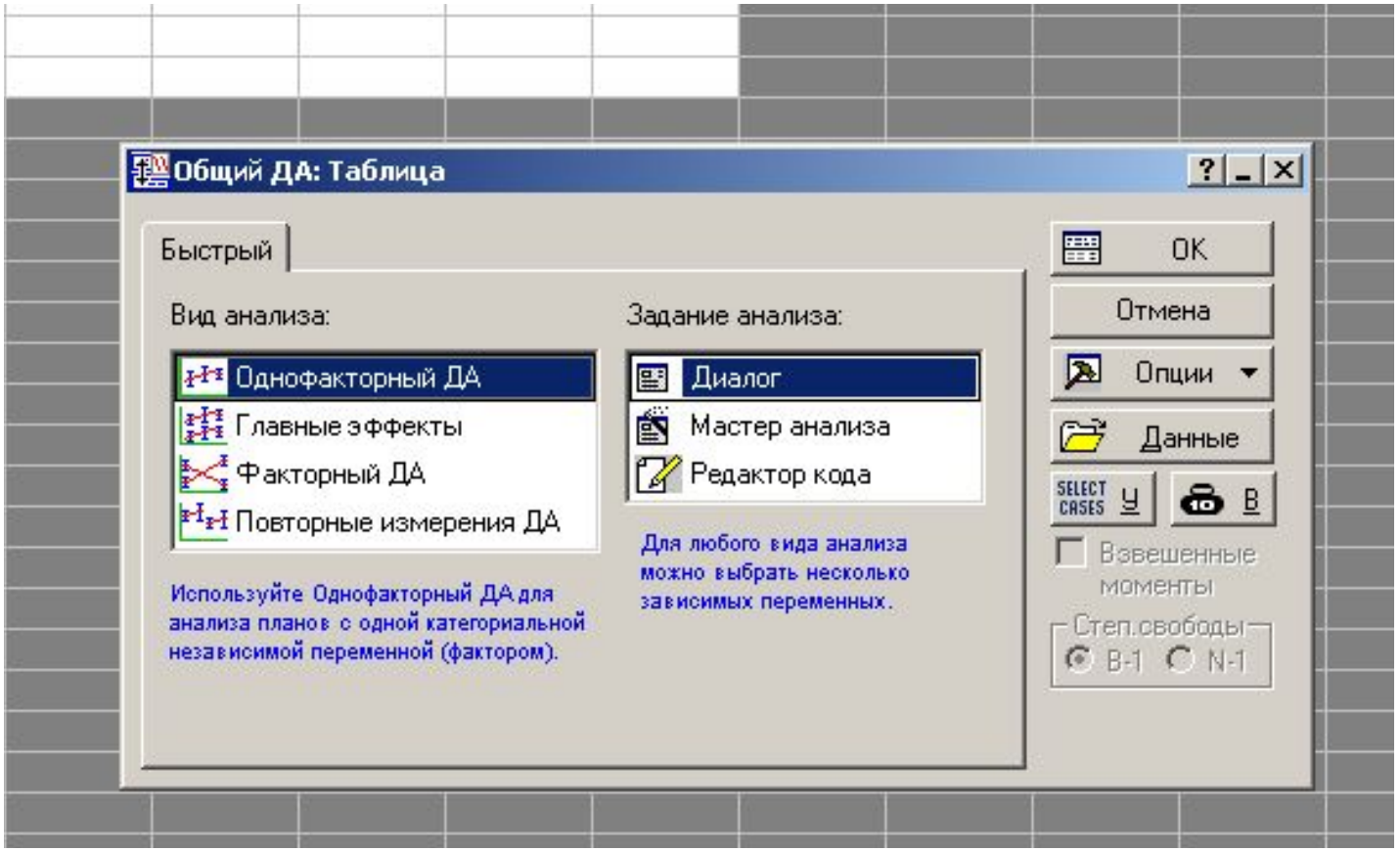
Лекція 4. Огляд методів статистичного моделювання

Однофакторний та багатофакторний дисперсійний аналіз

- Основною метою дисперсійного аналізу є дослідження значущості відмінності між середніми, тобто його використовують для перевірки статистичних гіпотез.

В пакеті Statistica «Анализ»- «Дисперсионный анализ»



РЕГРЕСІЙНИЙ АНАЛІЗ

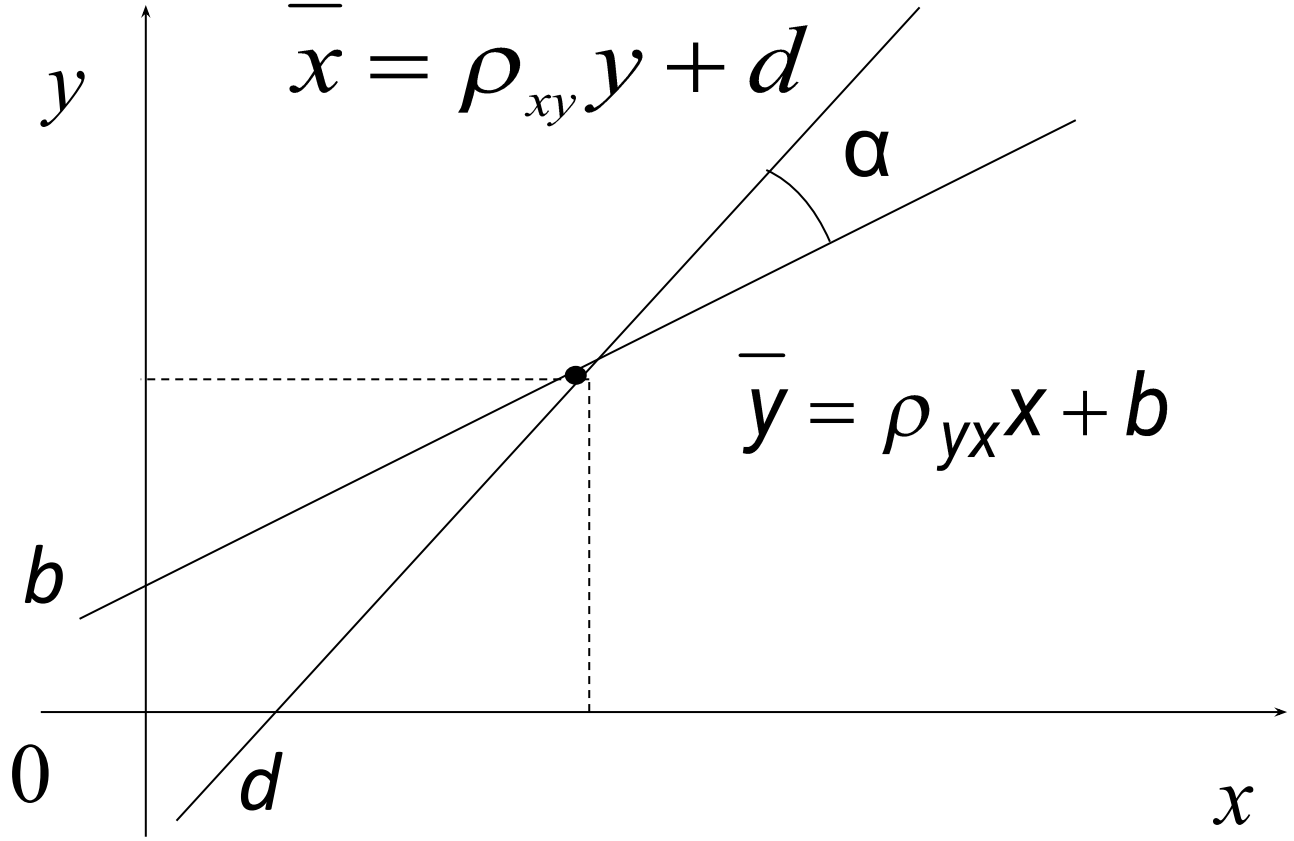
- Загальне призначення регресійного аналізу полягає в дослідженні зв'язку між однією або декількома незалежними змінними (званими також регресорами або предикторами) та залежною змінною.
- За видом залежності виокремлюють лінійну та нелінійну регресію. Нелінійні регресійні функції в свою чергу поділяються на ті, що можуть бути приведені до лінійної форми, та так звані «суттєво нелінійні».

