398017	100,972	0,034836
14		495,5800	579,2474	-83,667	0,79781	-0,92142	28,85684	2,162118	-93,067	0,021219
15		394,3600	495,2834	-100,924	0,22861	-1,11146	21,53539	0,774954	-106,939	0,015603
16		423,3200	429,5468	-6,227	-0,21702	-0,06857	23,23208	1,060537	-6,663	0,000070
17		712,2700	633,7743	78,496	1,16746	0,86447	35,05372	3,651182	92,243	0,030759
18		289,6600	325,4314	-35,771	-0,92283	-0,39395	34,07975	3,398017	-41,636	0,005924
19		881,2400	613,7950	267,445	1,03201	2,94535	31,69687	2,808711	304,556	0,274163
20		490,8800	579,2474	-88,367	0,79781	-0,97318	28,85684	2,162118	-98,295	0,023670
21		567,7900	509,6588	58,131	0,32606	0,64019	22,80135	0,985987	62,043	0,005888
22		665,9900	727,4848	-61,495	1,80273	-0,67724	51,68077	9,073365	-90,960	0,065014
23		621,4500	521,6794	99,771	0,40755	1,09877	26,49967	1,671520	109,059	0,024572
24		608,8000	565,4911	43,309	0,70456	0,47696	31,29287	2,713032	49,146	0,006958
25		473,6400	470,8286	2,811	0,06283	0,03096	44,27605	6,401889	3,688	0,000078
26		697,1400	766,2347	-69,095	2,06542	-0,76094	53,00418	9,594254	-104,807	0,090791
27		207,5100	203,6563	3,854	-1,74836	0,04244	38,08170	4,483803	4,676	0,000093
28		288,4800	229,8811	58,599	-1,57058	0,64535	37,66525	4,365200	70,777	0,020908
29		284,8800	304,0421	-19,162	-1,06783	-0,21103	37,80953	4,406142	-23,181	0,002260
30		280,3600	304,0421	-23,682	-1,06783	-0,26081	37,80953	4,406142	-28,649	0,003452
31		217,3800	203,6563	13,724	-1,74836	0,15114	38,08170	4,483803	16,653	0,001183
32		270,7100	304,0421	-33,332	-1,06783	-0,36708	37,80953	4,406142	-40,324	0,006839
Minimum		207,5100	203,6563	-157,894	-1,74836	-1,73887	21,53539	0,774954	-177,150	0,000006
Maximum		881,2400	766,2347	267,445	2,06542	2,94535	53,00418	9,594254	304,556	0,274163
Mean		461,5603	461,5603	-0,000	0,00000	-0,00000	35,05172	3,875000	-1,059	0,029507
Median		448,1050	472,6240	-2,493	0,07500	-0,02745	34,67706	3,552964	-2,625	0,007679

Отримуємо табличку залишків. Перевіряємо, чи виходять залишки за межі інтервалу $(-3s, 3s)$, де s - емпіричне середньоквадратичне відхилення залишків (на графіку залишок позначений *). Бачимо, що не виходять. Значення залишків знаходяться в третьому стовпчику таблиці. Середнє і медіана залишків=0.

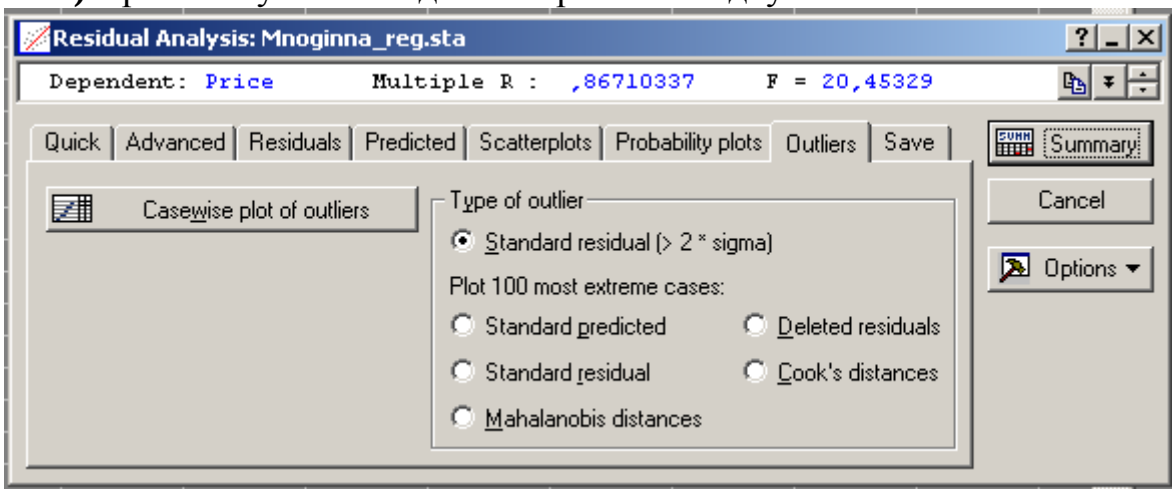
Далі натискаємо Probability plot ->Normal plot of residuals.



І отримуємо такий графік. Це є P-P діаграма порівняння залишків моделі з нормальним розподілом. Як ми бачимо, цілком вірогідно припустити, що залишки розподілені нормально – розкид не дуже великий.



3) Проаналізуємо викиди. Виберемо закладку Outliers.



Натиснемо Casewise plot outliers.

Standard Residual: Price (Mnoginna_reg.sta)
Outliers

Case	Standard Residuals −5.	−4.	−3.	±2.	3.	4.	5.	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val
19	.	.	.	.	*	.	.	881,2400	613,7950	267,4449	1,032014	2,945352	31,69687
Minimum	.	.	.	.	*	.	.	881,2400	613,7950	267,4449	1,032014	2,945352	31,69687
Maximum	.	.	.	.	*	.	.	881,2400	613,7950	267,4449	1,032014	2,945352	31,69687
Mean	.	.	.	.	*	.	.	881,2400	613,7950	267,4449	1,032014	2,945352	31,69687
Median	.	.	.	.	*	.	.	881,2400	613,7950	267,4449	1,032014	2,945352	31,69687

Як ми бачимо, 19-те спостереження ідентифікується як викид.

Якщо ми «відключимо» 19-те спостереження за допомогою SELECT CASES, то регресія теж зміниться і з'являться ще 2 викиди: спостереження 10 та 12.

Standard Residual: Price (Mnoginna_reg.sta)
Outliers

Case	Standard Residuals −5.	−4.	−3.	±2.	3.	4.	5.	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Er Pred.V
10	.	.	.	.*	.	.	.	690,1900	532,1943	157,996	0,608413	2,14422	26,909
12	.	.	.	*	.	.	.	402,5900	559,8158	-157,226	0,808068	-2,13377	26,162
Minimum	.	.	.	*	.	.	.	402,5900	532,1943	-157,226	0,608413	-2,13377	26,162
Maximum	.	.	.	.*	.	.	.	690,1900	559,8158	157,996	0,808068	2,14422	26,909
Mean	.	.	.	.*	.	.	.	546,3900	546,0050	0,385	0,708240	0,00522	26,535
Median	.	.	.	.*	.	.	.	546,3900	546,0050	0,385	0,708240	0,00522	26,535

Чи враховувати ці 3 спостереження в моделі чи ні? Однозначної відповіді немає. Можливо, замовник моделі, фахівець, який займався збором інформації щодо характеристик електростанцій, має свою думку щодо цього. Припустимо, він вважає, що 19 спостереження треба вилучити, а 10 та 12 залишити.

Regression Summary for Dependent Variable: Price (Mnoginna_reg.sta)
R= ,89591618 R?= ,80266581 Adjusted R?= ,77230670
F(4,26)=26,439 p<,00000 Std.Error of estimate: 73,685

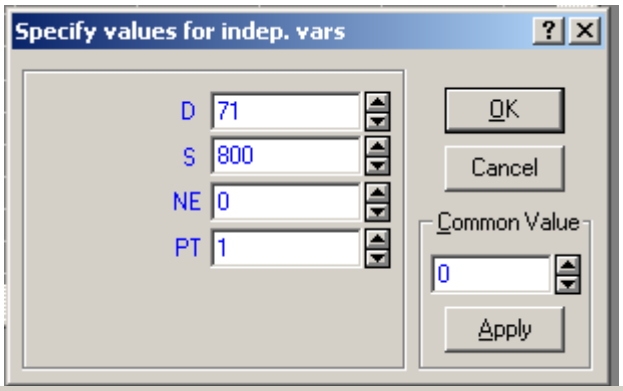
N=31	Beta	Std.Err. of Beta	B	Std.Err. of B	t(26)	p-level
Intercept			-4465,92	1045,983	-4,26959	0,000231
D	0,448076	0,100865	67,42	15,177	4,44233	0,000146
S	0,419162	0,087205	0,35	0,072	4,80663	0,000056
NE	0,333493	0,091128	115,78	31,636	3,65959	0,001128
PT	-0,303284	0,104793	-116,61	40,293	-2,89412	0,007598

Тоді остаточно маємо модель:

$$\Pr ice = -4465{,}92 + 67{,}42D + 0{,}35S + 115{,}78NE - 116{,}61PT + \varepsilon$$

Зауважимо, що при бажанні вилучити з нашої моделі вільний член, можна поступити наступним чином. На вкладці Advanced відмічаємо Advanced options (stepwise or ridge regression) (ми це робили раніше), натискаємо OK. У віконці Intercept вибираємо Set to zero. Далі працюємо за планом, вільний член буде відсутній.

4) Нам залишається прорахувати прогноз для заданих значень регресорів, а також знайти інтервал надійності. Переходимо у закладку Residual/assumptions/prediction і натиснемо Predict dependent variable. У відповідні віконця вставляємо значення змінних D, S, NE, PT. Натискаємо OK.



Predicting Values for (Mnoginna_reg.sta) variable: Price

Variable	B-Weight	Value	B-Weight * Value
D	67,421	71,0000	4786,93
S	0,348	800,0000	278,21
NE	115,777	0,0000	0,00
PT	-116,613	1,0000	-116,61
Intercept			-4465,92
Predicted			482,60
-95,0%CL			357,74
+95,0%CL			607,47

Отже, передбачуване значення для Price = 482,6. З ймовірністю 0,95 справжнє значення Price повинно потрапити в інтервал (357,74; 607,74).

5. Нелінійна регресія

Припустимо, що з діаграми розсіювання даних видно, або це випливає з теорії – регресійна модель нелінійна:

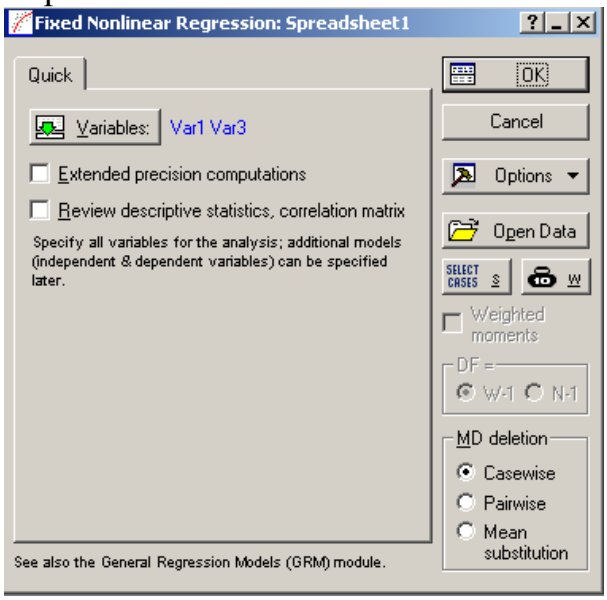
$$Y = g(X, b) + \varepsilon .$$

Як підганяти такі дані? Розглянемо на прикладі.

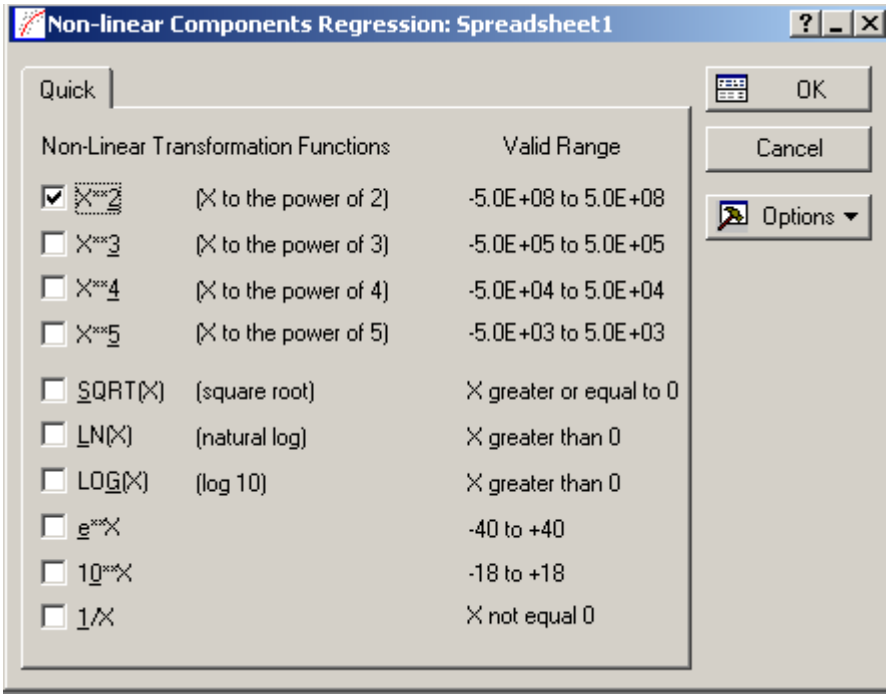
Приклад 5.1. Першу змінну заповнимо числами від 1 до 10, другу – випадковими значеннями від 0 до 1 (це в нас випадкова похибка ε, вводимо її суто для моделювання), третя змінна

$$v3 = 2 * v1 * v1 + v2 + 20 .$$

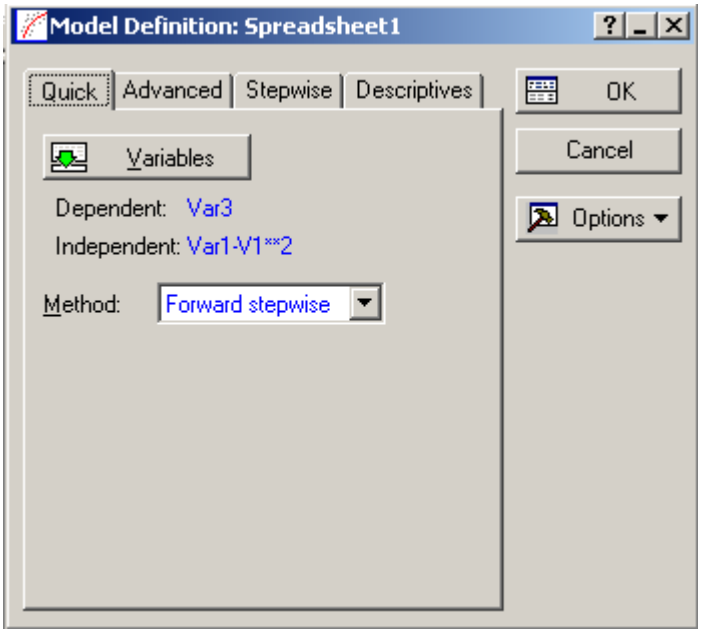
Потрібно отримати рівняння регресії для Var3 через Var1. Зрозуміло, що доцільно ввести в регресійну модель квадрат Var1. Переходимо Statistics->Advanced Linear/Nonlinear Models->Fixed nonlinear Regression. У вікні, що з'явилося, обираємо змінні Var1 та Var3.

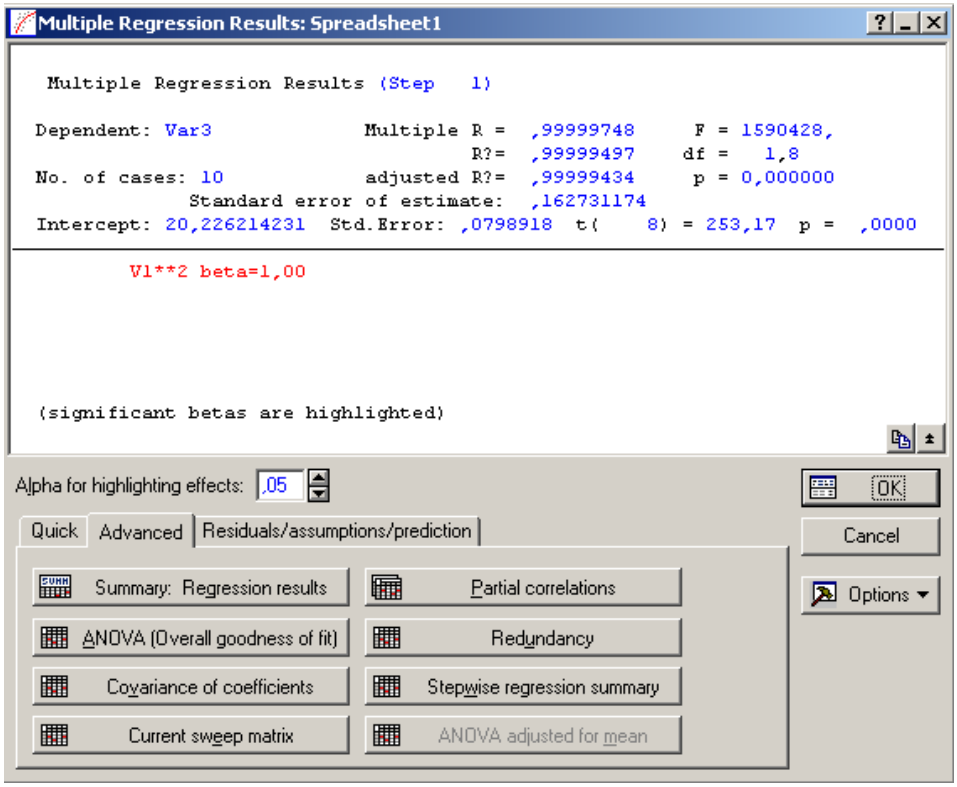


Натискаємо ОК. Оскільки ми знаємо, що залежність квадратична, обираємо X**2.



