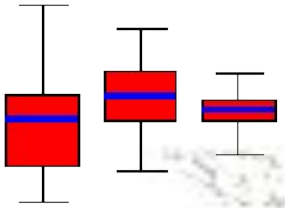
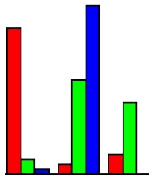
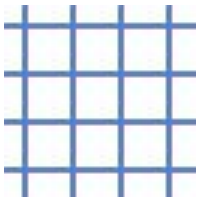
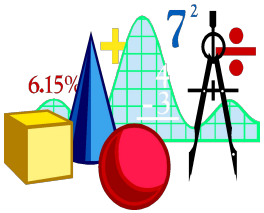




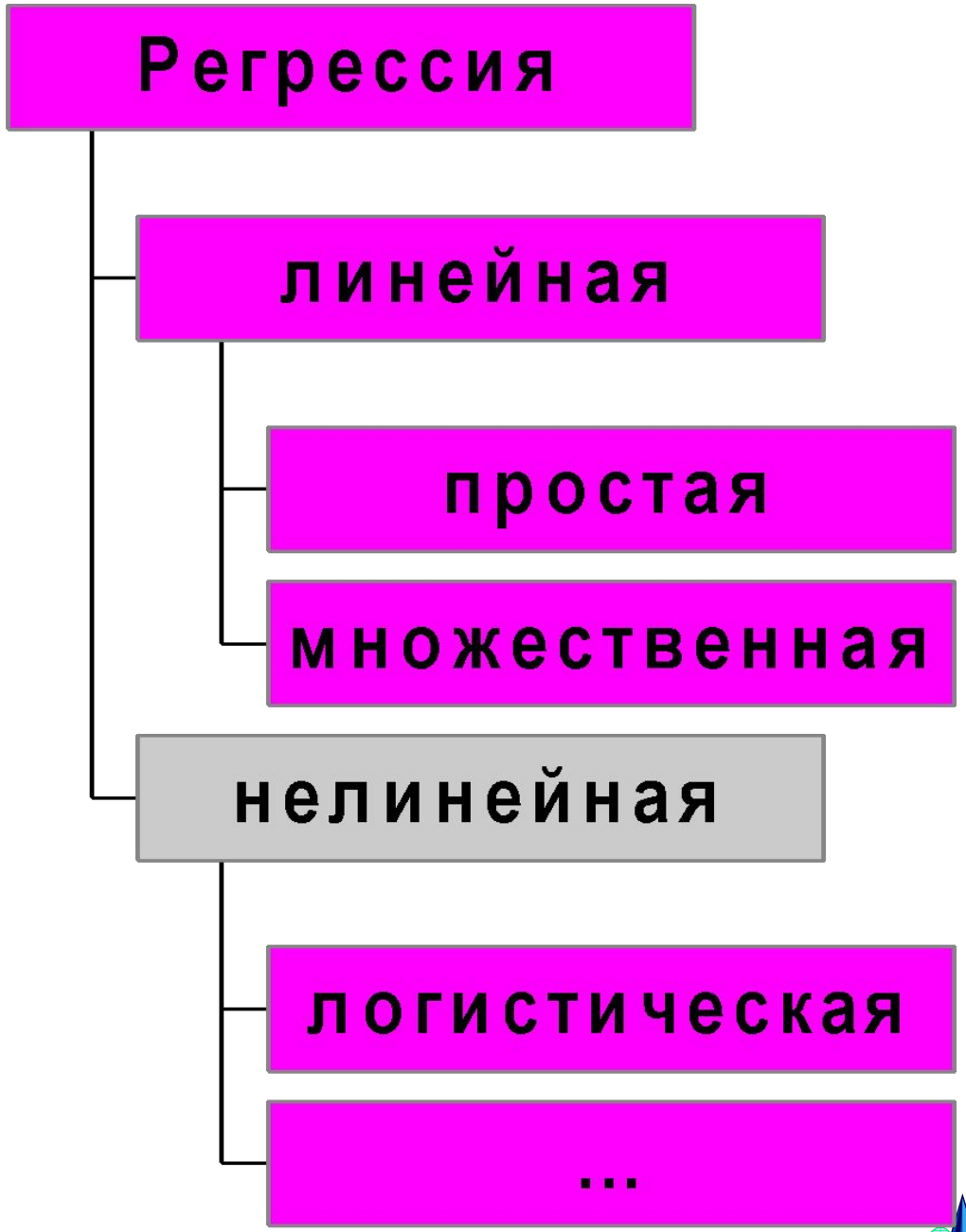
Нелинейная регрессия



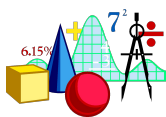
Стат. методы в психологии (Радчикова Н.П.)