РЕГРЕСІЙНИЙ АНАЛІЗ

- За кількістю незалежних змінних виокремлюють парну регресію та множинну. Таким чином найпростішою є парна лінійна регресія.

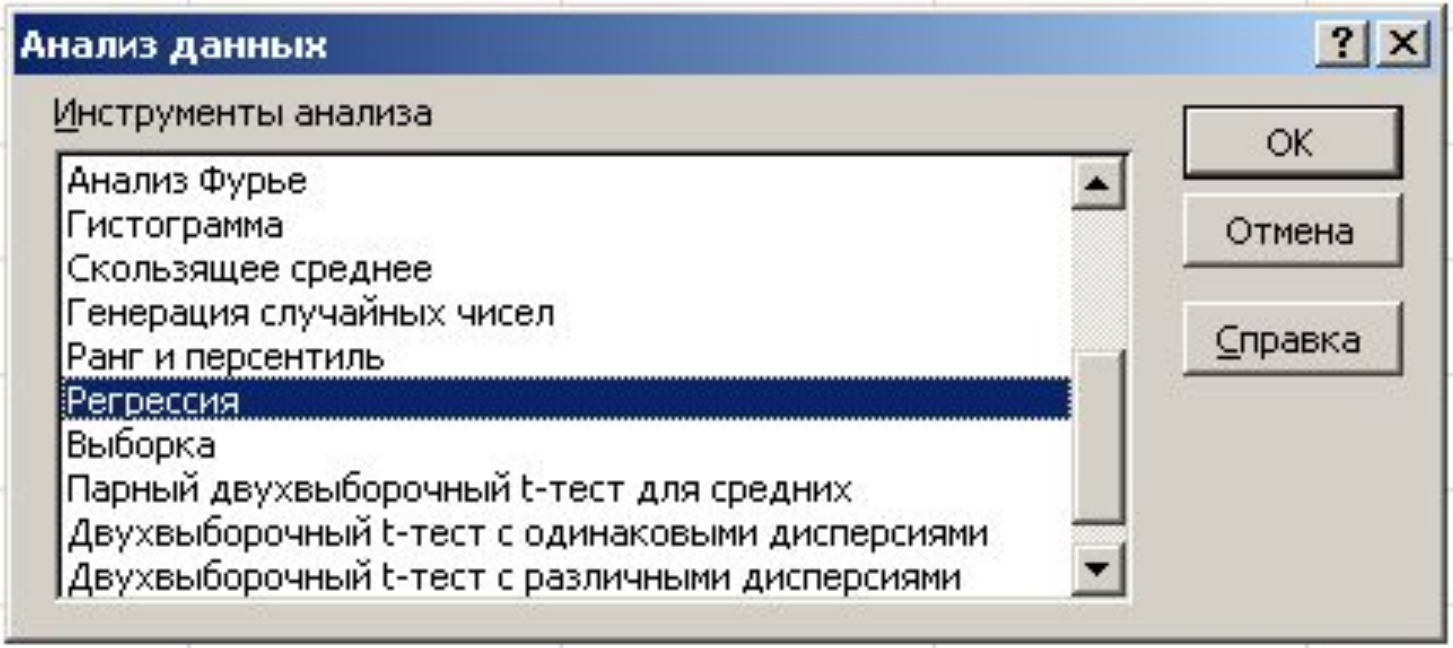
- рівняння парної лінійної регресії $\bar{y} = \rho_{yx} x + b$
- Y на X: $\bar{x} = \rho_{xy} y + d$
- X на Y: ρ_{yx} ρ_{xy} ,
- де , , b і d - коефіцієнти лінійної регресії, які знаходять методом найменших квадратів (МНК).