З'явиться вікно Model Definition. Перейдемо на закладку Quick і вкажемо залежну змінну Var3, незалежні змінні - Var1 та V1**2. Будуємо регресійну модель покроково (Forward stepwise) з відкиданням незначущих змінних. В закладці Stepwise вказуємо значення F to enter - 1.0, F to remove – 0.01.





Натискаємо на Summary: Regression results.

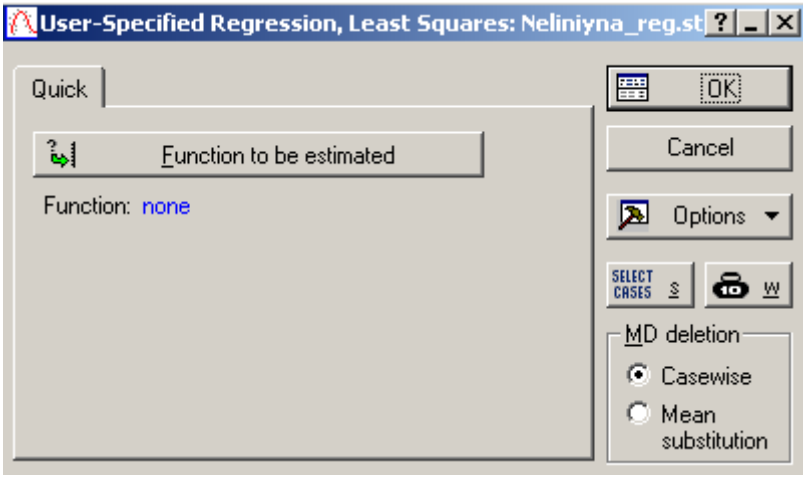
Regression Summary for Dependent Variable: Var3 (Spreadshe
R= ,99999748 R?= ,99999497 Adjusted R?= ,99999434
F(1,8)=1590E3 p<0,0000 Std.Error of estimate: ,16273

N=10	Beta	Std.Err. of Beta	B	Std.Err. of B	t(8)	p-level
Intercept			20,22621	0,079892	253,170	0,000000
V1**2	0,999997	0,000793	2,00178	0,001587	1261,122	0,000000

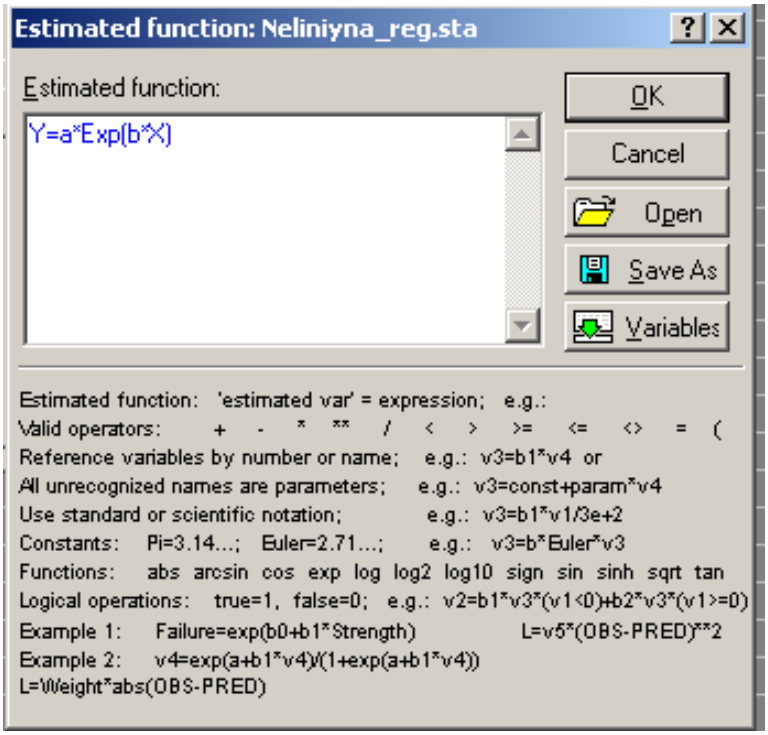
Отже, маємо той самий результат, на який сподівались.

Приклад 5.2. Створимо файл з наступними змінними: X містить числа від 1 до 10, Var2 задається як e^X, Var3 заповнимо випадковими значеннями від 0 до 1; Y задамо як Var2+Var3. Отже, ми задали штучно, що $Y \approx e^X$. Чи «розпізнає» таку залежність наш пакет, якщо ми задамо йому задачу підігнати дані у вигляді $Y = ae^{bX} + \varepsilon$, a,b обрати, виходячи з даних?

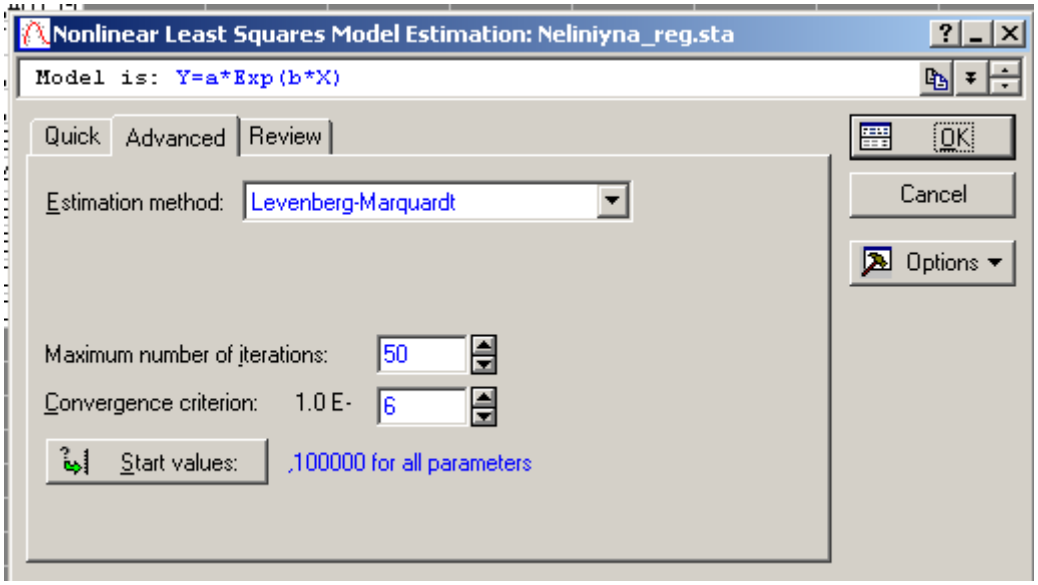
Переходимо Statistics->Advanced Linear/Nonlinear Models->Nonlinear Estimation. Далі обираємо User-specified regression, least squares. Натискаємо ОК.



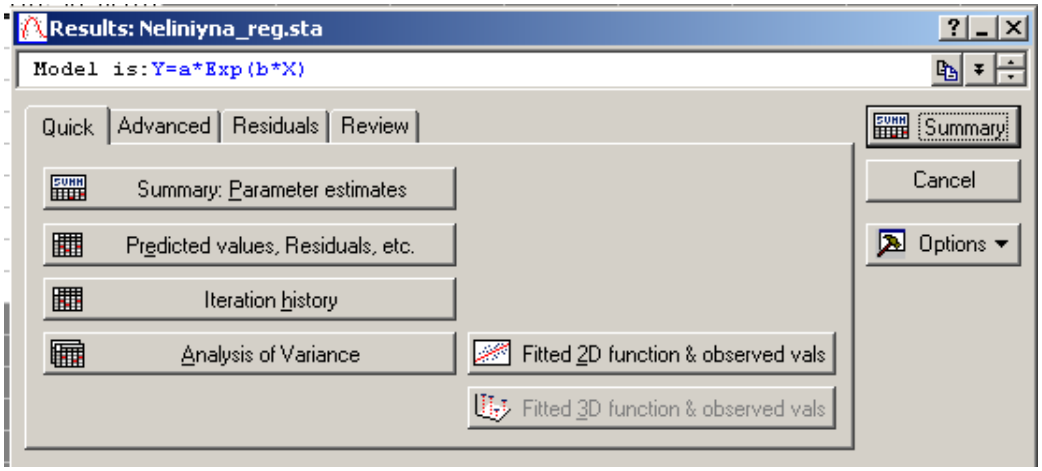
Натискаємо на Function to be estimated.



У віконці Estimated function набираємо Y=a*Exp(b*X), натискаємо ОК. І ще раз ОК.

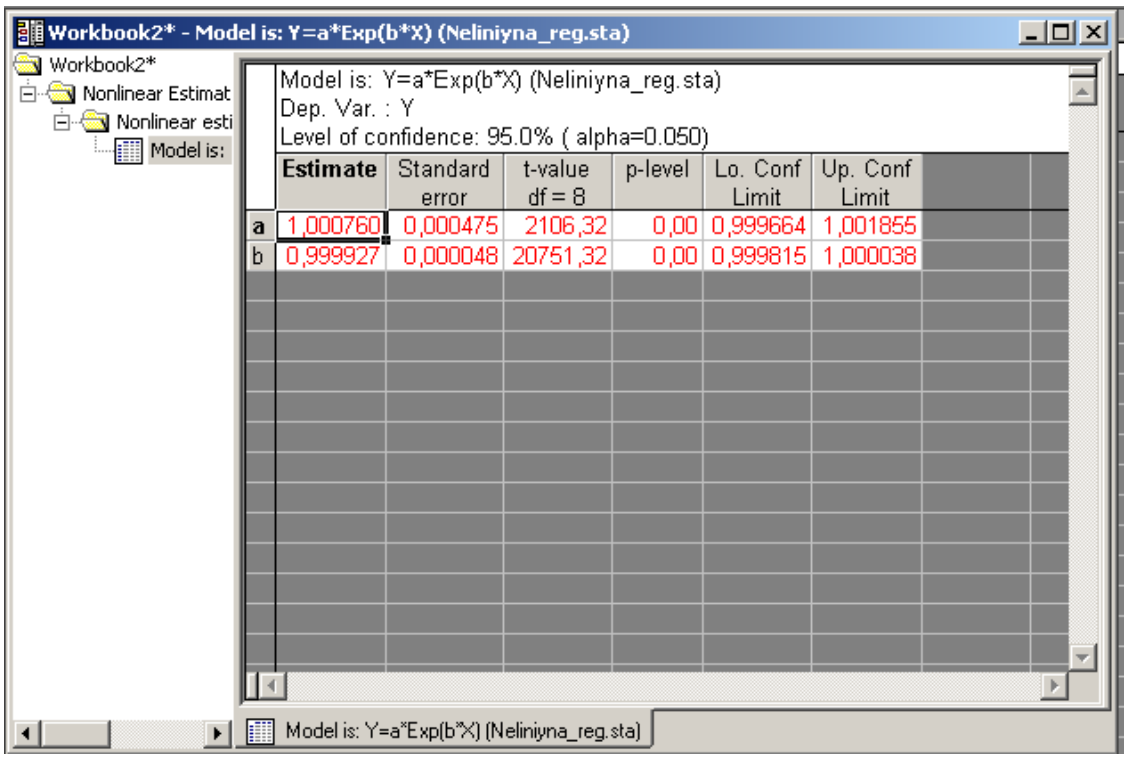


Нам пропонують, застосувавши МНК і отримавши відповідні трансцендентні рівняння, розв'язувати їх наближено за допомогою методу Левенберга-Марквардта, що насправді являє собою метод Ньютона. Оцінка похибки - 10^{-6}, Максимальна кількість ітерацій – 50. Початкові значення для а та в можна задати, клацнувши на кнопку Start Value.

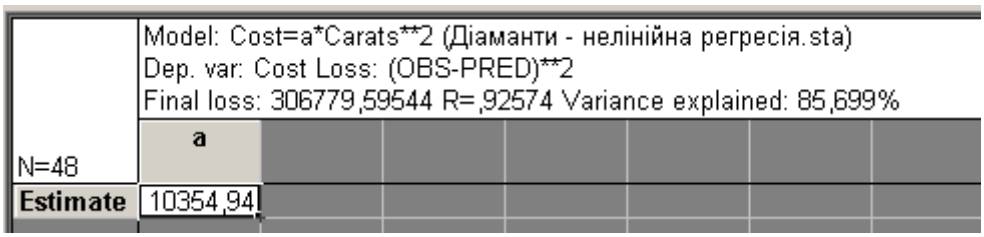


Потім натискаємо на Summary:Parameter estimates і маємо наступний результат.

Як ми бачимо, значення а та в дуже близькі до 1, як ми їх і задавали.



Зауваження 1. Можна за Estimation method взяти не Levenberq-Marquardt, а Rosenbrock and quasi-Newton, за loss function – (OBS-PRED)**2. В цьому випадку результатом підгонки будуть не тільки оцінені значення параметрів, а і відсоток поясненої дисперсії Variance explained – аналог коефіцієнта детермінації. Наприклад,



Відсоток поясненої дисперсії для цього прикладу становить 85,69 %. Він обчислюється для нелінійної моделі $Y_i = g(\beta, X_i) + \varepsilon_i$ за формулою

$$R^2 \times 100\%; \quad R^2 = 1 - \frac{ESS}{TSS} = 1 - \frac{\sum_{i=1}^{n} \left(Y_i - g\left(\hat{\beta}, X_i\right)\right)^2}{\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2}.$$

Зауваження 2. Для побудови відповідного графіку необхідно перейти на закладку Residuals, натиснути на кнопку Save predicted and residual values. Треба відмітити змінні, разом з якими ми хочемо зберегти передбачені моделлю та спостережувані значення. Пакет створить окремий файл з відповідними змінними.

Потім переходимо по ланцюжку Graphs->Scatterplots->Multiply. Відмічаємо змінні X як X, Y як Y та PREDICT. І графік побудований, нам заважають тільки 2 «підганяючі» прямі. Відмічаємо пряму чорним

кольором, клацаємо на праву кнопку миші і натискаємо на Properties. Натискаємо Delete.

Приклад 5.3. Відкриємо файл Program.sta. Маємо список службовців, які вивчали певний програмний продукт, а потім здавали тестування, на виконання завдання тесту був відведений певний час. EXPERNCE - витрачений працівниками час на вивчення програми, SUCCESS – результати тестування. Треба побудувати залежність успішності виконання від часу вивчення.

	1 EXPERNCE	2 SUCCESS
Frank	14	FAILURE
Henry	29	FAILURE
Tom	6	FAILURE
Beth	25	SUCCESS
Susan	18	SUCCESS
Harry	4	FAILURE
Paul	18	FAILURE
Pete	12	FAILURE
Diana	22	SUCCESS
Louise	6	FAILURE
Fred	30	SUCCESS
Hank	11	FAILURE
Steven	30	SUCCESS
Tod	5	FAILURE
Take	20	SUCCESS
Sam	13	FAILURE
Gail	9	FAILURE
Thomas	32	SUCCESS
Theodore	24	FAILURE
Charles	13	SUCCESS
Elizabeth	19	FAILURE
Lori	4	FAILURE
Ann	28	SUCCESS
Valerie	22	SUCCESS
Anke	8	SUCCESS

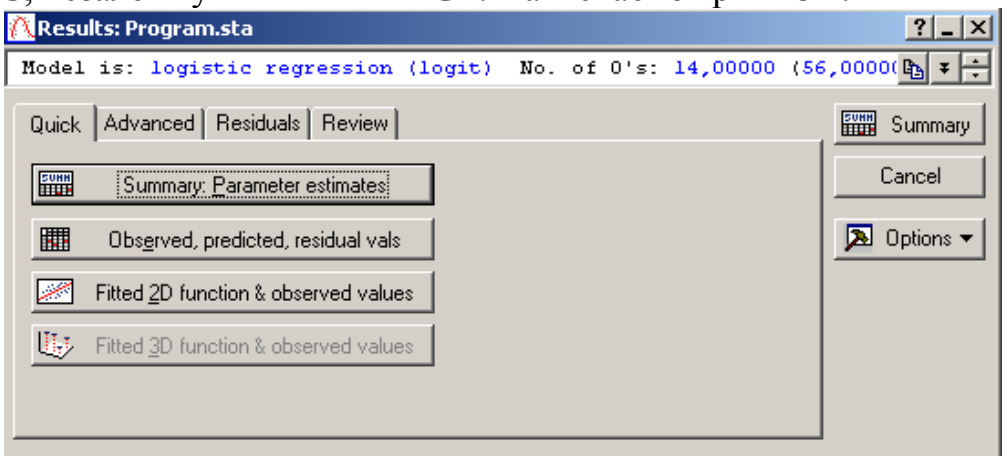
Будемо шукати залежність у вигляді логістичної регресійної моделі:

$$P\{Y = SUCCESS\} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)},$$

де Y – успішність, X – час навчання, b_0, b_1 – параметри, які необхідно оцінити. Тобто ми підганяємо змінну Y під біноміальний розподіл, ймовірність успіху в якому задається логістичною функцією, залежною від двох параметрів, а також від змінної – часу навчання.