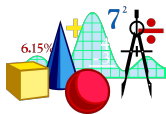
Может быть так, что зависимость между переменными нелинейная. Тогда применяем нелинейную регрессию



Бинарная логистическая регрессия позволяет исследовать зависимость дихотомических зависимых переменных от независимых переменных, имеющих любой вид шкалы

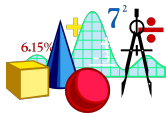


Бинарная логистическая регрессия от дискриминантного анализа отличается тем, что связь между зависимой и независимыми переменными нелинейная



Логистическая регрессия

Мы говорим о некотором событии, которое может произойти или не произойти. В этом случае вероятность наступления события рассматривается в зависимости от значений независимых переменных.



Математическая модель

$$p = \frac{1}{1 + e^{z}}$$

где $z=b_1x_1+b_2x_2+ \ldots+b_nx_n+ b_0$

p – вероятность наступления события, x – независимые переменные

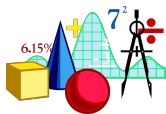
Если p больше 0.5, то можно предположить, что событие произойдет.

Математическая модель

$$p = \frac{1}{1 + e^{z}}$$

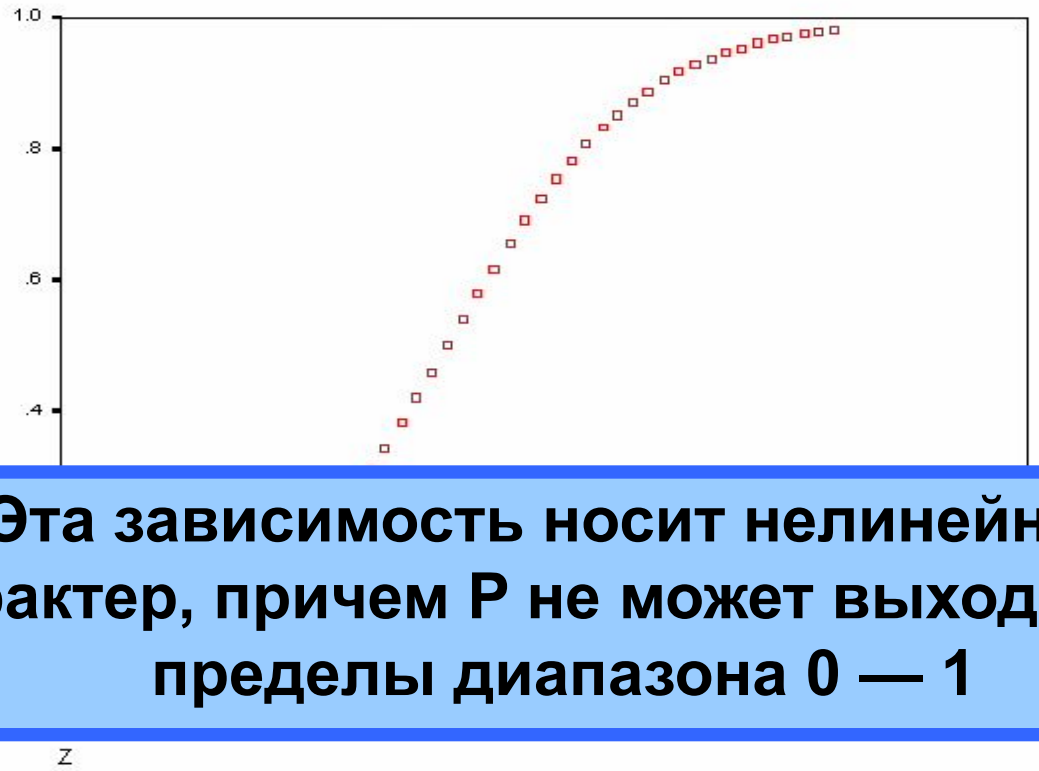
где $z = b_1 x_1 + b_2 x_2 + \ldots + b_n x_n + b_0$