Графічне представлення ліній регресії

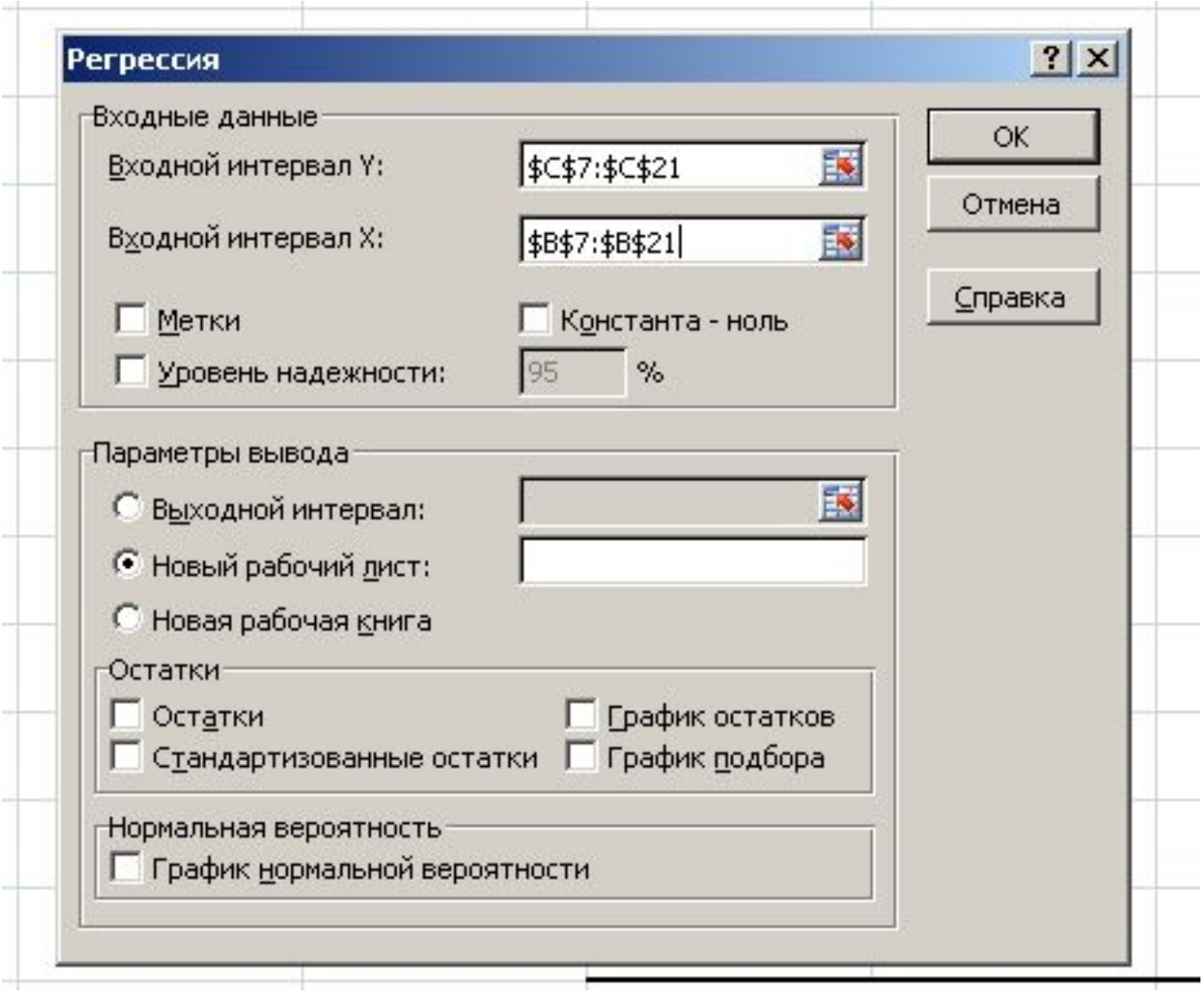


Регресійний аналіз в MS Excel

- Обираємо «Данные»
  -
- «Анализ данных» -
- «Ре



Регресійний аналіз в MS Excel



Регресійний аналіз в MS Excel

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,637553
R-квадрат	0,406474
Нормированный R-квадрат	0,360818
Стандартная ошибка	24,99363
Наблюдения	15

$$y = 6{,}189 + 0{,}959x$$

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	5561,537	5561,537	8,902994	0,010563089
Остаток	13	8120,863	624,6817		
Итого	14	13682,4			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	6,189128	18,90203	0,327432	0,748554	-34,64622604	47,02448
Переменная X 1	0,959183	0,321465	2,983788	0,010563	0,264700619	1,653666

Регресійний аналіз в MS Excel

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,93617
R-квадрат	0,876414
Нормированный R-кв	0,804985
Стандартная ошибка	24,18358
Наблюдения	15

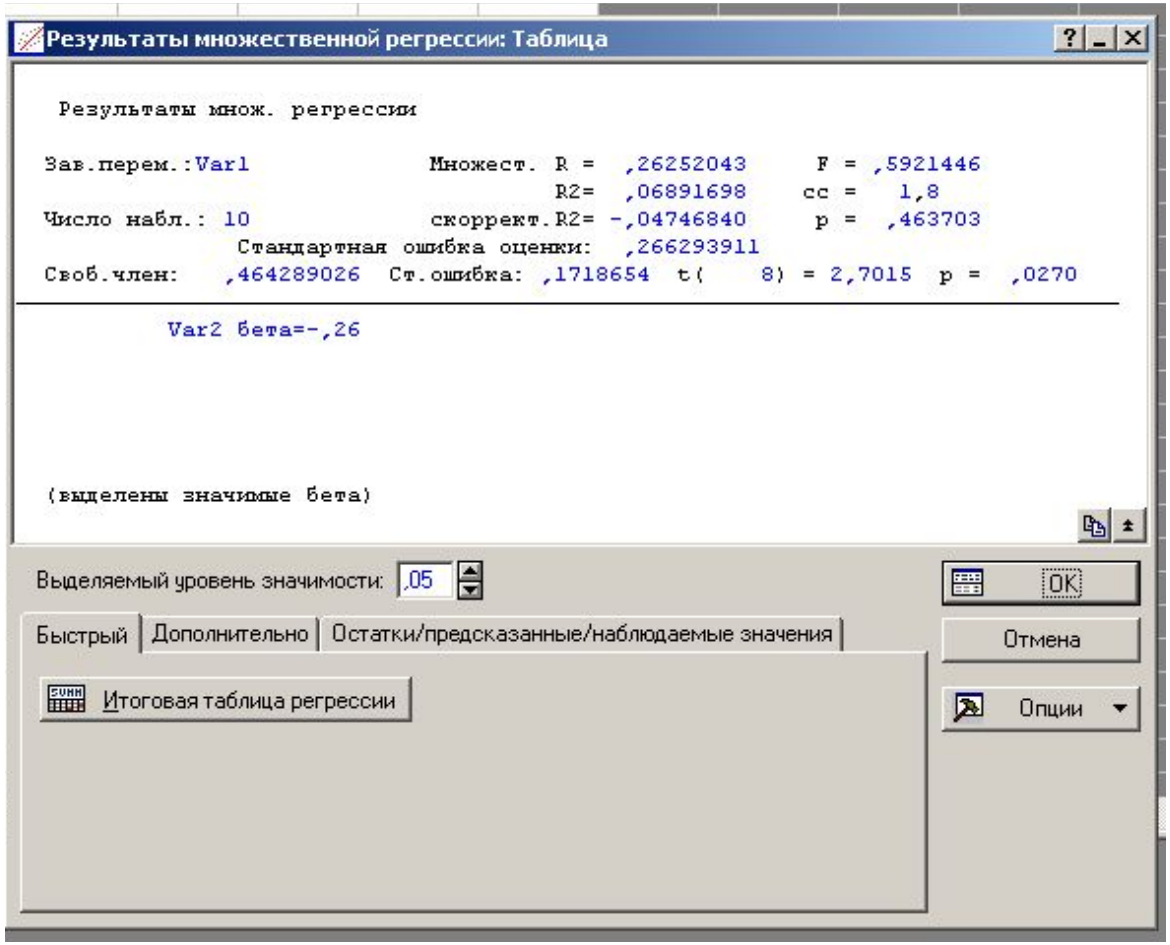
Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	58064,16	58064,16	99,28122	1,87611E-07
Остаток	14	8187,836	584,8454		
Итого	15	66252			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
Переменная X 1	1,058117	0,106194	9,963996	9,77E-08	0,83035335	1,28588