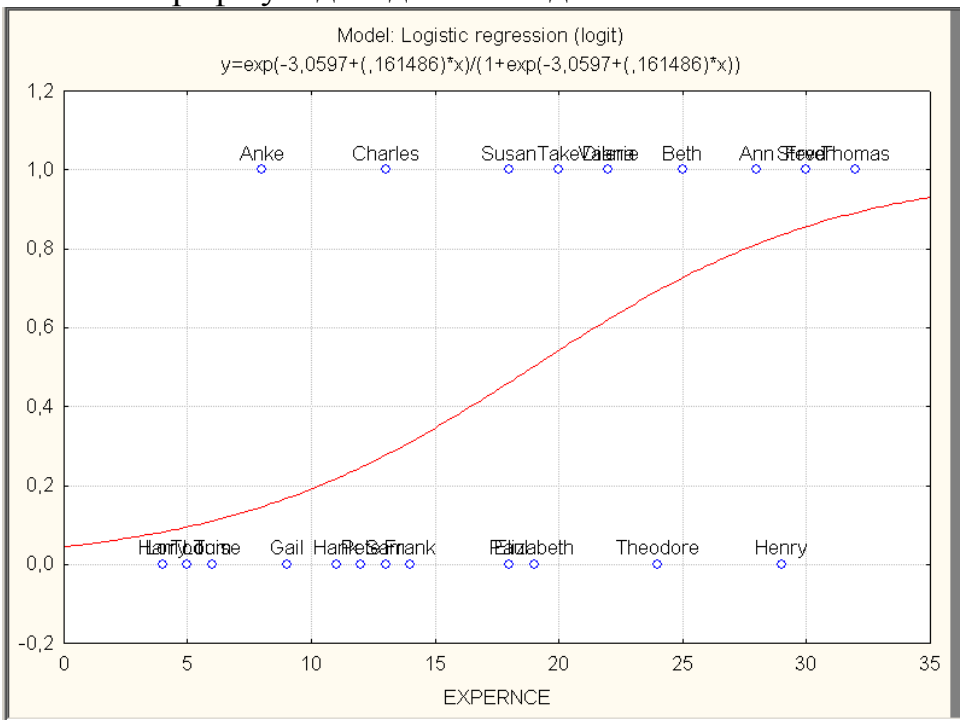
Переходимо Statistics->Advanced Linear/Nonlinear Models->Nonlinear Estimation.

Обираємо Quick Logit Regression->OK. Вкажемо як залежну змінну SUCCESS, незалежну - EXPERIENCE. Натискаємо тричі OK.

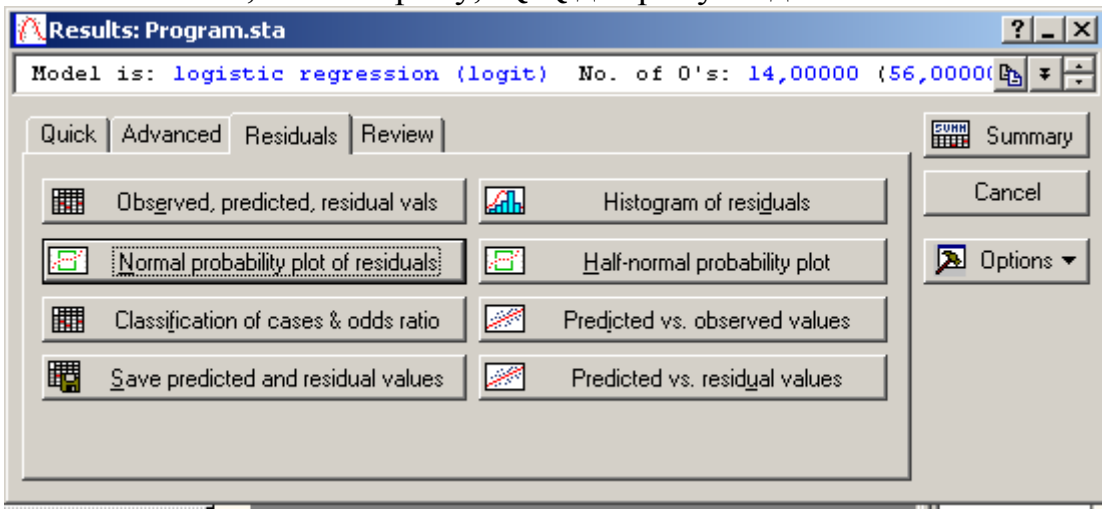


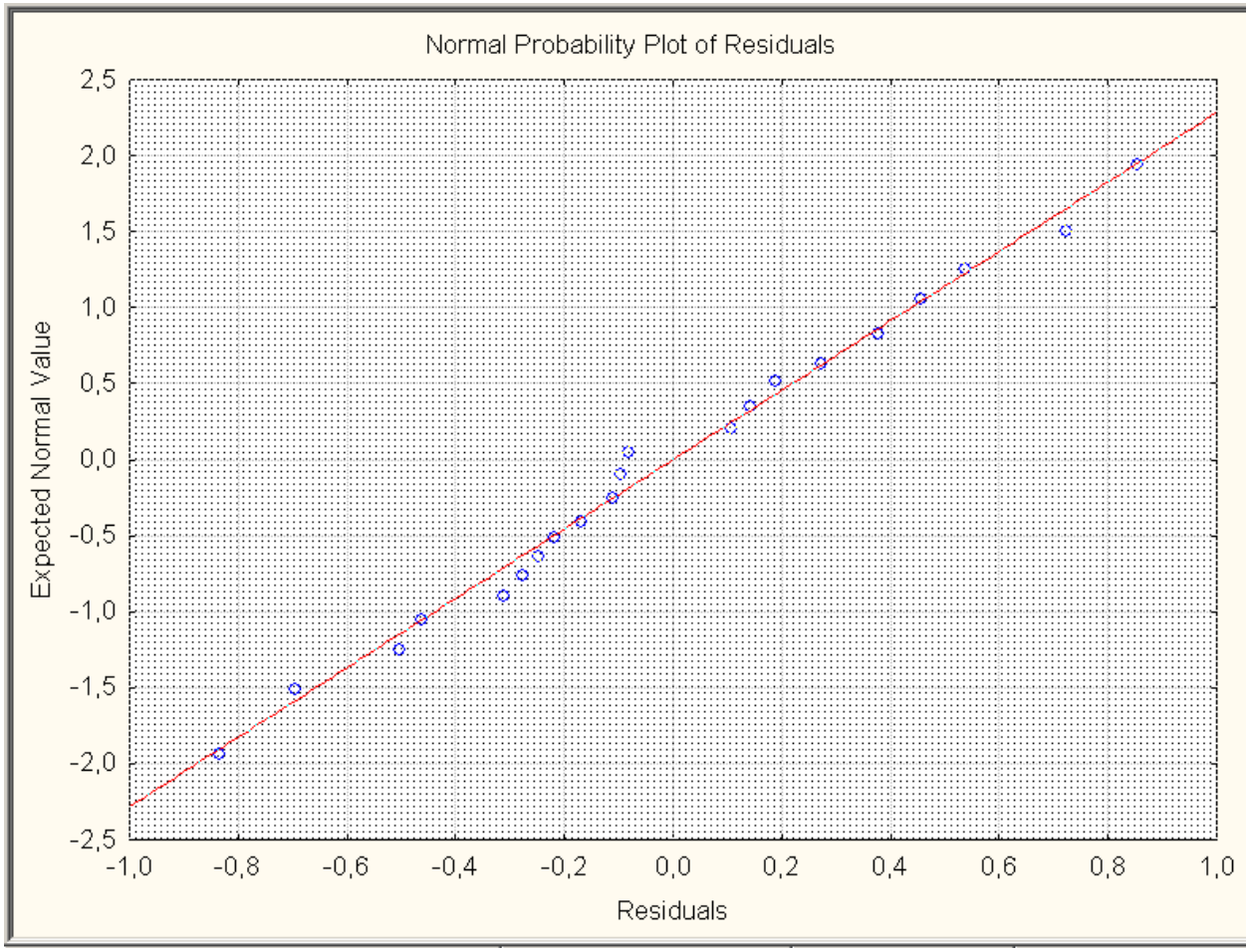
N=25	Model: Logistic regression (logit) N of 0's:14 1's:11 (Program.sta) Dep. var: SUCCESS Loss: Max likelihood Final loss: 12,712287040 Chi?(1)=8,8719 p=,00290 Const.B0	EXPERNCE
Estimate	-3,05970	0,16149
Odds ratio (unit ch)	0,04690	1,17526
Odds ratio (range)		91,98325

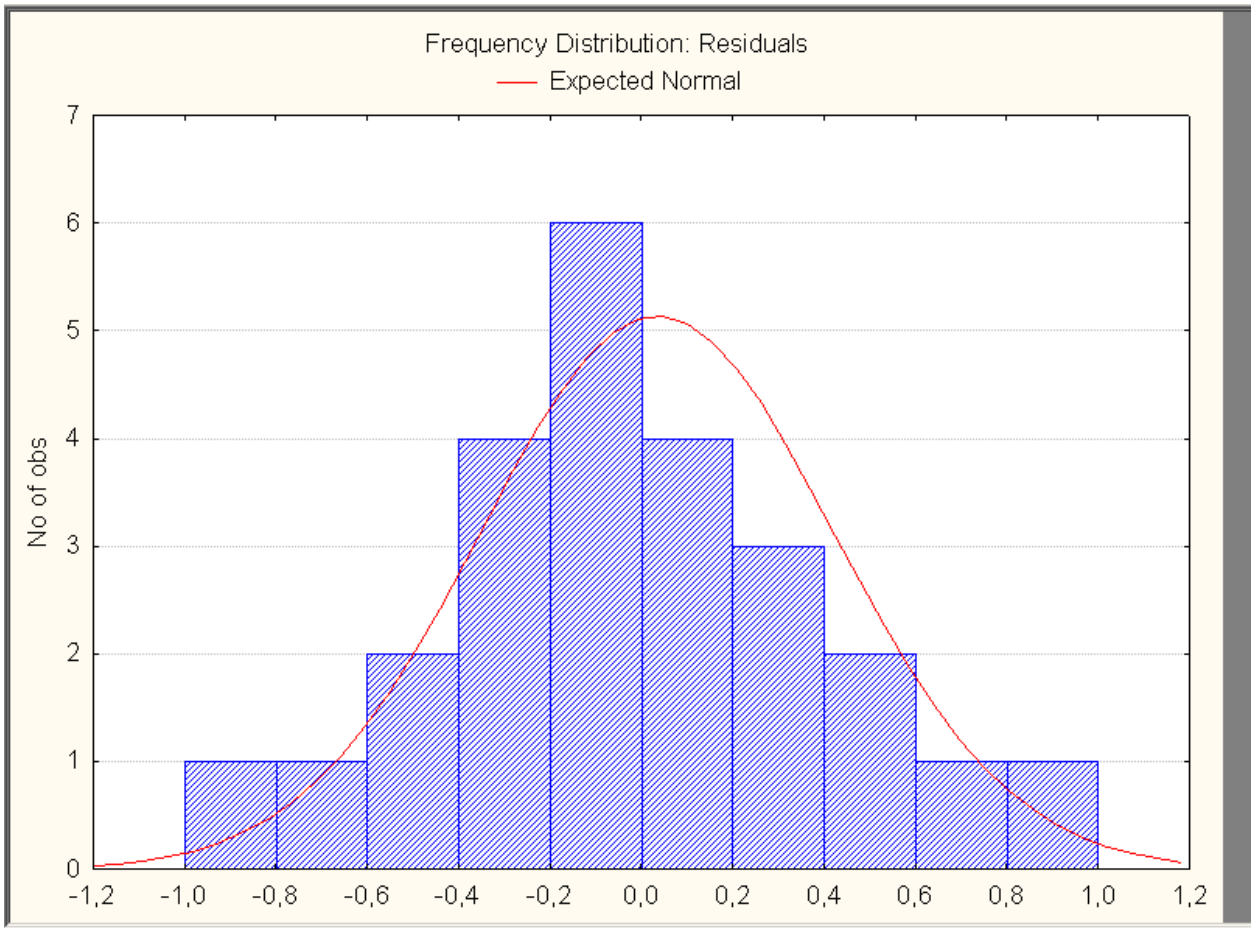
Отже, оцінка параметру b_0 - -3,0597, параметру b_1 – 0,16149. Сума квадратів залишків при підгонці – 12,71. В результаті перевірки за допомогою критерію максимальної правдоподібності досягнутий рівень значущості становить 0,0029 – змінна є значущою. Натиснувши Fitted 2D function & observed values, отримаємо графік логістичної регресійної функції, а також на графіку відкладені наші дані.



Натискаючи різні кнопки на закладці Residuals, можемо отримати значення залишків, їх гістограму, Q-Q діаграму і т.д.




Normal Probability Plot of Residuals
Expected Normal Value
2,5
2,0
1,5
1,0
0,5
0,0
-0,5
-1,0
-1,5
-2,0
-2,5
-1,0
-0,8
-0,6
-0,4
-0,2
0,0
0,2
0,4
0,6
0,8
1,0
Residuals


Frequency Distribution: Residuals
Expected Normal
No of obs
7
6
5
4
3
2
1
0
-1,2
-1,0
-0,8
-0,6
-0,4
-0,2
0,0
0,2
0,4
0,6
0,8
1,0
1,2

Приклад 5.4. **Поліноміальна регресія**

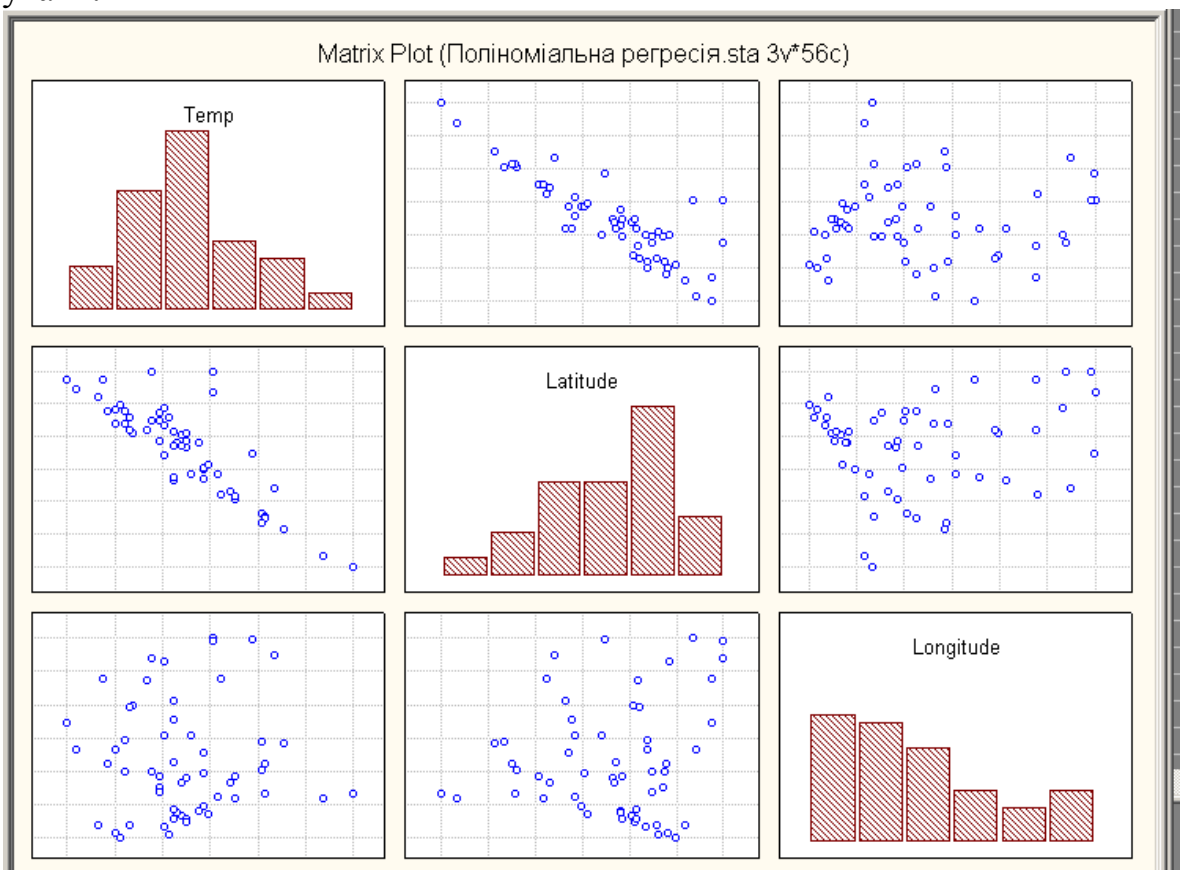
Наведені дані містять середню мінімальну температуру січня в градусах Фаренгейта разом з географічною довготою і широтою 56-ти міст Сполучених Штатів. (Мається на увазі, що для кожного року з 1931 по 1960 щоденні мінімальні температури січня були додані разом і поділені на 31. Потім ці середні для кожного року були усереднені по 30 роках.) Дані знаходяться у файлі Polinomialna_reg.sta.

Назви змінних:

1. City: Назва міста
2. Temp: Середня мінімальна температура січня в градусах Фаренгейта за 1931-1960 рр.
3. Lat: Географічна широта в градусах (на північ від екватора).
4. Long: Географічна довгота в градусах (на захід від нульового меридіана)

Завдання. Виявити і побудувати оптимальну модель залежності Temp від Lat та Long.

Перше, що треба зробити для вивчення зв’язків між змінними – **скористатися технікою візуального аналізу. Можна побудувати MatrixPlot**. Видно, що між Temp та Lat – лінійна залежність, між Temp та Long – нелінійна або взагалі відсутня. Регресори вважаємо незалежними, мультиколінеарності немає на графіку, також це зрозуміло з теоретичних міркувань.



Для дослідження на мультиколінеарність побудуємо вибіркову кореляційну матрицю. Здійснимо шлях Statistics->Basic Statistics/Tables->Correlation matrices. OK. Відмічаємо One variable list. Відмічаємо Latitude та Longitude. OK. Натискаємо на Summary.