Наша задача, как всегда, - оценить коэффициенты b_i



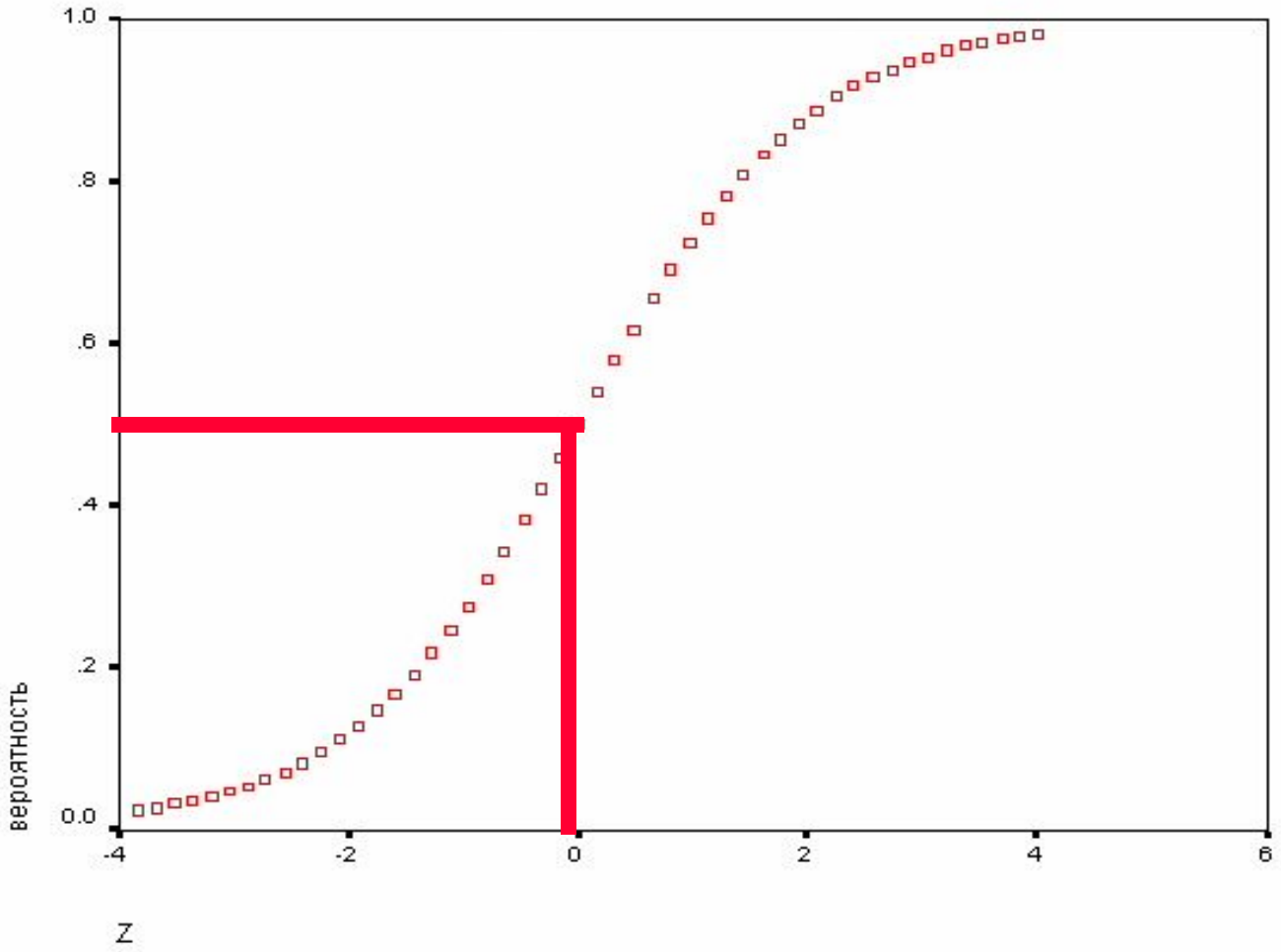
Математическая модель

Зависимость, связывающая вероятность события и величину Z, показана на следующей диаграмме:

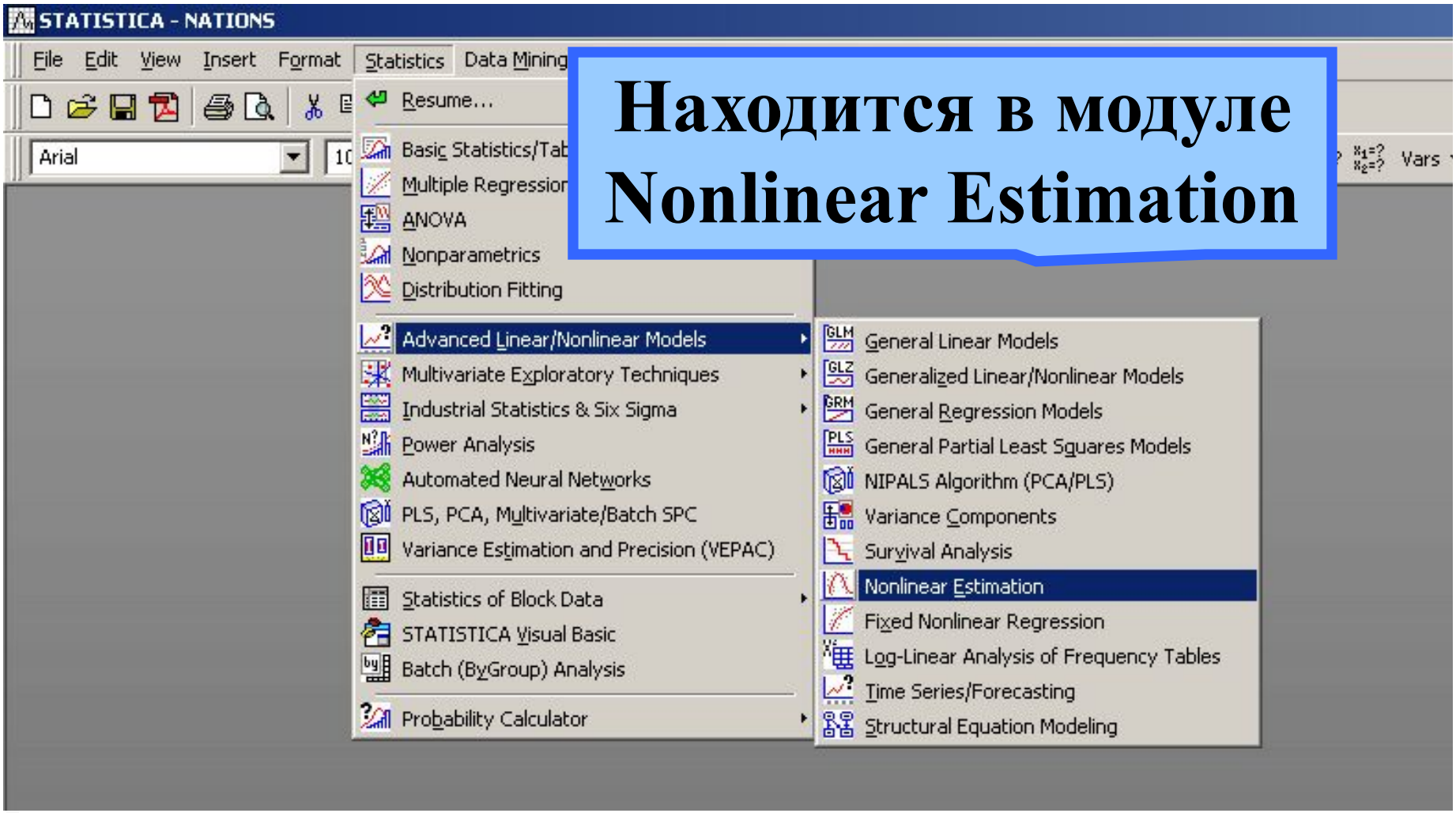


Эта зависимость носит нелинейный характер, причем P не может выходить за пределы диапазона 0 — 1