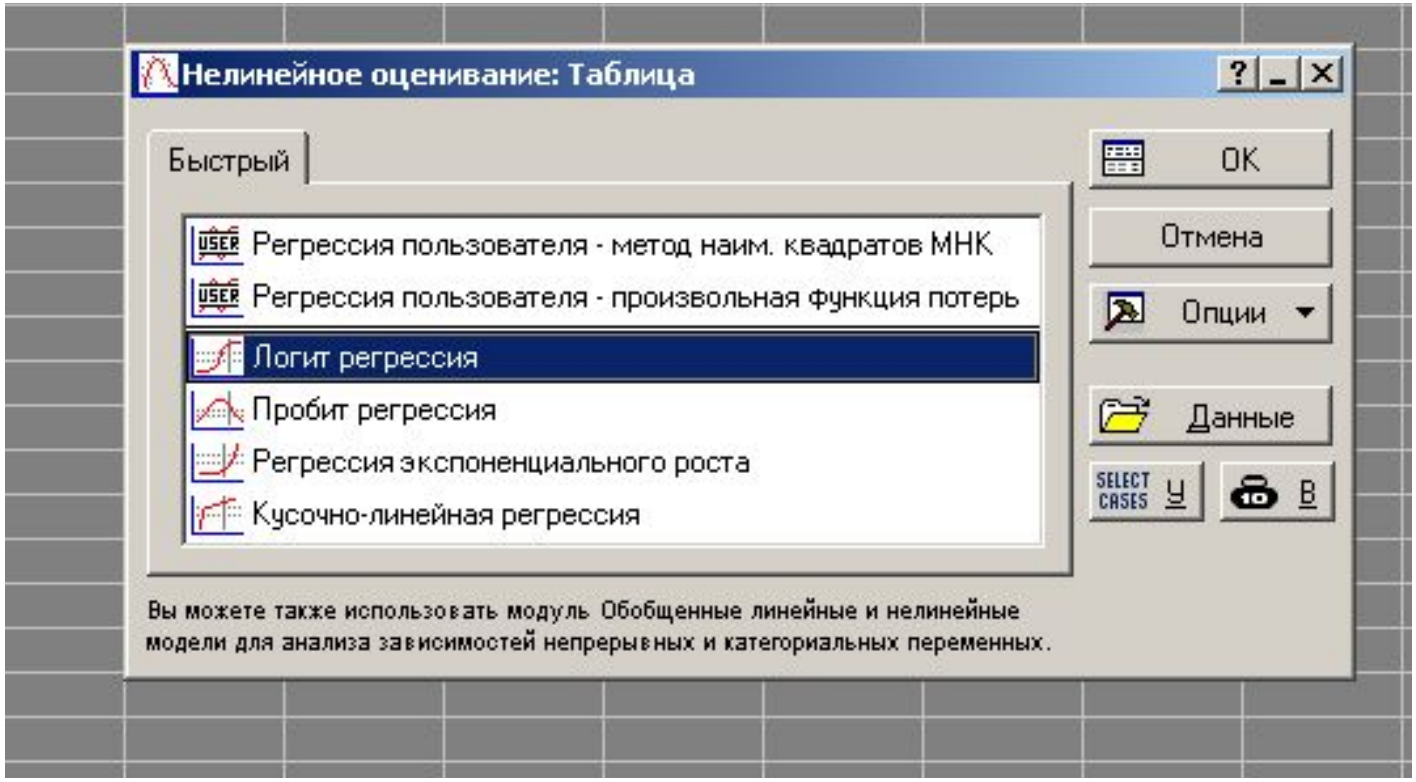
$$y = 1{,}058x$$

Регресійний аналіз в Statistica: лінійна регресія
«Анализ» – «Множественная регрессия»

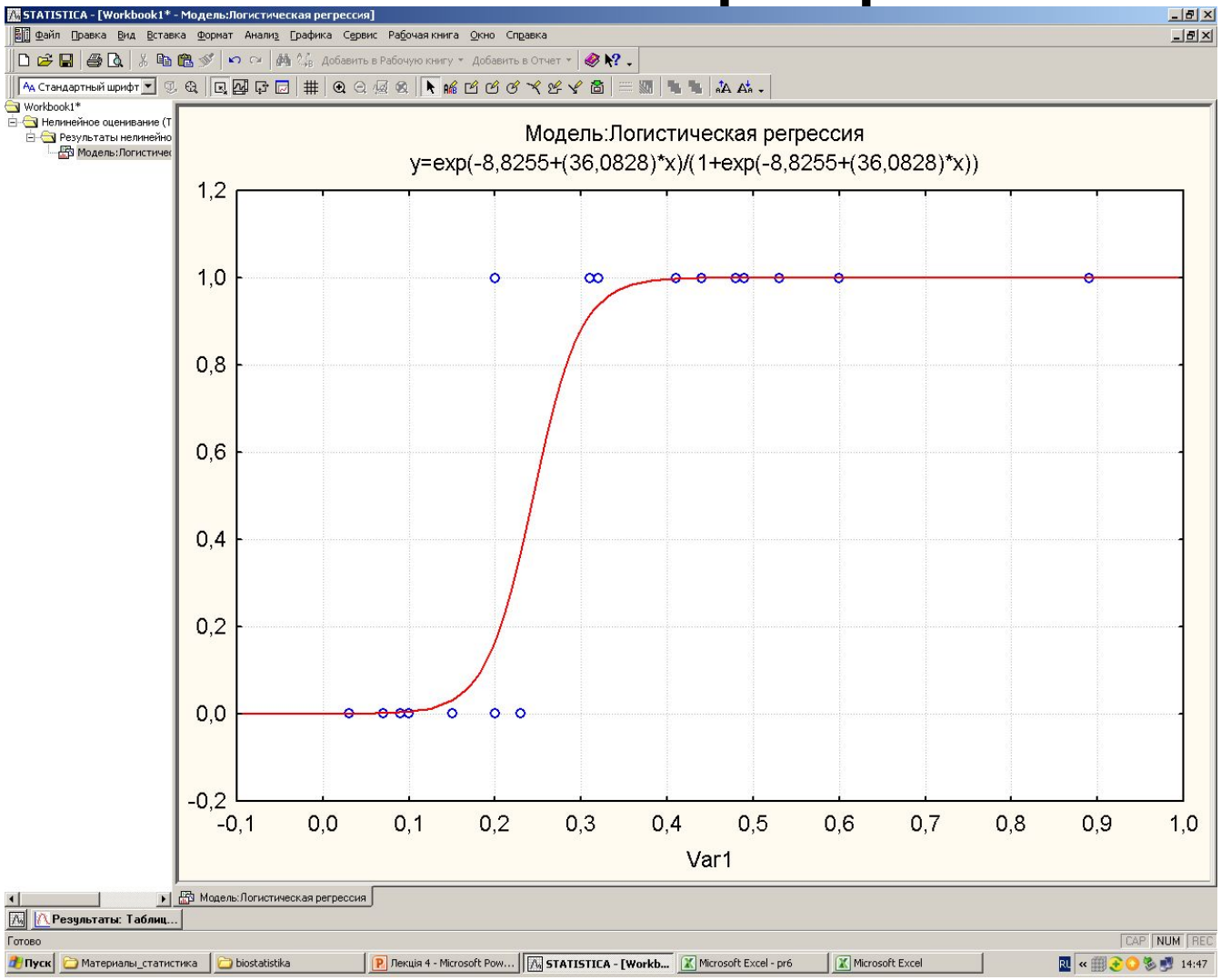


Регресійний аналіз в Statistica: логістична регресія

- Анализ – Углубленные методы анализа – Нелинейное оценивание – логит-регрессия



Регресійний аналіз в Statistica: логістична регресія



Методи статистичної класифікації: кластерний та дискримінантний аналіз

- **Кластерний аналіз** (англ. *Data clustering*) — задача розбиття заданої вибірки об'єктів (ситуацій) на підмножини, що називаються кластерами, так, щоб кожен кластер складався зі схожих об'єктів, а об'єкти різних кластерів істотно відрізнялися.
- в пакеті Statistica кластерний аналіз здійснюється
- *«Анализ» - «Многомерный разведочный анализ» – «Кластерный анализ»*

Дискримінантний аналіз

- Дискримінантний аналіз — це статистичний метод, призначений для вивчення відмінностей між двома або більшою кількістю груп об'єктів з використанням даних про різноманітність кількох ознак, що відрізняють ці об'єкти один від одного. Типове для дискримінантного аналізу завдання — визначення тих ознак, які найкраще дискримінують (відрізняють) об'єкти, що відносяться до різних груп. Після того, як визначені найкращі способи дискримінації наявних груп (тобто проведена інтерпретація відмінностей між ними), цей спосіб аналізу дозволяє проводити класифікацію об'єктів, належність яких до тієї чи іншої групи заздалегідь невідома.