Variable	Correlations (Поліноміальна регресія.sta) Marked correlations are significant at p < ,05000 N=56 (Casewise deletion of missing data) **Latitude**	Longitude
Latitude	1,00	0,14
Longitude	0,14	1,00

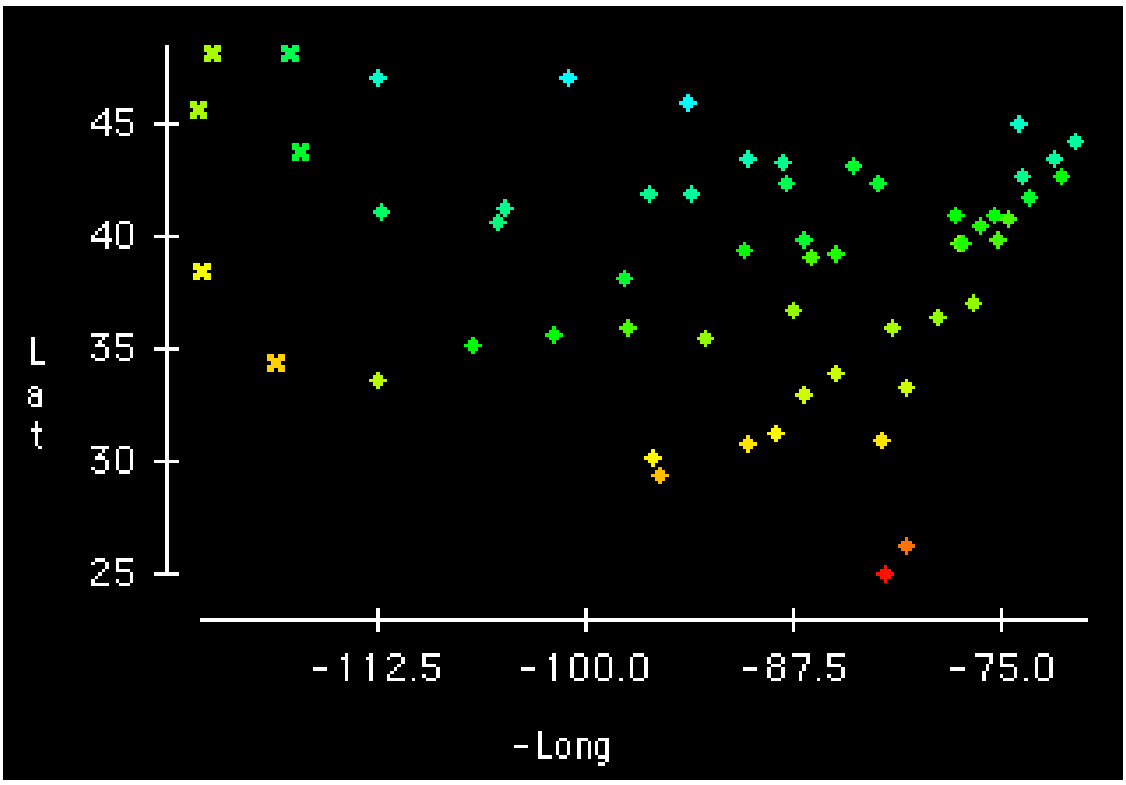
Як ми бачимо, вибірковий коефіцієнт кореляції між Latitude та Longitude дорівнює 0,14. Це також свідчить про відсутність мультиколінеарності.

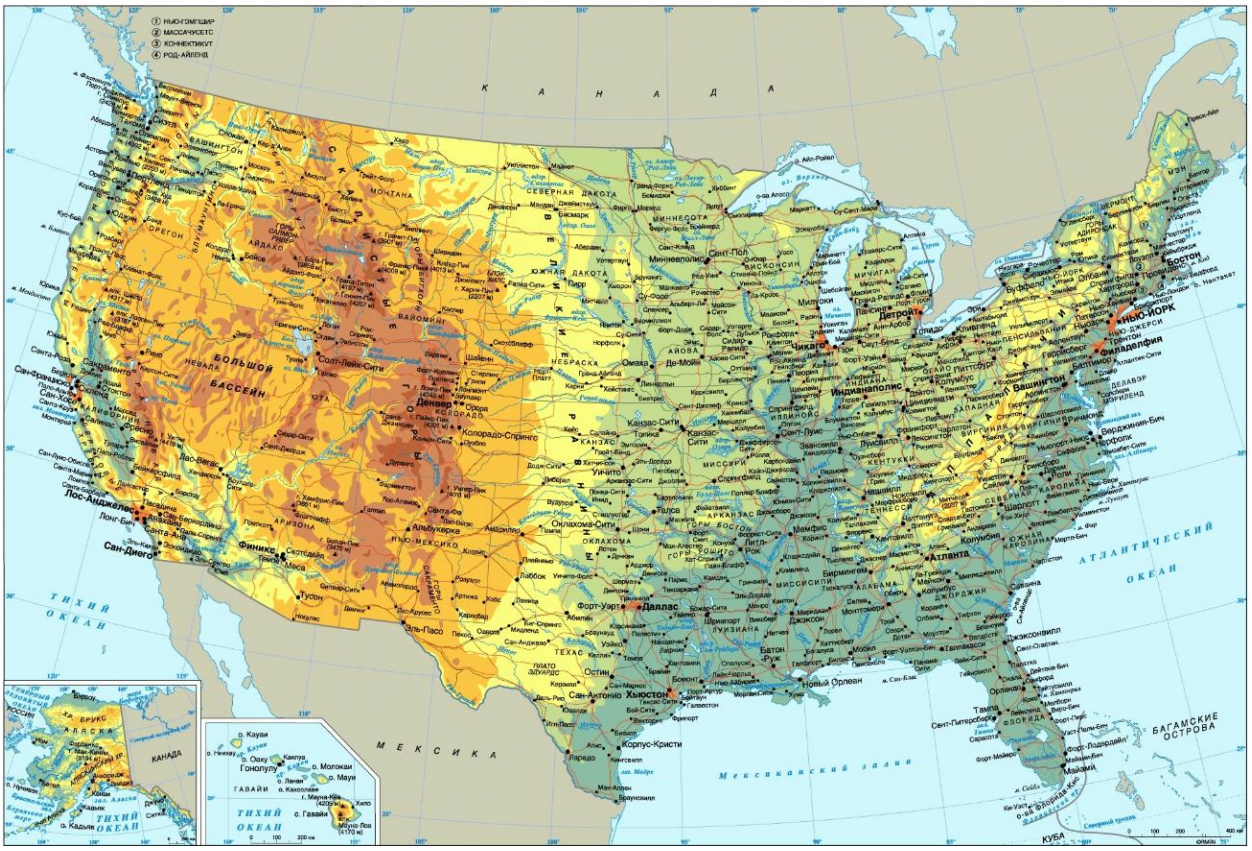
Рробимо лінійну «експрес-модель», попередньо розглядаючи змінні як потенційні регресори – чи будуть змінні значущими чи ні?

N=56	Regression Summary for Dependent Variable: Temp (Поліноміальна регресія.sta) R= ,86089982 R?= ,74114850 Adjusted R?= ,73138052 F(2,53)=75,875 p<,00000 Std.Error of estimate: 6,9345 **Beta**	Std.Err. of Beta	B	Std.Err. of B	t(53)	p-level
Intercept			98,64523	8,327083	11,8463	0,000000
Latitude	-0,869729	0,070630	-2,16355	0,175699	-12,3139	0,000000
Longitude	0,149851	0,070630	0,13396	0,063141	2,1216	0,038560

Як ми бачимо, модель досить непогана, обидві змінні значущі, але Latitude (широта) краще підходить на роль регресора, що входить до моделі лінійним чином. Коефіцієнт детермінації = 0,741.

Міркуємо: який вигляд залежності задати? Якщо ми побудуємо залежність широти від довготи, а яскравістю відмітимо теплішу або холоднішу температуру січня, то картина буде така:

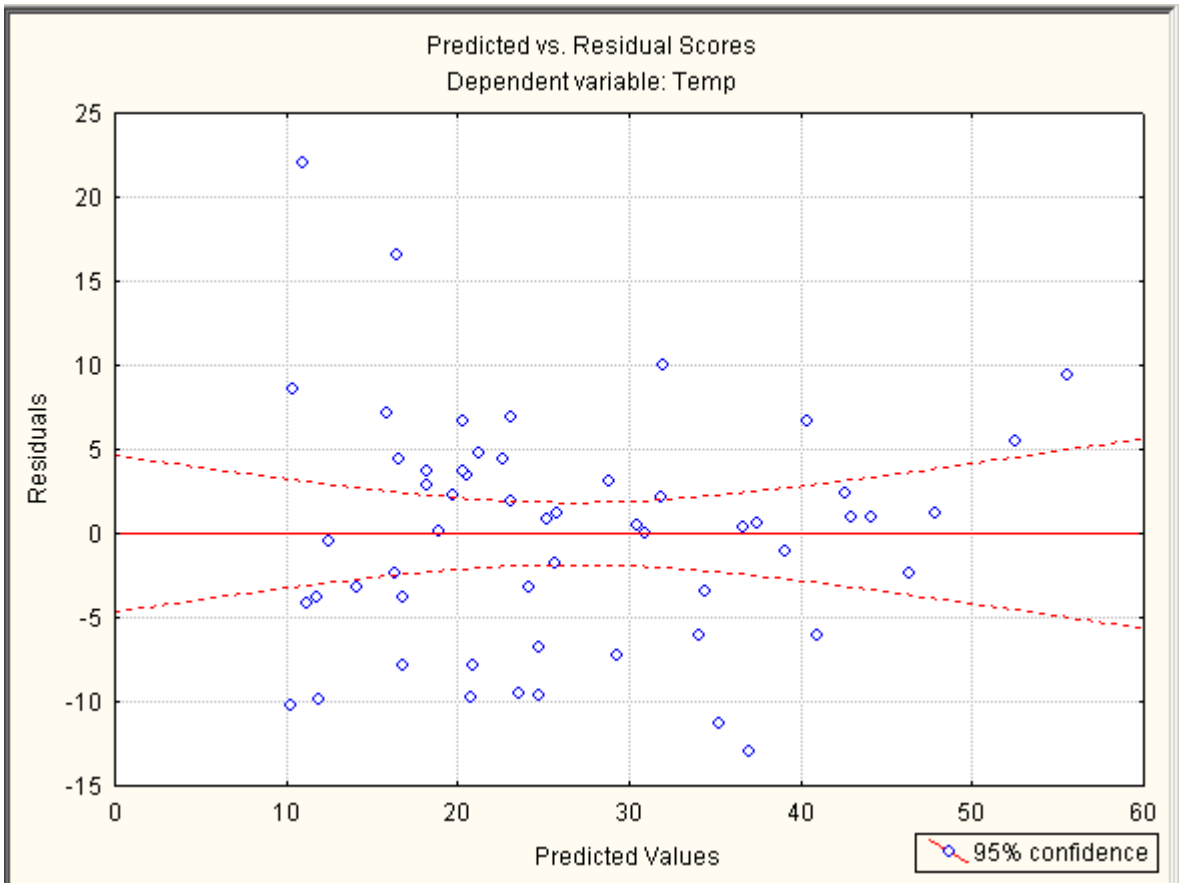




Щодо вигляду залежності маємо такі теоретичні міркування. Від широти середня температура січня повинна залежати лінійно – чим далі на північ, тим холодніше.

Щодо довготи – то інша справа. Америка з заходу омивається Тихим, а зі сходу - Атлантичним океаном, і через цей вплив мінімальна температура січня на узбережжі вища. Тому слід очікувати нелінійну залежність Temp від довготи. Як задати цю нелінійну залежність? На діаграмі розсіювання даних видно, що можна спробувати задати поліноміальну залежність зі степенем полінома не нижче, ніж 3.

Також для аналізу залишків моделі використовують графік Predicted versus residuals. На закладці Residual/assumptions/prediction натискаємо на кнопку Perform residual analysis, а потім на Predicted vs. residuals. Отримуємо графік, на якому відкладаються залишки та значення Temp, передбачені моделлю. Вважається, що якщо модель побудована правильно, то дана діаграма розсіювання не буде містити ніякої закономірності, точки утворюватимуть безформну хмару. Якщо точки розташовані вздовж якоїсь кривої, то не всі закономірності залучені до моделі, ми повинні спробувати підібрати залежність в іншій формі.



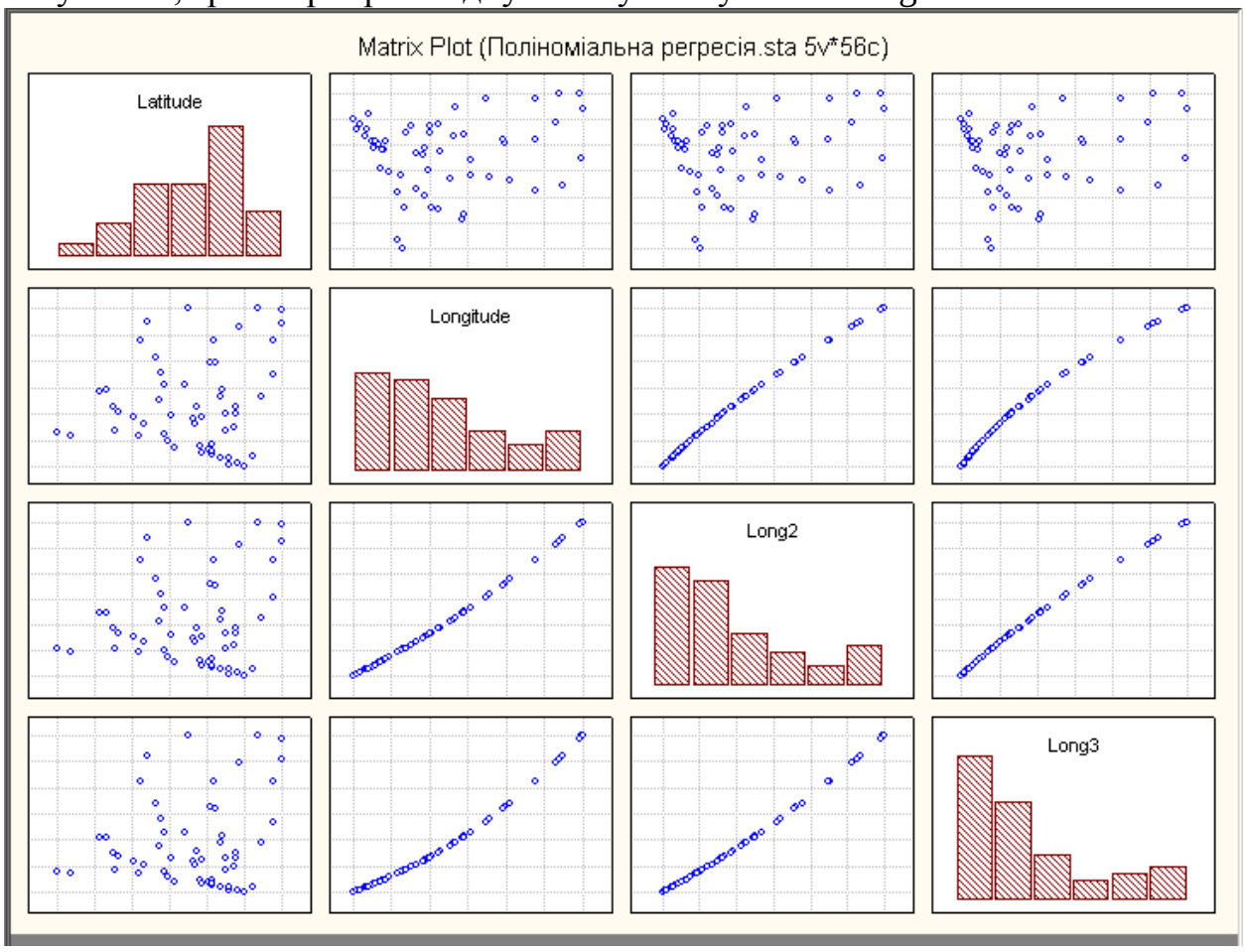
Здається, що залишки розташовані вздовж кривої.
Будуємо модель залежності у вигляді

$$Temp = b_1 + b_2 Lat + b_3 Long + b_4 (Long)^2 + b_5 (Long)^3 + \varepsilon .$$

Застосуємо метод лінеаризації. Введемо ще 2 змінні: $Long2 = (Long)^2$ та $Long3 = (Long)^3$. Побудуємо модель лінійної залежності Temp від Lat, Long, Long2, Long3.

Але пакет не хоче виконувати цю дію, скаржиться на наявність мультиколінеарності. Будуємо MatrixPlot для всіх регресорів, бачимо таку невтішну картину.

Безумовно, треба прибрати одну змінну. Вилучаємо Long3.



Regression Summary for Dependent Variable: Temp (Поліноміальна регресія.sta)
R= ,93404951 R?= ,87244849 Adjusted R?= ,86508975
F(3,52)=118,56 p<0,0000 Std.Error of estimate: 4,9144

N=56	Beta	Std.Err. of Beta	B	Std.Err. of B	t(52)	p-level
Intercept			323,9515	31,35546	10,3316	0,000000
Latitude	-1,01376	0,053786	-2,5218	0,13380	-18,8479	0,000000
Longitude	-4,92345	0,695229	-4,4014	0,62151	-7,0818	0,000000
Long2	5,10901	0,698305	0,0237	0,00324	7,3163	0,000000

Маємо модель, яка краща, ніж просто лінійна, всі змінні значущі, коефіцієнт детермінації підвищився до 0,872.

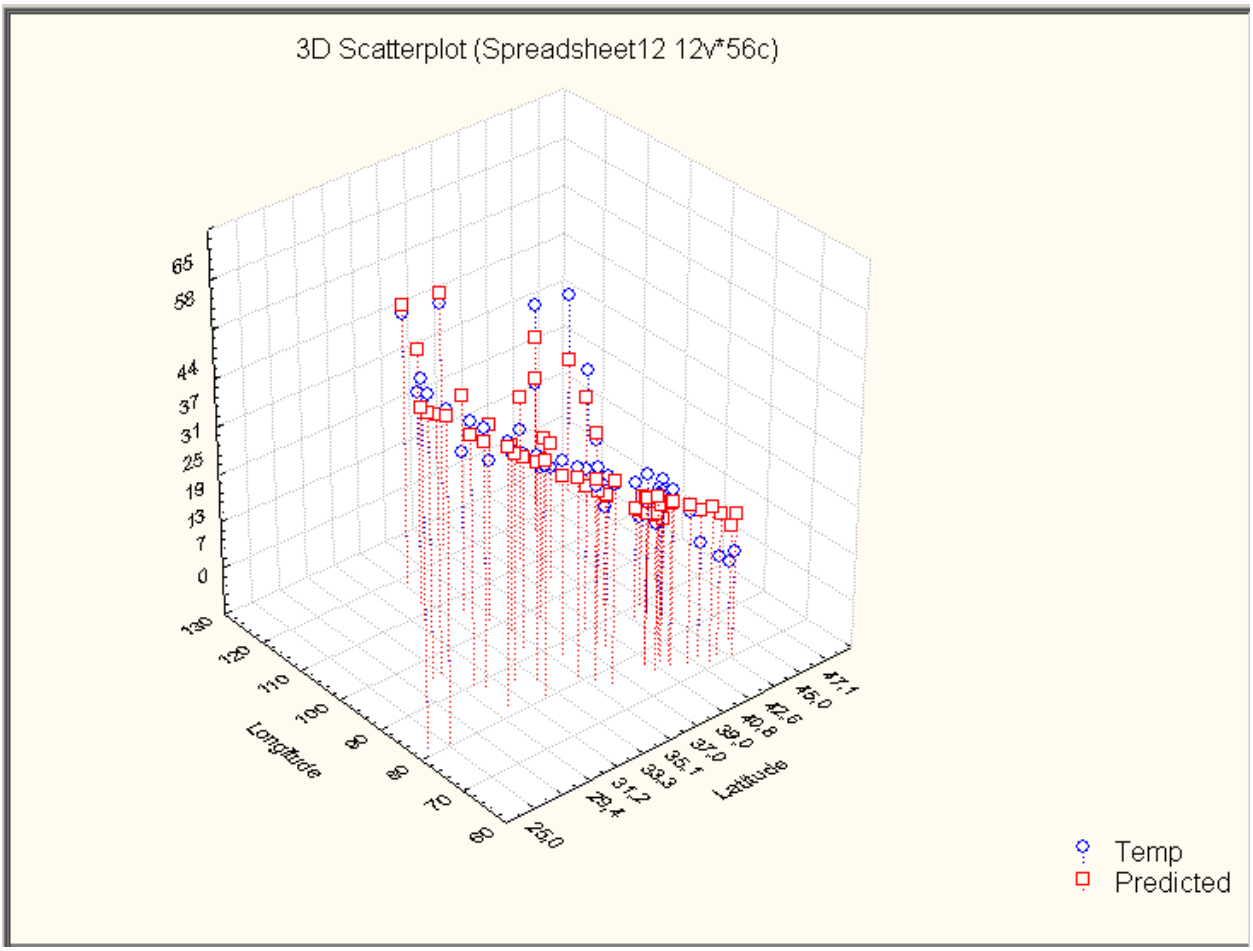
Випишемо модель – формулу для прогнозування.

$$Temp = 323{,}95 - 2{,}52Lat - 4{,}4Long + 0{,}024(Long)^2$$

Побудуємо графік. Перейдемо на закладку Residuals/assumptions/prediction->Perform residual analysis->Save->Save residuals & predicted. Відмітити змінні, які будуть збережені в окремому файлі разом зі змінною Predicted – змінні Temp, Longitude, Latitude. OK.

Потім переходимо за шляхом: Graphs->3D XYZ Graphs->Scatterplots->Scatterplot. OK. Відмічаємо змінні:

X:Latitude; Y:Longitude; Z:Temp, Predicted. Отримуємо графік.

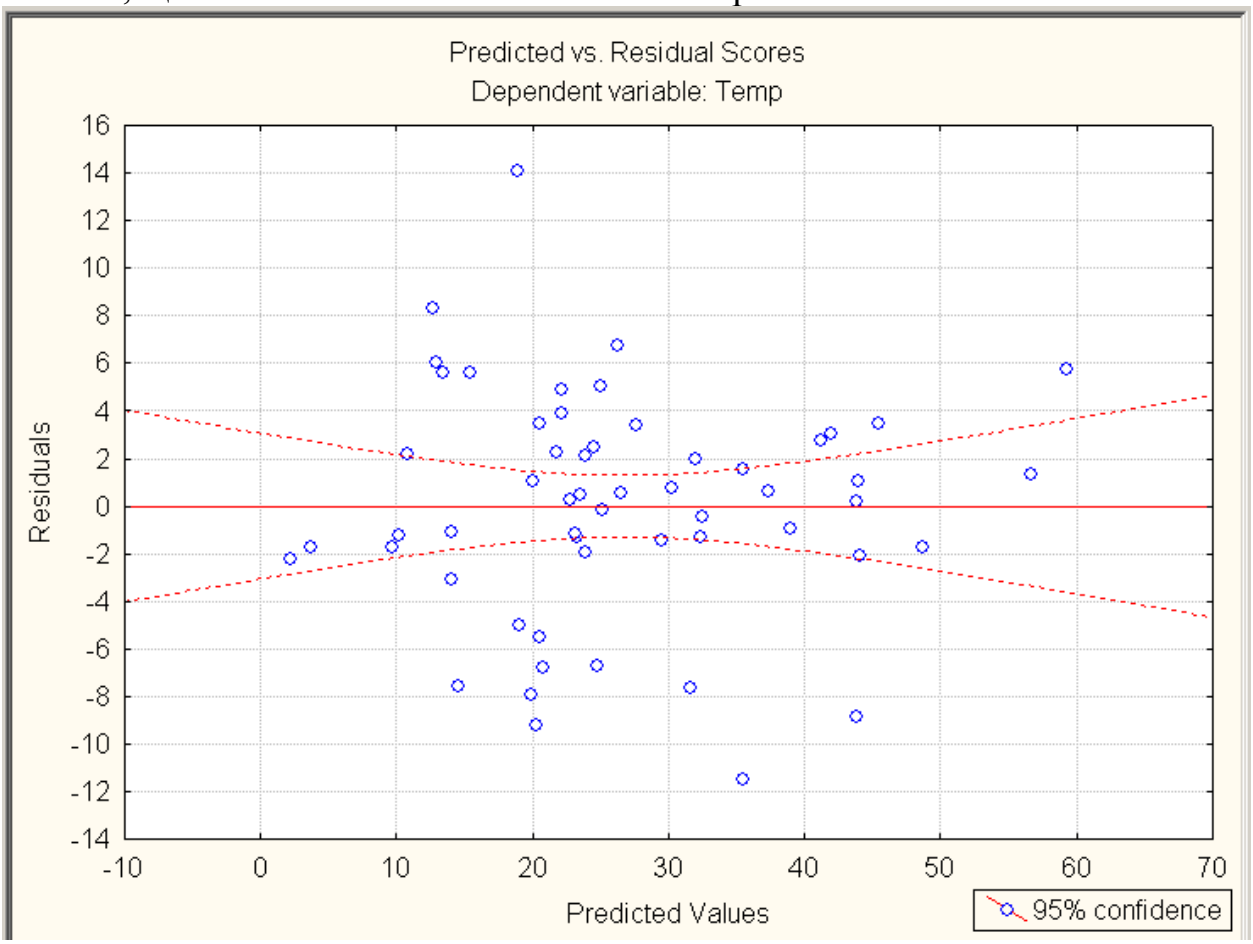


Схоже, що побудована модель непогана.

Якщо ми хочемо на діаграмі розсіювання подивитись – які саме спостереження спричинюють викид, то клацаємо на панелі меню на збільшувальне скло (Brushing). Справа з'являється панелька Brushing. В області Selection Brush відмічаємо Point. Далі клацаємо на діаграмі відповідну точку->На панельці натискаємо Update->На діаграмі відмічається номер (або назва) спостереження.

Також можна змінити точку, з якої ми дивимось на 3D-Scatterplot. Для цього клацаємо на головній панелі  (3D-Rotation control). З'являється окрема картинка діаграми. Пересуваючи квадратик знизу, зліва чи зверху, можемо обертати графік навкруги вісі OZ, площини OXY та дивитись на графік на більшій чи меншій відстані відповідно. Для ефекту обертання треба натиснути на Analytic exploratory spin options.

Ще раз будуємо графік Predicted vs. residuals для знайденої моделі. Бачимо, що в залишках немає ніякої закономірності.



Припустимо, що місто знаходиться на широті 20^0 і довготі 80^0. На закладці Residuals/assumptions/prediction натискаємо на Predict dependent variables. Задаємо Lat 20, Long 80, Long2 теж доводиться задати вручну 6400.

Variable	Predicting Values for (Поліноміальна регресія.sta) variable: Temp				
	B-Weight	Value	B-Weight * Value		
Latitude	-2,52184	20,000	-50,437		
Longitude	-4,40141	80,000	-352,113		
Long2	0,02369	6400,000	151,599		
Intercept			323,951		
Predicted			73,001		
-95,0%CL			67,766		
+95,0%CL			78,237		

Результат – 73^0 передбачена мінімальна температура січня.

6. Критерій χ^2 (Пірсона)

а) Критерій згоди χ^2

Критерій χ^2 дозволяє перевірити узгодженість результатів вимірювань з тим чи іншим статистичним розподілом.

Приклад 6.1. Наводимо величини e, знайдені Р. Міллікеном при визначенні заряду електрона, який дорівнює $e \cdot 10^{-10}$ од. СГСЕ.

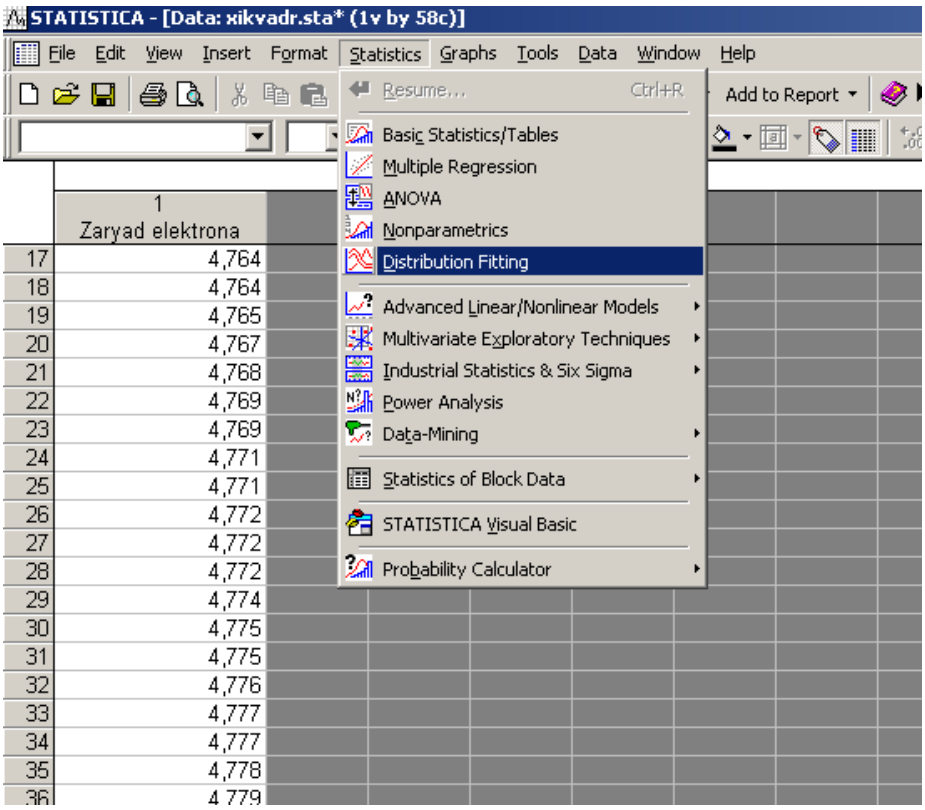
4,7 4,74 4,747 4,749 4,758 4,761 4,764 4,764 4,764 4,765
4,767 4,768 4,769 4,769 4,771 4,771 4,772 4,772 4,772 4,774
4,775 4,775 4,776 4,777 4,777 4,778 4,779 4,779 4,779 4,781
4,781 4,782 4,783 4,783 4,785 4,785 4,785 4,788 4,788 4,789
4,789 4,79 4,79 4,791 4,791 4,791 4,792 4,792 4,795 4,797
4,799 4,799 4,801 4,805 4,806 4,808 4,809 4,81

Перевірити гіпотезу про нормальний розподіл результатів вимірювань величини e при визначенні заряду електрона.

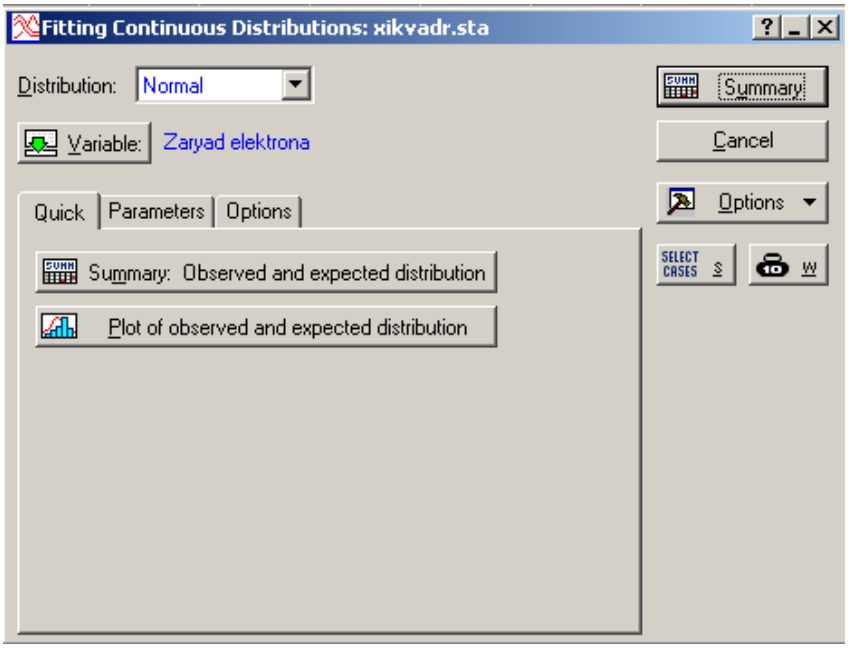
Висуваємо гіпотезу

H_0: генеральна сукупність розподілена нормально.

Створюємо файл з даними, задаючи тільки одну змінну (Можна скористатись файлом xikvadr(zaryad).sta). Натискаємо Statistics->Distribution Fitting.



Відмічаємо розподіл Normal. Відмічаємо нашу змінну. Потім натискаємо на Summary: Observed and expected distribution



Upper Boundary	Variable: Zaryad elektrona, Distribution: Normal (xikvadr.sta) Chi-Square = 3,11239, df = 3 (adjusted) , p = 0,37462								
	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 4,69000	0	0	0,00000	0,0000	0,00004	0,00004	0,00008	0,0001	-0,00004
4,70000	1	1	1,72414	1,7241	0,00052	0,00057	0,00090	0,0010	0,99948
4,71000	0	1	0,00000	1,7241	0,00495	0,00552	0,00854	0,0095	-0,00495
4,72000	0	1	0,00000	1,7241	0,03535	0,04086	0,06094	0,0705	-0,03535
4,73000	0	1	0,00000	1,7241	0,18982	0,23068	0,32727	0,3977	-0,18982
4,74000	1	2	1,72414	3,4483	0,76728	0,99796	1,32289	1,7206	0,23272
4,75000	2	4	3,44828	6,8966	2,33499	3,33294	4,02584	5,7465	-0,33499
4,76000	1	5	1,72414	8,6207	5,35086	8,68381	9,22563	14,9721	-4,35086
4,77000	9	14	15,51724	24,1379	9,23499	17,91880	15,92240	30,8945	-0,23499
4,78000	15	29	25,86207	50,0000	12,00522	29,92402	20,69865	51,5931	2,99478
4,79000	14	43	24,13793	74,1379	11,75566	41,67968	20,26839	71,8615	2,24434
4,80000	9	52	15,51724	89,6552	8,67094	50,35062	14,94989	86,8114	0,32906
4,81000	6	58	10,34483	100,0000	4,81726	55,16788	8,30562	95,1170	1,18274
4,82000	0	58	0,00000	100,0000	2,01558	57,18346	3,47513	98,5922	-2,01558
< Infinity	0	58	0,00000	100,0000	0,81654	58,00000	1,40783	100,0000	-0,81654

Бачимо, що досягнутий рівень значущості дорівнює $p = 0{,}37462$. Гіпотезу про нормальність приймаємо.

Тепер аналогічним чином перевіряємо гіпотезу про рівномірний розподіл вибірки (Rectangular). Досягнутий рівень значущості дорівнює $p = 0{,}00000$. Гіпотезу про рівномірну розподіленість відхиляємо.

б) Критерій χ^2 для гіпотези незалежності випадкових величин

Приклад 6.2. У таблиці наведені дані про 1426 ув'язнених, яких було класифіковано щодо алкогольної залежності (алкоголік, неалкоголік) і характеру злочинів, за які їх засудили. Чи можна на підставі цих даних дійти висновку про наявність зв'язку між алкоголізмом і характером злочину?