Етапи дискримінантного аналізу

- Для проведення дискримінантного аналізу введемо позначення
- x_{ik} - значення k-тої ознаки у i-го пацієнта основної групи ($i = 1 \ldots n_1 \quad k = 1 \ldots K$)
- y_{jk} - значення k-тої ознаки у j-го пацієнта контрольної групи ($j = 1 \ldots n_2$)
- 1. знаходимо середні значення $\overline{x_k}$ та $\overline{y_k}$

Етапи дискримінантного аналізу

- 2. обраховуємо коваріаційні матриці S_x та S_y
- «Данные» - «Анализ данных» - «Ковариация»

- 3. розраховуємо сумарну коваріаційну матрицю:

$$\hat{S} = \frac{1}{n_1 + n_2 - 2}\left[n_1 S_x + n_2 S_y\right]$$

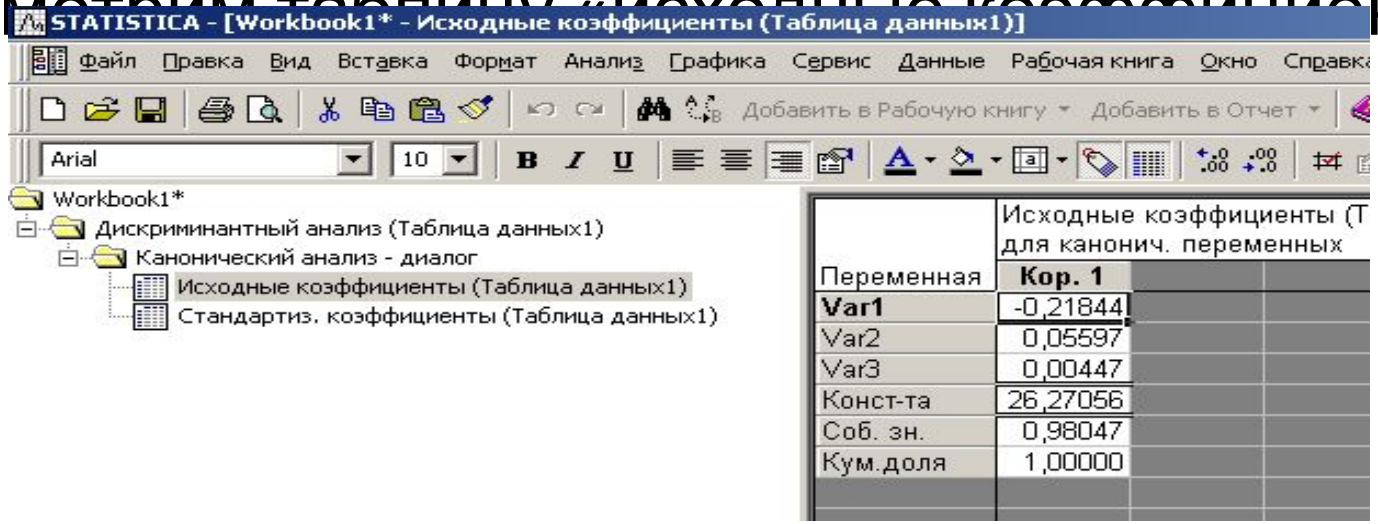
- 4.обчислюємо обернену матрицю $\hat{S}^{-1}$
- Функція “МОБР”, F2, Shift+Ctrl+Enter

Етапи дискримінантного аналізу

- 5. Обчислюємо вектор оцінок коефіцієнтів дискримінантної функції $$a = \hat{S}^{-1}(\bar{x} - \bar{y})$$
- Функція “МУМНОЖ”, F2, Shift+Ctrl+Enter
- 6. Знаходимо оцінки дискримінантної функції для кожного пацієнта основної та контрольної груп
- 7. Обчислюємо середні значення оцінок $\bar{Z}_x$ та $\bar{Z}_y$
- 8. Знаходимо константу (межу) дискримінації $$c = (\bar{Z}_x + \bar{Z}_y)/2$$

Дискримінантний аналіз в Statistica

- Анализ – Многомерный разведочный анализ – Дискриминантный анализ – Выбрать переменные – Дополнительно – Канонический анализ – Коэффициенты для канонических переменных
- Смотрим таблицу «исходные коэффициенты»



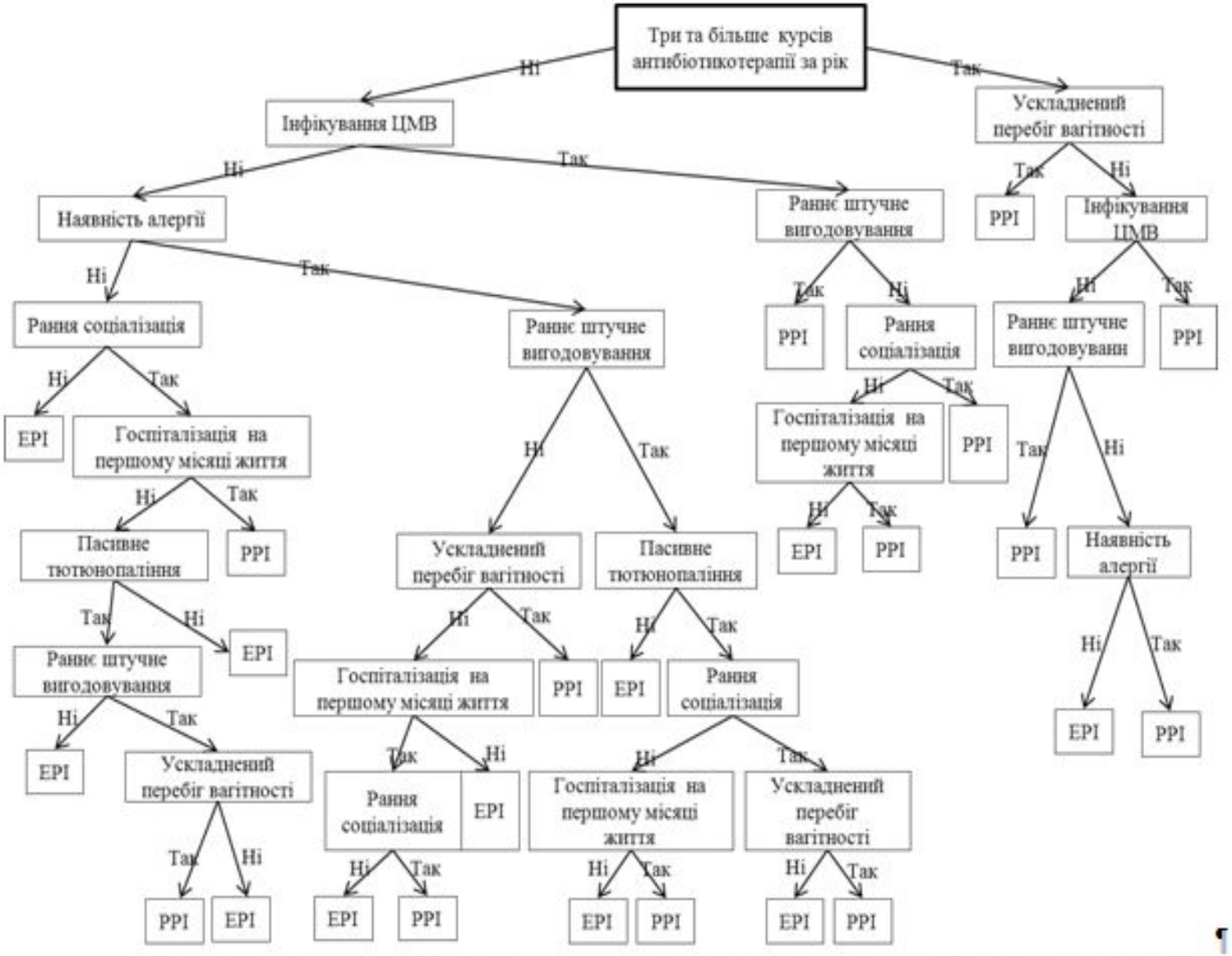
Методи статистичної класифікації: факторний аналіз

- ***факторний аналіз,*** зокрема метод головних компонентів та канонічний аналіз, - багатовимірний метод, застосовуваний для вивчення взаємозв'язків між значеннями змінних. Передбачається, що відомі змінні залежать від меншої кількості невідомих змінних і випадкової помилки.
- в пакеті Statistica факторний аналіз здійснюється
- *«Анализ» - «Многомерный разведочный анализ» – «Факторный анализ»*

побудова дерева рішень

- Дерева рішень (*decision trees*) є одним з найбільш популярних методів вирішення завдань класифікації та прогнозування. Дерева рішень дозволяють візуально і аналітично оцінити результати вибору різних рішень. Дерева рішень використовують, коли потрібно прийняти рішення в умовах невизначеності, коли кожне рішення залежить від результату попередніх рішень або деяких заданих умов, що з'являються з певною ймовірністю.
- в пакеті Statistica *«Анализ»* - «*Многомерный разведочный анализ» – «Деревья классификации»*

Приклад дерева рішень

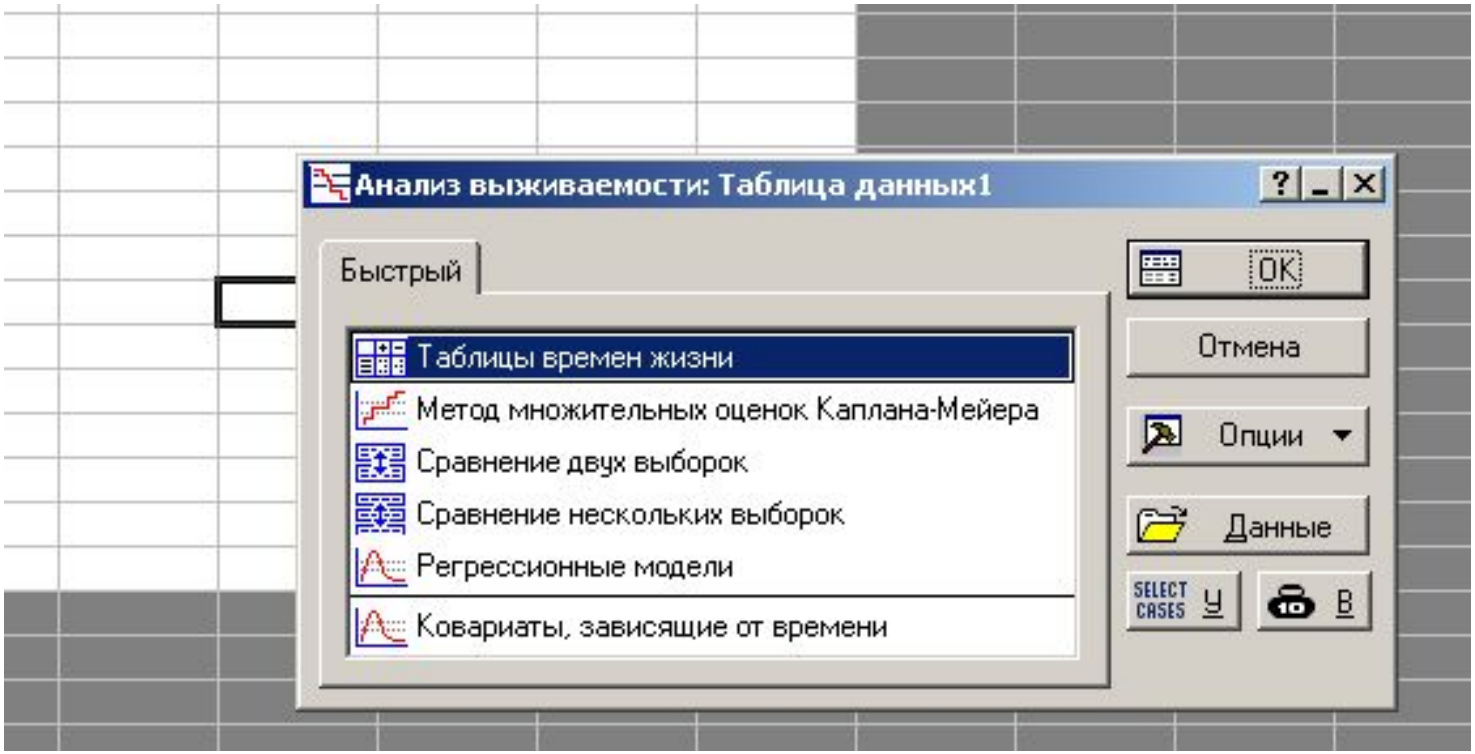


Прогнозування ймовірності появи досліджуваного результату в певний період часу (аналіз дожиття).