Вид злочину	Алкоголіки	Неалкоголіки
1.Підпал	50	43
2.Згвалтування	88	62
3.Насильницькі дії	155	110
4.Крадіжка	379	300
5.Виготовлення фальшивих грошей	18	14
6.Шахрайство	63	144

Висуваємо гіпотезу

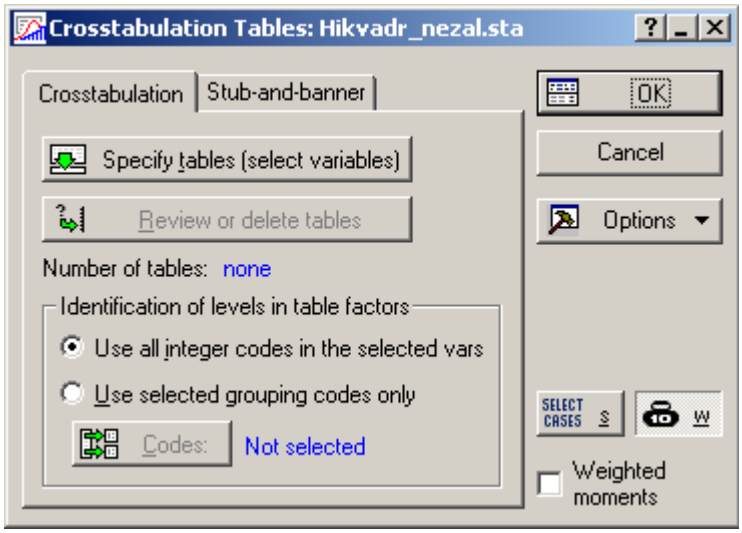
H_0: вид злочину і алкогольна залежність не пов'язані між собою.

Створюємо файл з даними, задаючи три змінні : Vid, Alcohol і n. (Можна скористатись файлом Hikvadr_nezal.sta). Перша змінна може приймати значення від 1 до 6 – задає вид злочину, друга від 1 до 2 – несе інформацію про алкогольну залежність, а третя n задає відповідні частоти.

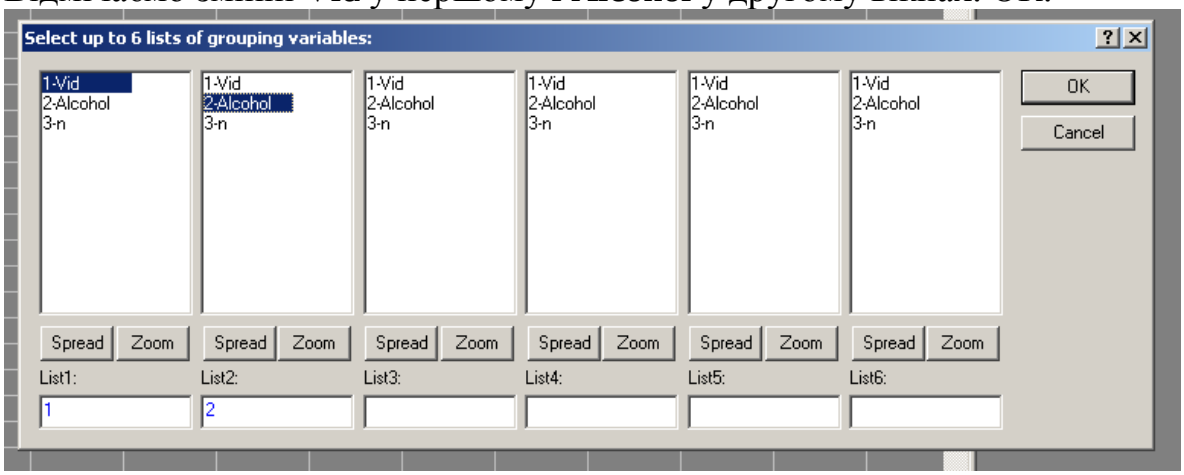
Data: Hikvadr_nezal.sta (3v by 12c)

	1 Vid	2 Alcohol	3 n
1	1	1	50
2	2	1	88
3	3	1	155
4	4	1	379
5	5	1	18
6	6	1	63
7	1	2	43
8	2	2	62
9	3	2	110
10	4	2	300
11	5	2	14
12	6	2	144

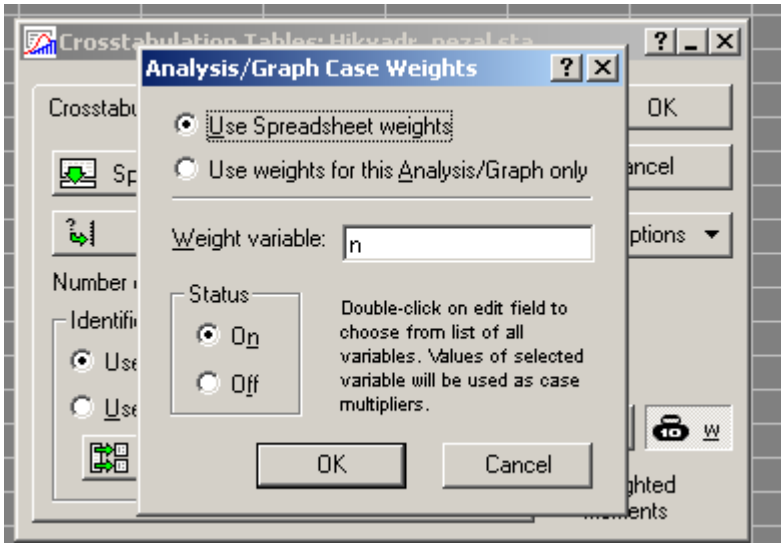
Для того, щоб перевірити гіпотезу за критерієм χ^2, здійснюємо наступні дії. Натискаємо Statistics->Basic Statistics/Tables->Tables and Banners->OK->Specify tables



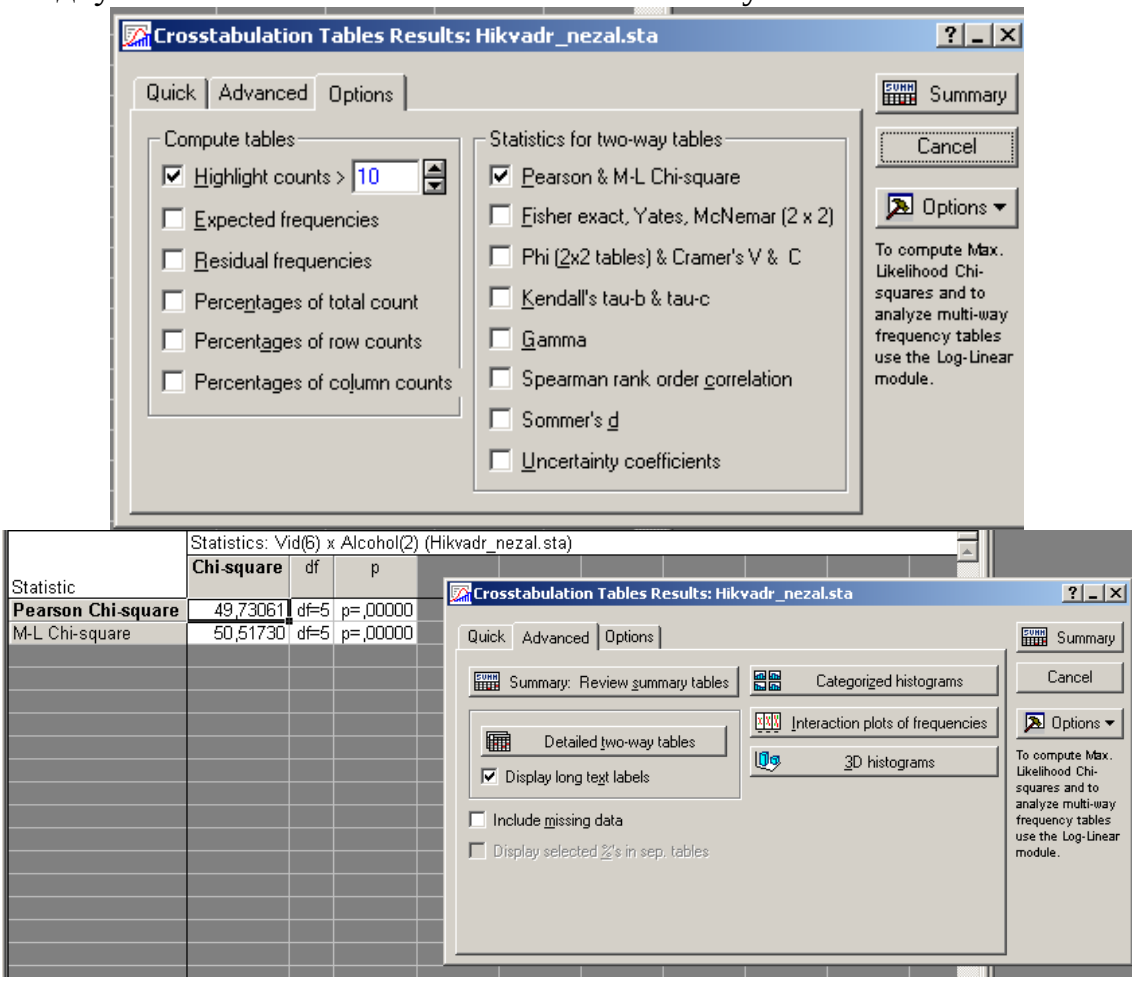
Відмічаємо змінні Vid у першому і Alcohol у другому вікнах. ОК.



Потім натискаємо на гирю, яка знаходиться у правому нижньому куточку з написом W. Виставляємо Status в «on», Weight variable задаємо як n, потім натискаємо ОК, знов ОК.



На закладці Options відмічаємо Pearson & M-L Chi-square. Переходимо на закладку Advanced і натискаємо Detailed two-way tables.

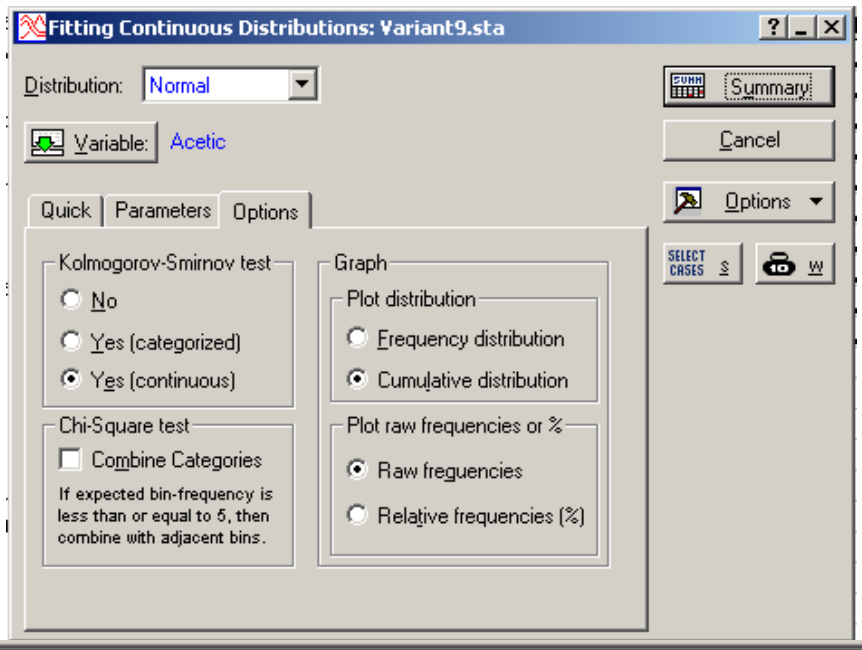


В першому рядочку бачимо, що значення статистики χ^2=49,73061, число ступенів свободи=5, досягнутий рівень значущості 0. Отже, гіпотезу про відсутність зв’язку відхиляємо, зв’язок є.

7. Критерій Колмогорова

Критерій Колмогорова, як і критерій χ^2, дозволяє перевірити узгодженість результатів вимірювань з тим чи іншим статистичним розподілом, але цей розподіл обов'язково повинний бути неперервним.

Для застосування цього критерію треба перейти Statistics-> Distribution Fitting. Потім відмітити розподіл, на узгодженість з яким ви перевіряєте ваші дані. ОК. Відмітити змінну. На закладці Options відмітити: Kolmogorov-Smirnov test – Yes(continuous), Plot distribution – Cumulative distribution, Chi-Square test – зняти галочку. Ще треба вказати кількість інтервалів групування на закладці Parameters – Number of categories (напр., 15 штук).



Variable: Acetic, Distribution: Normal (Variant9.sta)
Kolmogorov-Smirnov d = 0,15780, p = n.s., Lilliefors p = n.s.
Chi-Square = 11,96204, df = 17, p = 0,80243

Upper Boundary	**Observed Frequency**	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	C E
<= 4,40000	0	0	0,00000	0,0000	0,383699	
4,50000	0	0	0,00000	0,0000	0,187353	
4,60000	1	1	10,00000	10,0000	0,252850	
4,70000	1	2	10,00000	20,0000	0,329134	
4,80000	0	2	0,00000	20,0000	0,413227	
4,90000	1	3	10,00000	30,0000	0,500395	
5,00000	0	3	0,00000	30,0000	0,584446	
5,10000	0	3	0,00000	30,0000	0,658390	
5,20000	1	4	10,00000	40,0000	0,715367	
5,30000	1	5	10,00000	50,0000	0,749691	
5,40000	1	6	10,00000	60,0000	0,757780	
5,50000	0	6	0,00000	60,0000	0,738773	
5,60000	0	6	0,00000	60,0000	0,694683	
5,70000	1	7	10,00000	70,0000	0,630042	

Оскільки в другому рядочку написано про рівень значущості $p = n.s.$ - несуттєвий, то гіпотезу про узгодженість розподілів приймаємо.

8. Дисперсійний аналіз

Дисперсійний аналіз – це сукупність методів, які дозволяють перевіряти: чи змінюється середнє значення характеристик деяких об'єктів в залежності від дії деякого фактора або кількох факторів.

а) Однофакторний дисперсійний аналіз

Нехай ми маємо k рівнів фактора. Специфікація моделі має вигляд

$$x_{ij} = a_i + \varepsilon_{ij}; \quad i = \overline{1,k}; \quad j = \overline{1,n},$$

де a_i - середнє значення, обчислене при i – му рівні фактора; ε_{ij} - випадкова величина, результат впливу неврахованих факторів, ε розподілена $N(0,\sigma^2)$.

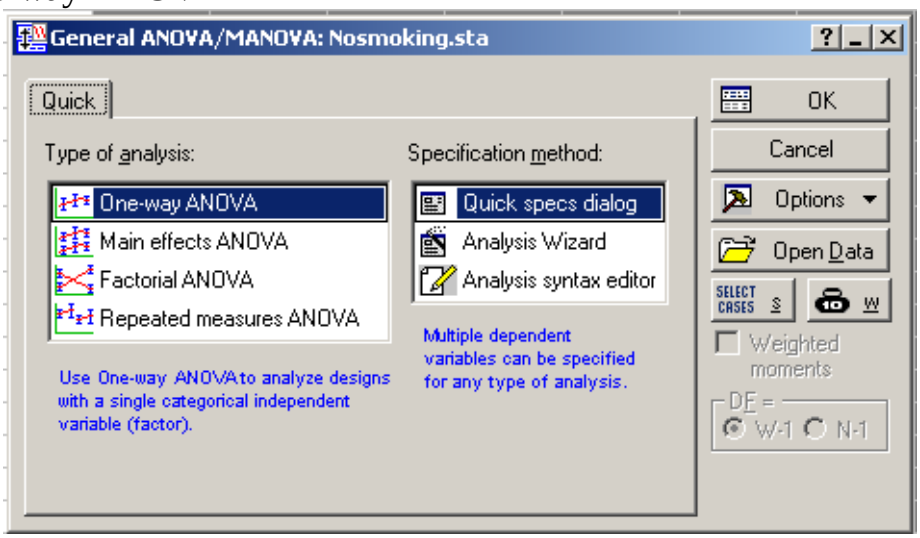
Приклад 8.1. Куріння вважається головним фактором, що сприяє хронічним захворюванням легенів. А пасивне куріння таким фактором зазвичай не вважається. Для характеристики стану дихальних шляхів взяли один з показників функції дихання – максимальну об'ємну швидкість середини видиху, яку вимірювали під час профілактичного огляду співробітників підприємства. Зменшення цього показника – ознака порушення прохідності дихальних шляхів. Дані обстеження наведені в таблиці.

Група (фактор)	**Максимальна об'ємна швидкість середини видиху (л/с) (Vidih)**
1. Некурящі, працюють в помешканні, де не палять (n=10)	3,548; 2,1874; 3,1646; 3,4153; 3,5168; 2,6718; 3,4502; 2,5873; 3,4662; 3,3884
2. Некурящі, працюють в накуреному помешканні (n=10)	2,7941; 2,5126; 2,3435; 3,1944; 2,5801; 3,5152; 3,1748; 3,1443; 2,8229; 2,9739
3. Курці, що викурюють невелику кількість цигарок (n=10)	1,8486; 3,1194; 2,961; 1,4257; 2,747; 2,0553; 1,6891; 1,6228; 1,933; 2,8235
4. Курці, що викурюють середню кількість цигарок (n=10)	2,5115; 2,285; 1,3353; 2,9383; 2,3571; 1,3424; 2,0629; 2,7382; 2,3747; 2,1532
5. Курці, що викурюють велику кількість цигарок (n=10)	1,7544; 2,4652; 1,6647; 1,8542; 2,9761; 2,5873; 2,8684; 3,5921; 0,9386; 2,2418

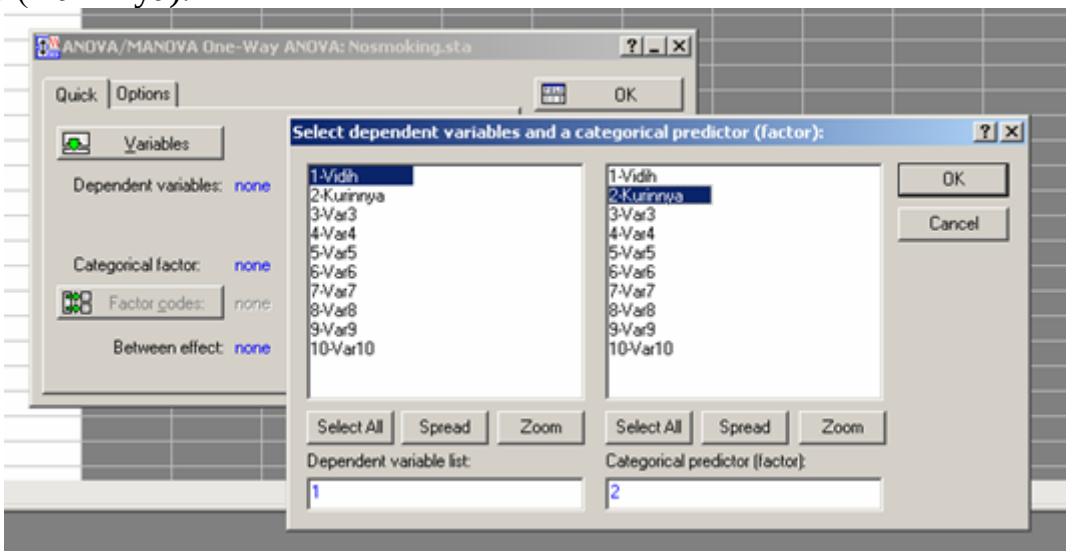
Перевіряємо гіпотезу про рівність середніх в кожній з 5-ти груп, що відповідають п'яти рівням фактора. Тобто, якщо приймаємо гіпотезу H_0– впливу фактора немає.

$$H_0 : a_1 = a_2 = a_3 = a_4 = a_5.$$

Створюємо файл з даними, змінна Vidih відповідає максимальній об’ємній швидкості середини видиху; змінна Kurinnya відповідає рівню фактора і приймає значення 1-5. Заходимо в модуль ANOVA, обираємо тип аналізу One-way ANOVA



Відмічаємо відповідні змінні – залежну (Vidih) і ту, що відповідає за фактор (Kurinnya).