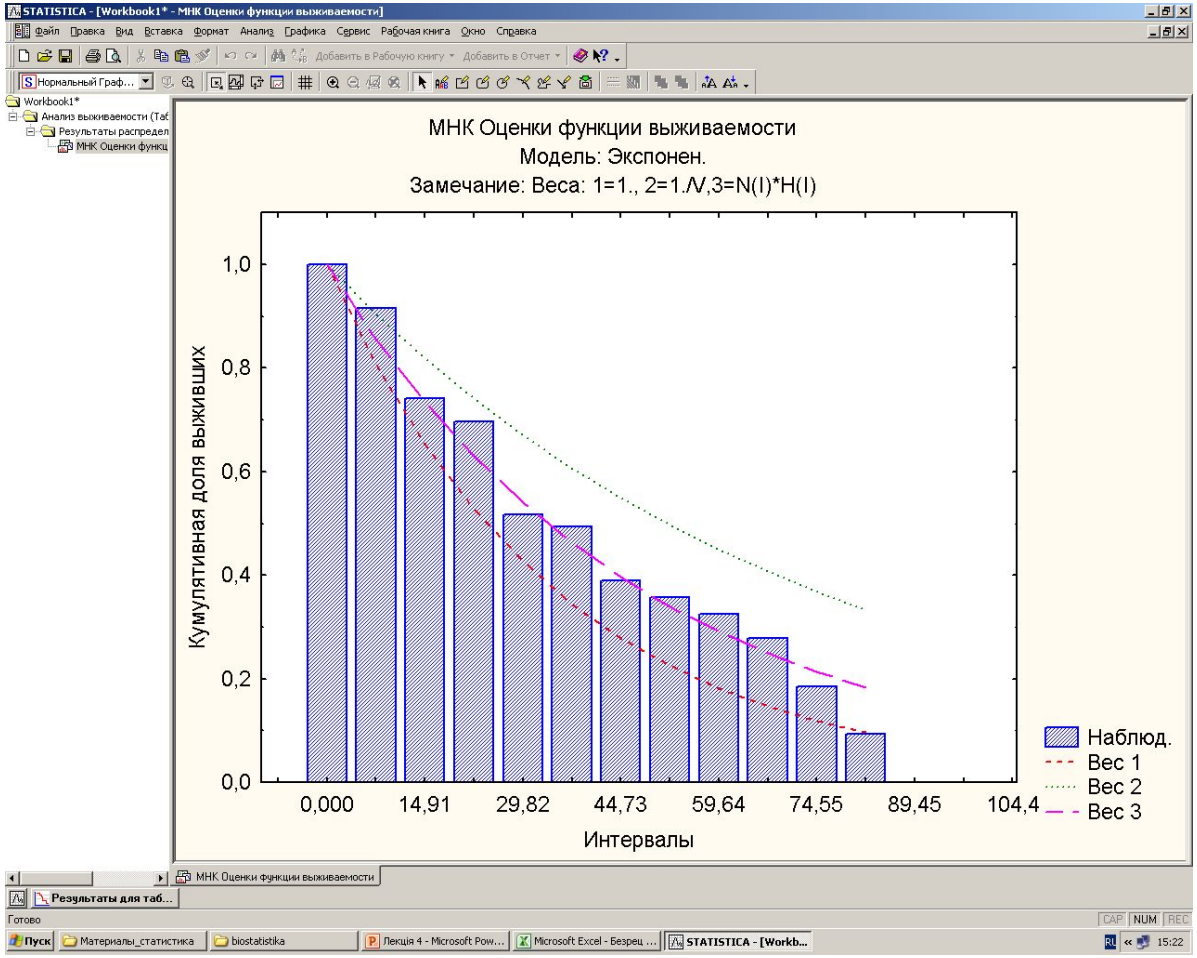
- Аналіз дожиття використовується, коли у дослідника неповні данні. Спостереження, які містять неповну інформацію, називаються неповними або цензурованими. Спостереження до настання досліджуваної події називається повним.

аналіз дожиття

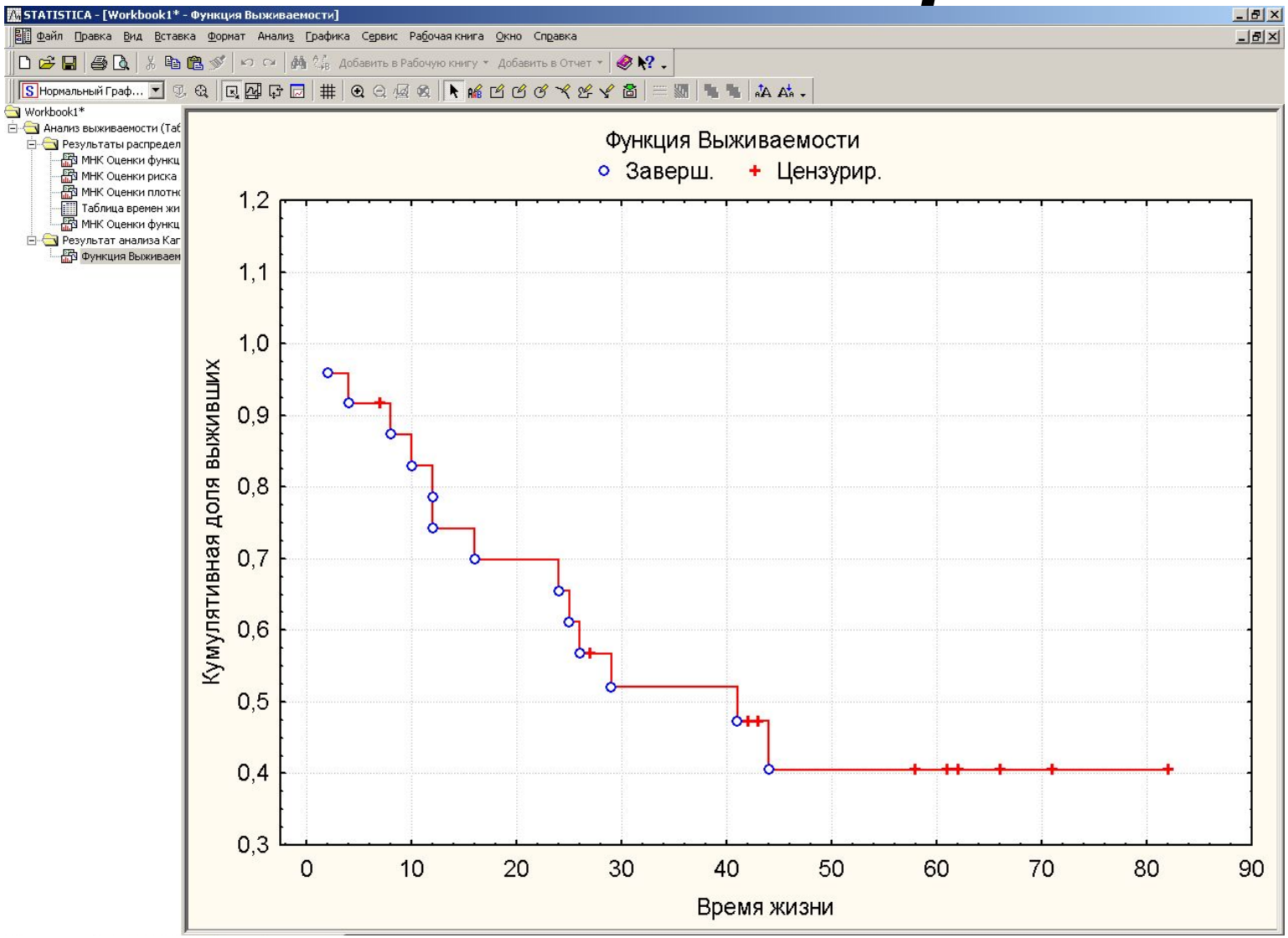
- Анализ – Углубленные методы анализа – Анализ выживаемости



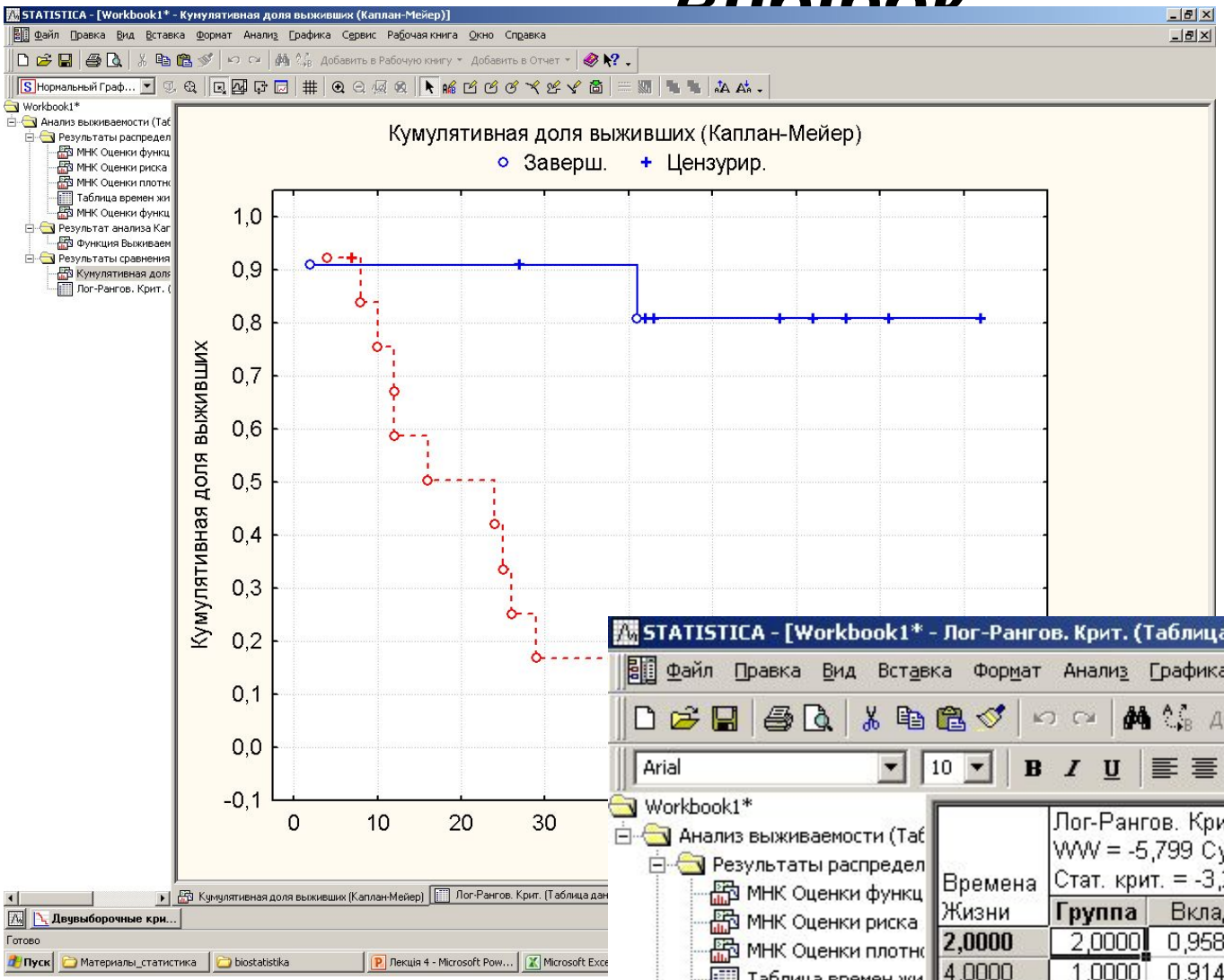
аналіз дожиття: таблиці часу життя

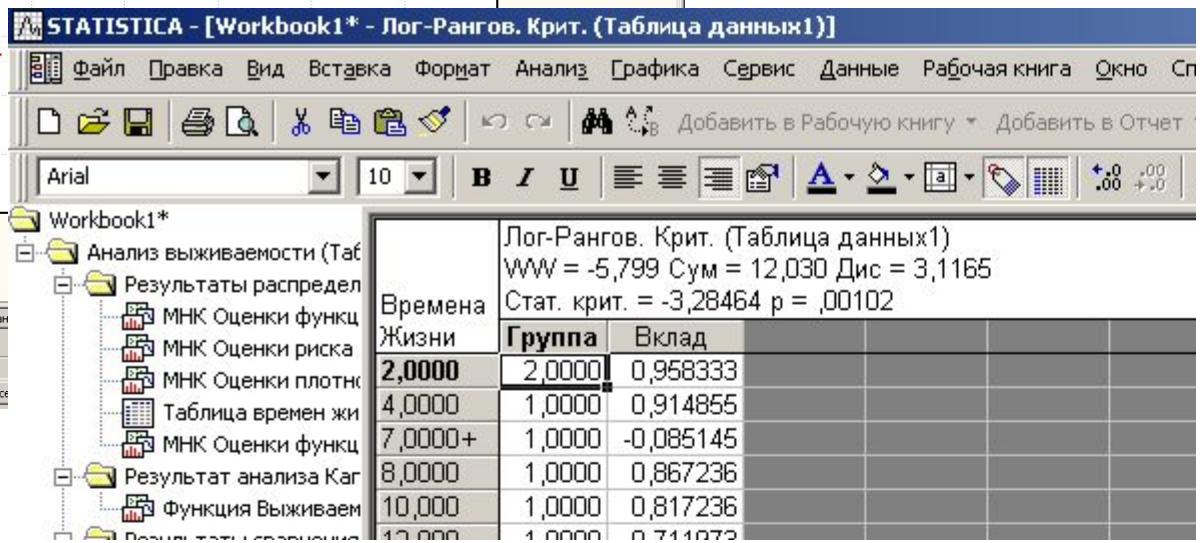


аналіз дожиття: метод Каплана-Майера



аналіз дожиття: порівняння двох вибірок





Модель пропорційних ризиків Кокса

- Модель пропорційних інтенсивностей або ризиків Кокса - найбільш загальна регресійна модель, оскільки вона не пов'язана з якимись припущеннями щодо розподілу часу виживання. Модель може бути записана у наступному вигляді:
- $h\{(t), (z_1, z_2, ..., z_m)\} = h_0(t)*\exp(b_1*z_1 + ... + b_m*z_m)$
- де *h(t,...)* позначає результуючу інтенсивність, при заданих для відповідного спостереження значеннях m коваріат
- (z_1, z_2, ..., z_m) та відповідному часі життя (*t*). Множник $h_0(t)$ називається базовою функцією інтенсивності; вона дорівнює інтенсивності у випадку, коли всі незалежні змінні дорівнюють нулю.

Мета-аналіз

- Мета-аналіз (англ. meta-analysis) — поняття наукової методології. Означає об'єднання результатів декількох досліджень методами статистики для перевірки однієї або кількох взаємопов'язаних наукових гіпотез.

- У мета-анализі використовують або первинні дані оригінальних досліджень, або опубліковані (вторинні) дані, які узагальнюють результати досліджень, присвячених одній проблемі.

Аналіз потужності

- в пакеті Statistica аналіз потужності здійснюється
- *«Анализ» - «Анализ мощности»*
- У модулі «Аналіз потужності» доступні графічні та аналітичні процедури, що дозволяють оцінити потужність і обсяг вибірки для різних процедур статистичного аналізу.