Натискаємо ОК, і знов ОК. Далі натискаємо All effects. Отримуємо:

Univariate Tests of Significance for Vidih (Nosmoking.sta)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	326,2408	1	326,2408	1005,797	0,000000
Kurinnya	7,6229	4	1,9057	5,875	0,000683
Error	14,5962	45	0,3244		

Оскільки досягнутий рівень значущості $p = 0{,}000683 < 0{,}01$ - гіпотезу H_0 відхиляємо, приймаємо альтернативну гіпотезу про нерівність середніх при різних рівнях фактора. Фактор впливає на досліджувану змінну.

б) Двофакторний дисперсійний аналіз

Нехай на досліджувану ознаку впливають одразу 2 фактори А і В. Ми маємо p рівнів фактора А і q рівнів фактора В. Таким чином, кожне спостереження x_{ijk} залежить від трьох індексів – рівня фактору А, рівня фактору В і номера спостереження. Специфікація моделі має вигляд

$$x_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}; \quad i = \overline{1,p}; \quad j = \overline{1,q}; \quad k = \overline{1,n},$$

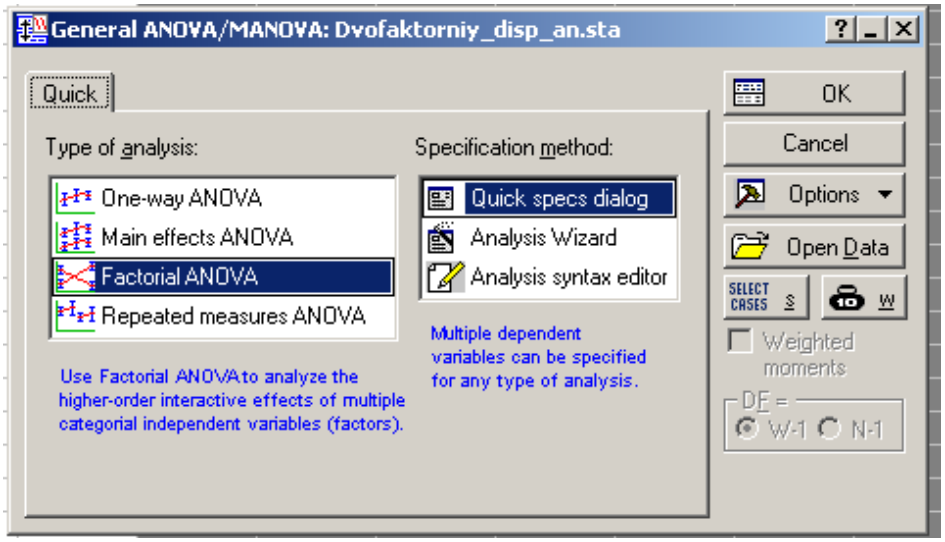
де μ - загальне середнє, α_i, β_j - ефекти впливу факторів А і В відповідно; γ_{ij} - ефект взаємодії; ε_{ijk} - випадкова величина, результат впливу неврахованих факторів (неспецифічна компонента), ε розподілена $N(0, \sigma^2)$.

Приклад8.2. Досліджувався вплив факторів на рейтинг правих політичних партій (у %). Фактор А відповідає віку респондентів (A_1-в діапазоні 20-35 років; A_2 - 36-50 років; A_3 – 51-70 років). Фактор В відповідає регіону, де проводилося опитування (B_1-західний регіон; B_2-центральний; B_3-східний).

B\A	A_1	A_2	A_3
B_1	25,2 10,2 5,4 13,2 18,2 5,2 13,4 15,2 4,5 19,2	4,3 10,5 20,3 32,4 5,6 12,4 6,2 9,8 16,8 18,4	14,3 10,6 28,4 10,8 7,4 6,5 4,5 26,3 30,2 11,8
B_2	10,6 8,4 11,2 4,6 5,8 18,2 16,4 13,2 4,8 8,9	12,4 4,3 13,2 5,6 8,9 14,8 22,3 6,8 7,2 11,4	6,2 7,5 3,5 12,4 13,5 16,4 7,9 8,9 15,4 10,8
B_3	2,5 6,4 12,5 14,8 12,3 8,5 5,9 8,9 15,4 12,8	4,5 4,9 12,3 15,6 7,9 8,9 9,8 13,9 4,2 6,9	14,8 2,9 5,9 10,6 8,5 13,4 2,2 19,5 7,9 9,9

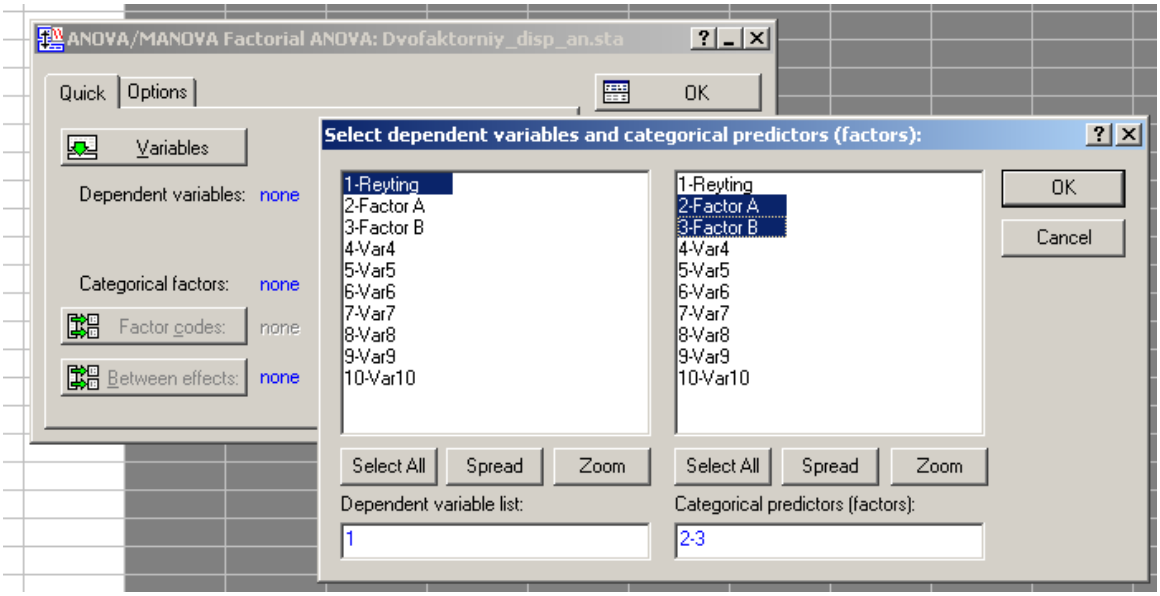
Створюємо файл з даними, змінна Reyting відповідає відносній кількості мешканців, які на наступних виборах збираються голосувати за праву політичну партію; також задаємо змінні Factor A і Factor B.

Заходимо в модуль ANOVA, обираємо тип аналізу Factorial ANOVA:



Відмічаємо відповідні змінні – залежну (Reyting) і обидві, що відповідають за фактори (Factor A, Factor B).

Потім натискаємо ОК, знову ОК і All effects. Отримуємо



Effect	Univariate Tests of Significance for Reyting (Dvofaktorniy_disp_an.st Sigma-restricted parameterization Effective hypothesis decomposition SS	Degr. of Freedom	MS	F	p
Intercept	11406,38	1	11406,38	299,6700	0,000000
Factor A	6,24	2	3,12	0,0820	0,921342
Factor B	327,90	2	163,95	4,3073	0,016684
Factor A*Factor B	24,53	4	6,13	0,1611	0,957390
Error	3083,11	81	38,06		

Бачимо, що досягнутий рівень значущості $p_A = 0{,}921342 > 0{,}01$ - отже, гіпотезу про відсутність впливу фактора А на рейтинг приймаємо, рейтинг правих політичних сил не залежить від віку опитаних. Для фактору В досягнутий рівень значущості $p_B = 0{,}016684$. Отже, якщо ми приймемо рівень значущості 0,01, то вважаємо, що впливу фактора В немає, якщо приймемо $\alpha = 0{,}05$ - вважаємо, що вплив фактора В на Reyting присутній.

9. Факторний аналіз

Факторний аналіз – метод багатовимірного аналізу, який дозволяє дослідити структуру зв'язку між змінними. В деяких задачах він дозволяє суттєво зменшити число регресорів в моделі і вилучити з моделі мультиколінеарність.

Нехай в нашій вибірці присутні n об'єктів, кожний має d змінних-характеристик $\left(x_i^j\right)\ i=\overline{1,n};\ j=1,d$. Розглядаємо модель з такою специфікацією

$$x_i^j = \bar{x}^j + \sum_{k=1}^{d} a_{jk} f_i^k + \varepsilon_{ij} \quad i=\overline{1,n};\ j=1,d\ .$$

Тут $\bar{x}^j$ - середнє по j-тій змінній, a_{jk} називається навантаженням на j-ту змінну, f_i^k - фактором, ε_{ij} - стохастична компонента - називається специфічністю.

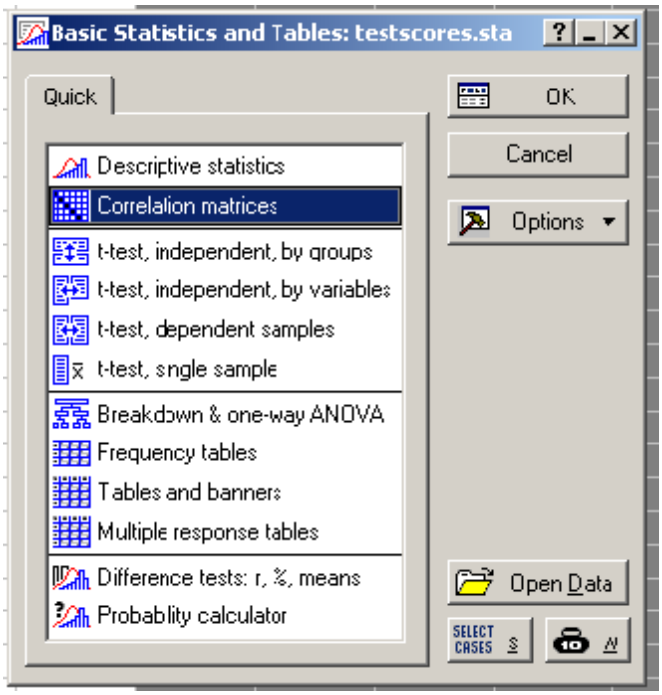
Приклад 9.1. Розкриємо файл testscores.sta. Маємо дані про оцінки 25 студентів з дисциплін «Диференціальна геометрія» (difgeom), «Комплексний аналіз» (complex), «Алгебра» (algebra), «Математичний аналіз» (reals), «Статистика» (statistics). Ми хочемо побудувати лінійну модель залежності оцінки зі статистики від інших оцінок. Статистика вивчається на 5 курсі, в магістратурі. А всі інші предмети здаються раніше, в бакалавраті. Нам треба по оцінках студента-бакалавра спрогнозувати – чи буде він мати достатньо високі оцінки зі статистики в магістратурі, чи доцільно його туди приймати?

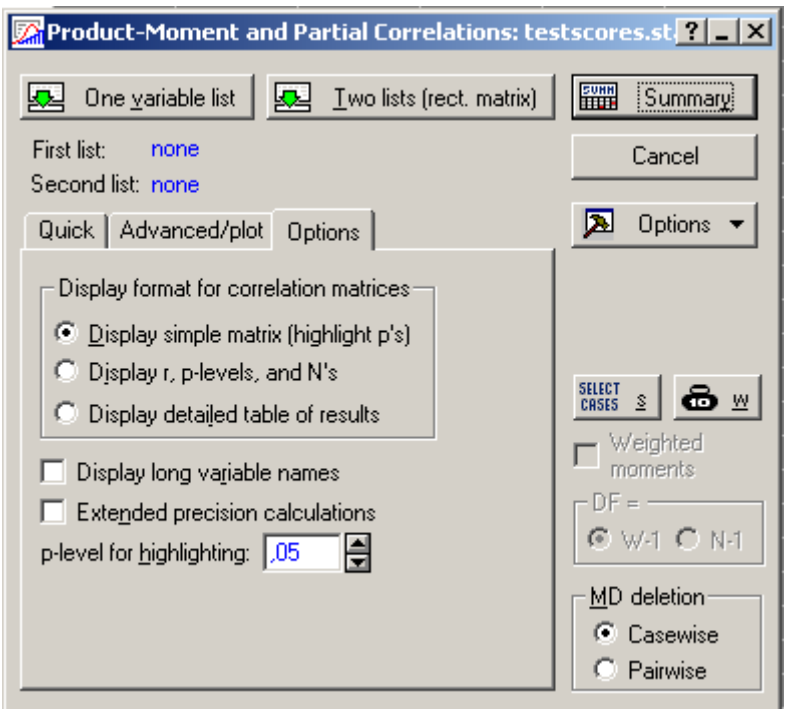
Спершу побудуємо модель множинної регресії

$$statistics = \beta_0 + \beta_1 difgeom + \beta_2 complex + \beta_3 a\lg ebra + \beta_4 reals + \varepsilon$$

N=25	Regression Summary for Dependent Variable: statistics (testscores.sta) R= ,86546555 R?= ,74903061 Adjusted R?= ,69883674 F(4,20)=14,923 p<,00001 Std.Error of estimate: 9,0619					
	Beta	Std.Err. of Beta	B	Std.Err. of B	t(20)	p-level
Intercept			-17,7161	11,96971	-1,48008	0,154433
difgeom	0,242951	0,208338	0,2148	0,18421	1,16614	0,257275
complex	0,008726	0,209399	0,0119	0,28641	0,04167	0,967173
algebra	0,487502	0,238615	0,8242	0,40342	2,04305	0,054449
reals	0,191123	0,215631	0,2442	0,27557	0,88634	0,385967

Парадокс: коефіцієнт детермінації досить великий $R^2 = 0{,}749$, а при цьому жодна змінна не є значущою. Це свідчить про високий рівень мультиколінеарності в моделі. Наявна сильна попарна кореляція. Подивимося на кореляційну матрицю змінних. Перейдемо Statistics->Basic Statistics/Tables->Correlation matrices





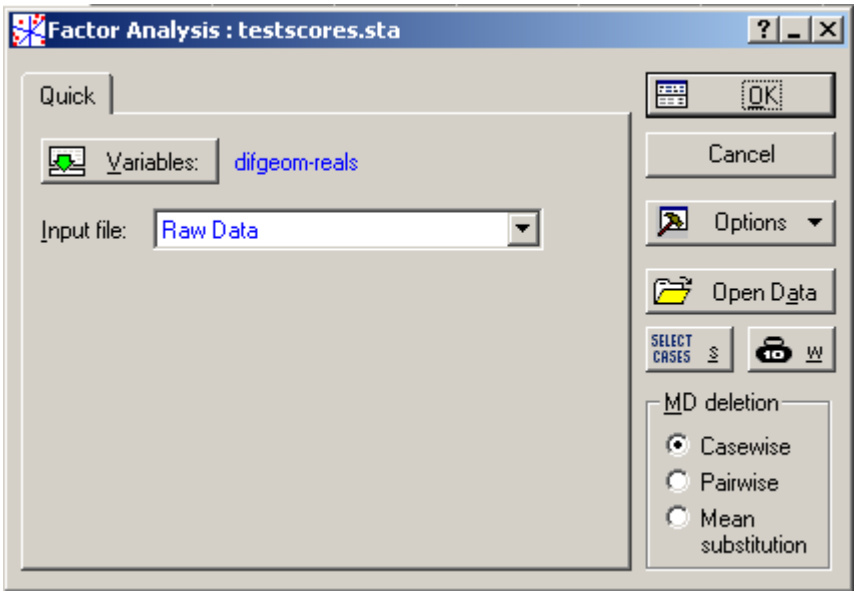
Відмічаємо One variable list, відмічаємо список з чотирьох змінних, натискаємо ОК, потім Summary, маємо:

Correlations (testscores.sta)
Marked correlations are significant at p < ,05000
N=25 (Casewise deletion of missing data)

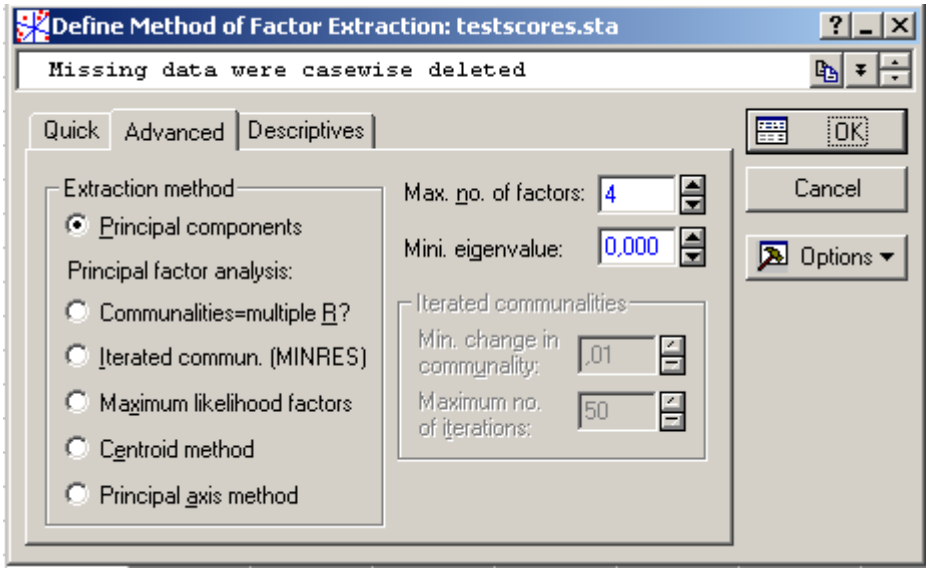
Variable	**difgeom**	complex	algebra	reals
difgeom	1,00	0,81	0,75	0,73
complex	0,81	1,00	0,77	0,70
algebra	0,75	0,77	1,00	0,84
reals	0,73	0,70	0,84	1,00

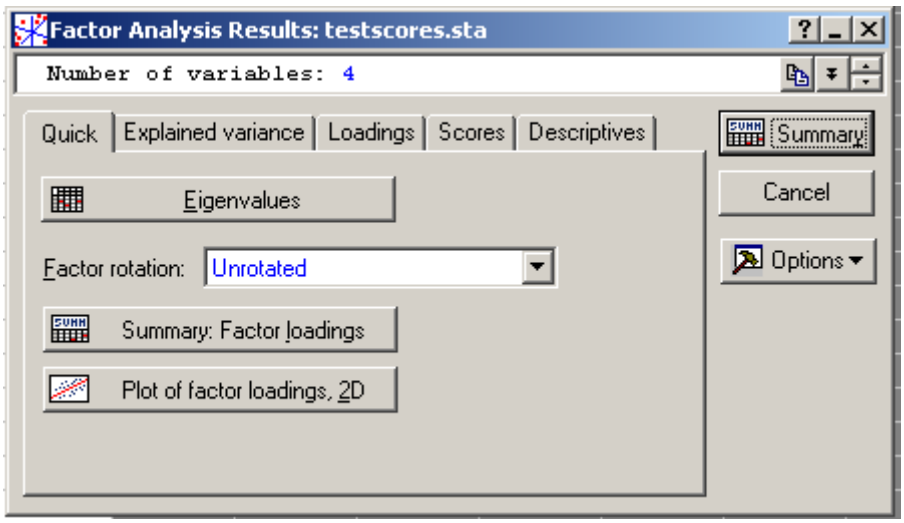
Спробуємо розв’язати задачу по-іншому, спираючись на модель з факторного аналізу. Ми припускаємо, що наші 4 змінні залежать від факторів, яких ми не можемо спостерігати.

Переходимо Statistics-> Multivariate Exploratory Techniques->Factor Analysis. Відмічаємо наші 4 пояснюючі змінні.



На закладці Advanced відмічаємо метод Principal components (метод головних компонент) , Максимальна кількість факторів Max. no of factors: 4; мінімальне власне значення матриці навантажень Mini. eigenvalue: 0. Натискаємо ОК.





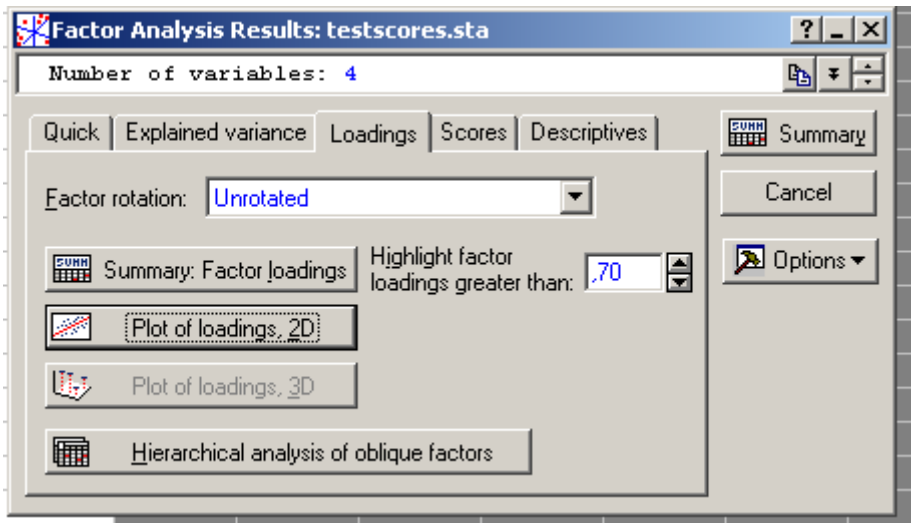
На закладці Quick натискаємо Eigenvalues.

Value	Eigenvalues (testscores.sta) Extraction: Principal components			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,302269	82,55673	3,302269	82,5567
2	0,352380	8,80951	3,654649	91,3662
3	0,206875	5,17186	3,861524	96,5381
4	0,138476	3,46190	4,000000	100,0000

Бачимо, що 4 фактори пояснюють всю дисперсію, але вже перший фактор пояснює 82,56% дисперсії. Спробуємо задати 2 фактори.

Value	Eigenvalues (testscores.sta) Extraction: Principal components			
	Eigenvalue	% Total variance	Cumulative Eigenvalue	Cumulative %
1	3,302269	82,55673	3,302269	82,55673
2	0,352380	8,80951	3,654649	91,36624

2 фактори пояснюють 91% дисперсії. Мабуть, на такій кількості факторів треба зупинитись. На закладці Loadings натискаємо на Summary: Factor loadings



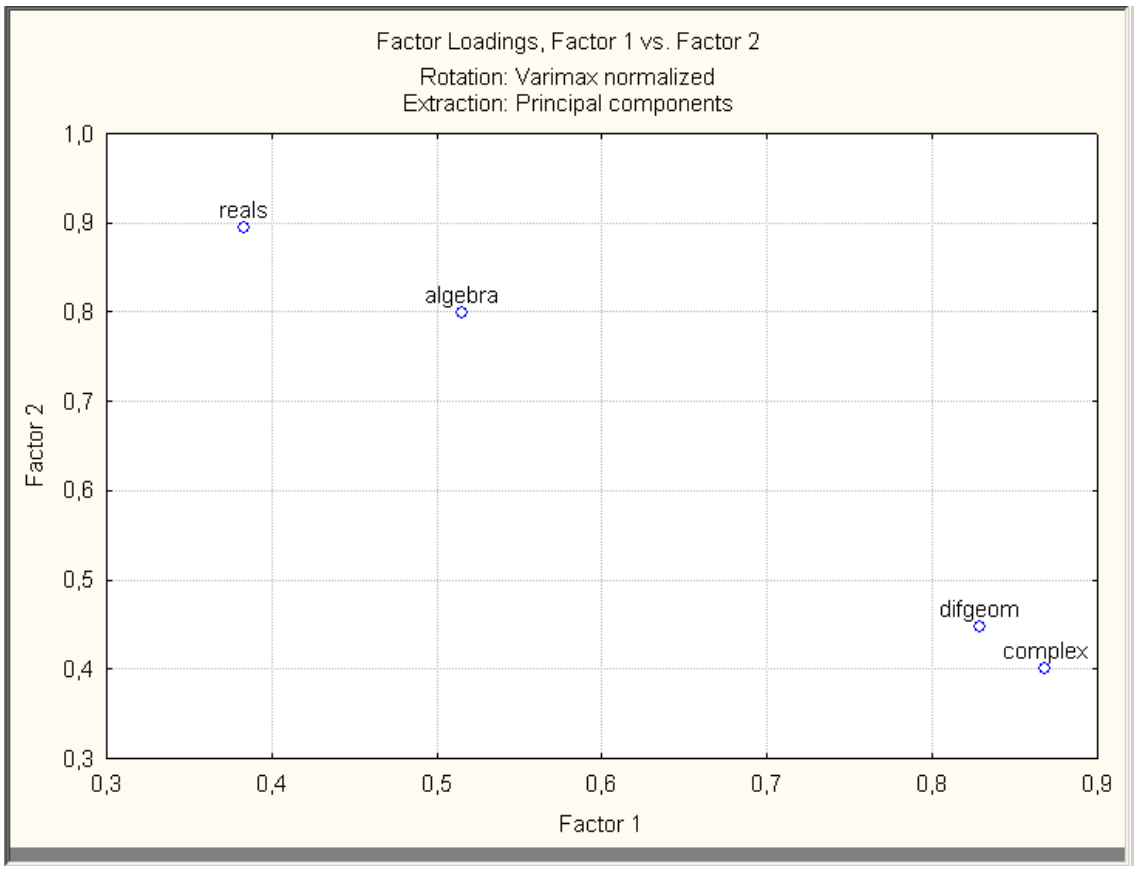
Variable	Factor Loadings (Unrotated) (testscores.sta) Extraction: Principal components (Marked loadings are > ,700000) **Factor 1**	Factor 2
difgeom	-0,905697	-0,261060
complex	-0,900800	-0,322308
algebra	-0,928140	0,209635
reals	-0,899499	0,369322
Expl.Var	3,302269	0,352380
Prp.Totl	0,825567	0,088095

Бачимо, що основне навантаження сконцентроване на факторі 1.

Тепер застосовуємо обертання. Задаємо Factor rotation як Varimax normalized на закладці Quick і натискаємо на Summary.

Variable	Factor Loadings (Varimax normalized) (testscores.sta) Extraction: Principal components (Marked loadings are > ,700000) **Factor 1**	Factor 2
difgeom	0,829285	0,448023
complex	0,868689	0,400878
algebra	0,515627	0,799699
reals	0,383340	0,893615
Expl.Var	1,855155	1,799495
Prp.Totl	0,463789	0,449874

Як бачимо, difgeom, complex тісно пов’язані з першим фактором, algebra, reals – з другим. Подивимось на відповідний графік. На закладці Loadings натиснемо на Plot of load 2D. Отримуємо



Для отримання значень факторів на закладці Scores переходимо до Factor Scores.

Factor Scores (testscores.sta)
Rotation: Varimax normalized
Extraction: Principal components

Case	**Factor 1**	Factor 2
1	1,07217	-1,48753
2	1,21782	-0,63125
3	-0,34905	-0,62201
4	2,15850	0,86053
5	0,28203	0,43455
6	-0,92622	-0,61569
7	-0,57210	0,73430
8	-0,55338	0,75087
9	-0,92842	-0,82192
10	-1,04438	-0,49371
11	0,01126	-0,29858
12	1,15795	-0,91296
13	1,46960	1,25524
14	-1,72942	2,12106
15	-1,03300	0,63196
16	0,14276	0,42856
17	-0,08794	-0,26571
18	-0,99223	-1,71200
19	1,34660	-1,42983
20	1,07069	1,46013
21	-0,19295	-0,15594
22	-0,70847	-0,04964
23	-0,10560	0,72403
24	0,25190	1,13345
25	-0,95813	-1,03793

Наша мета – знайти формулу для прогнозування оцінки зі статистики. Копіюємо значення факторів у дві нові змінні – Factor1 та Factor2 і будуємо модель з двома регресорами. Як ми бачимо, модель вдала – всі змінні є значущими, коефіцієнт детермінації майже не змінився, $R^2 = 0{,}737$.

5 statistics	6 Factor1	7 Factor2	8 Var8	9 Var9	10 Var10
37	1,07217	-1,48753			
52	1,21782	-0,63125			
29	-0,34905	-0,62201			
81	2,15850	0,86053			
40	0,28203	0,43455			

gression Summary for Dependent Variable: statistics (testscores.sta)

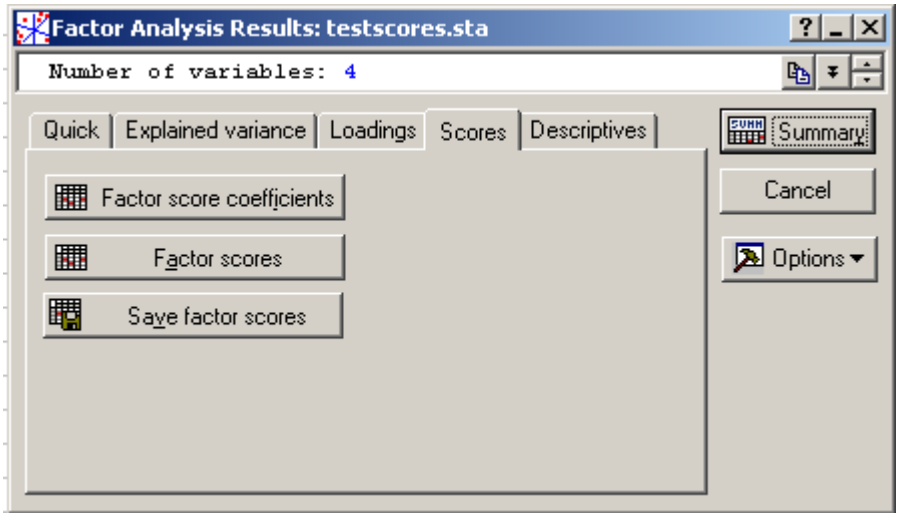
Regression Summary for Dependent Variable: statistics (testscores.sta)
R= ,85891966 R?= ,73774299 Adjusted R?= ,71390144
F(2,22)=30,944 p<,00000 Std.Error of estimate: 8,8323

N=25	Beta	Std.Err. of Beta	B	Std.Err. of B	t(22)	p-level
Intercept			43,80000	1,766460	24,79535	0,000000
Factor1	0,533691	0,109182	8,81263	1,802886	4,88807	0,000069
Factor2	0,672991	0,109182	11,11285	1,802886	6,16392	0,000003

Прогноз можна здійснювати за формулою

$$statistics = 43{,}8 + 0{,}533691\,Factor1 + 0{,}672991\,Factor2.$$

Проблема виникає тільки в тому, що ми не знаємо значень *Factor*1 та *Factor*2 для нових об'єктів. Їх треба виразити через оцінки difgeom, complex, algebra, reals. Для цього на закладці Scores треба клацнути на кнопку Factor score coefficients.



Variable	Factor Score Coefficients (testscores.sta) Rotation: Varimax normalized Extraction: Principal components **Factor 1**	Factor 2
difgeom	0,714648	-0,336681
complex	0,835328	-0,461777
algebra	-0,216073	0,621473
reals	-0,539662	0,938844

Значення факторів обчислюємо за формулами

$$Factor1 = 0,714648\frac{(difgeom - \overline{difgeom})}{\sigma_{difgeom}} + 0,835328\frac{(complex - \overline{complex})}{\sigma_{complex}} -$$

$$-0,216073\frac{(a\lg ebra - \overline{a\lg ebra})}{\sigma_{a\lg ebra}} - 0,539662\frac{(reals - \overline{reals})}{\sigma_{reals}};$$

$$Factor2 = -0,33668\frac{(difgeom - \overline{difgeom})}{\sigma_{difgeom}} - 0,46178\frac{(complex - \overline{complex})}{\sigma_{complex}} +$$

$$+0,62147\frac{(a\lg ebra - \overline{a\lg ebra})}{\sigma_{a\lg ebra}} + 0,938844\frac{(reals - \overline{reals})}{\sigma_{reals}}.$$

Тут $\sigma_{difgeom}, \sigma_{complex}, \sigma_{a\lg ebra}, \sigma_{reals}$ - відповідні вибіркові середньоквадратичні відхилення.

Зміст

Література

1. Кокс Д., Снелл Э. Прикладная статистика. Принципы и примеры. - М.: "Мир", 1984.
2. Турчин В.М. Теорія ймовірностей і математична статистика. Основні поняття, приклади, задачі: Підручник. – Д.: Вид-во Дніпропетр. нац. ун-ту, 2006. – 476 с.
3. Гланц С. Медико-биологическая статистика. – М.:Практика, 1998. – 459 с.
4. Майборода Р.Є. Регресія: Лінійні моделі: Навчальний посібник. – К.:ВПЦ «Київський університет», 2007. – 296 с.
5. Оленко А.Я. Комп’ютерна статистика: Навчальний посібник. – К.:ВПЦ «Київський університет», 2007. – 174 с.
6. Мамчич Т., Оленко А., Осипчук М., Шпортюк В. Статистичний аналіз даних з пакетом STATISTICA. – Дрогобич: Відродження, 2006.
7. StatSoft’s Electronic Statistics Textbook (http://www.statsoft.com/textbook/stathome.html)

Навчальне видання

МАЙБОРОДА *Ростислав Євгенович*
СУГАКОВА *Олена Володимирівна*

СТАТИСТИЧНИЙ АНАЛІЗ ДАНИХ ЗА ДОПОМОГОЮ ПАКЕТУ STATISTICA

Підписано до друку 12.10.2012. Формат 60x80[16].
Гарнітура Times. Папір офсетний. Друк офсетний.
Наклад 30 примірників. Ум. друк. арк. 4.

Видавнича лабораторія радіофізичного факультету
Київського національного університету імені Тараса Шевченка