Математическая модель

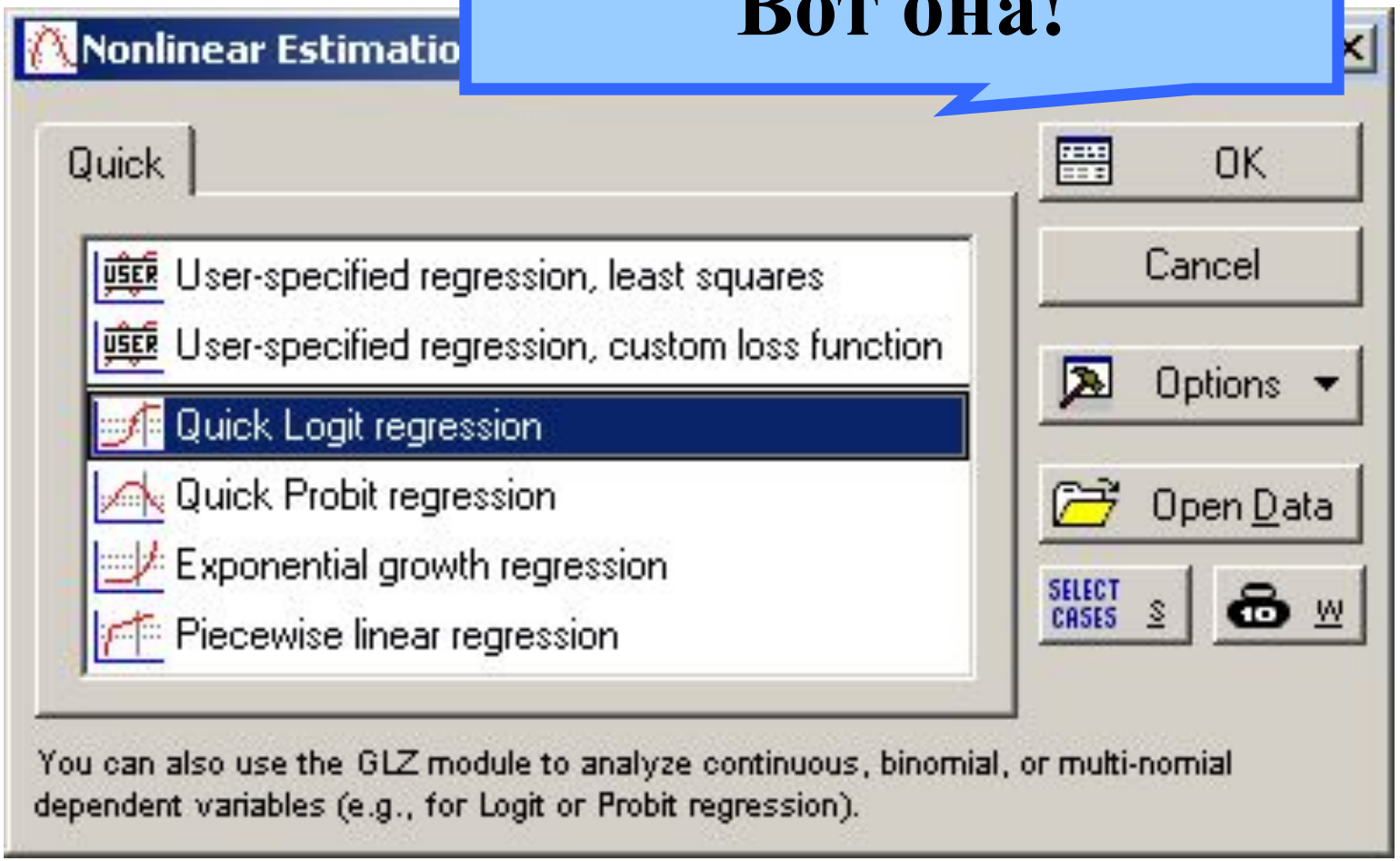


Логистическая регрессия

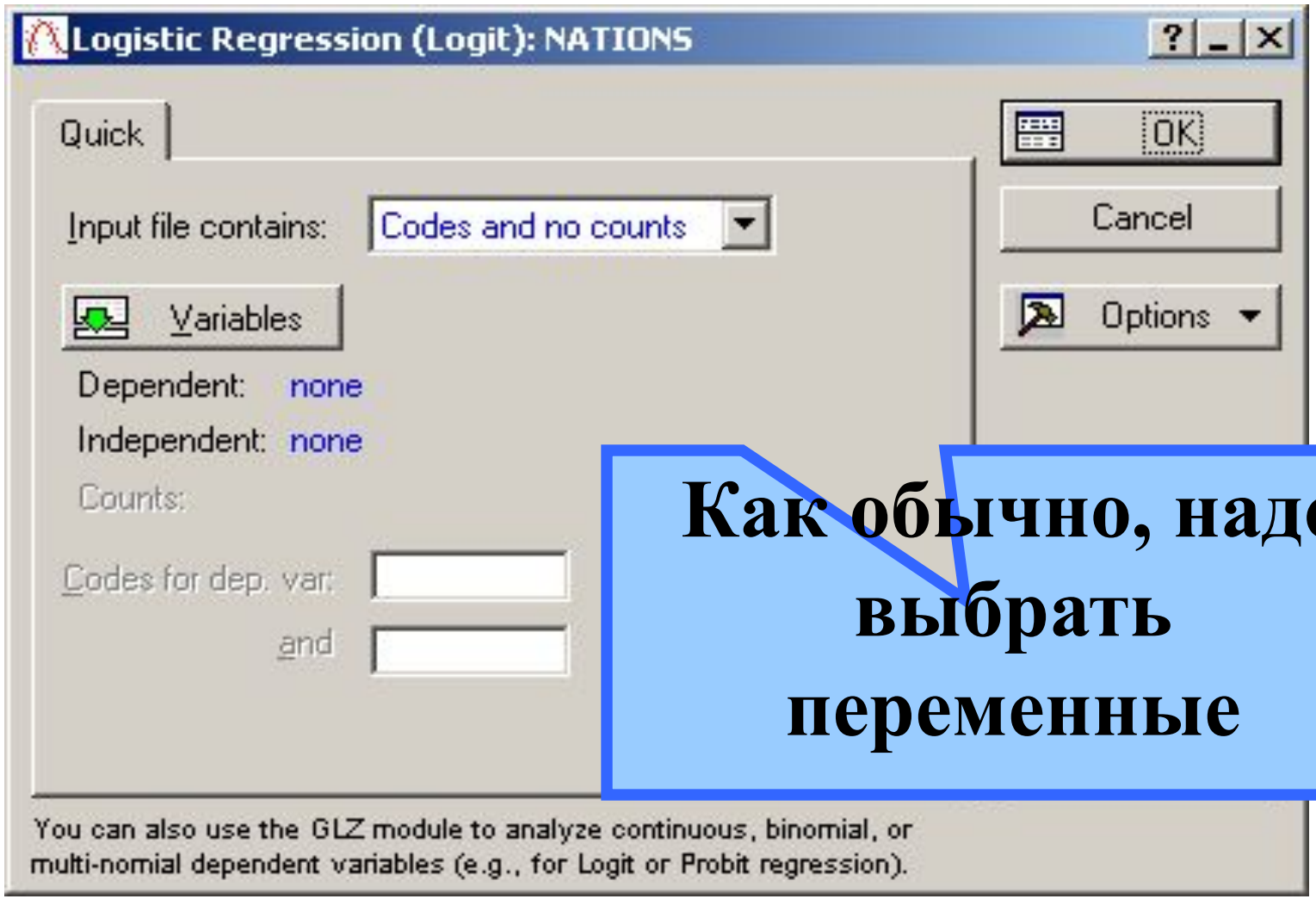


Логистическая регрессия

Вот она!



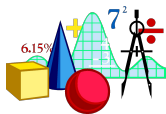
Логистическая регрессия



Как обычно, надо выбрать переменные

Пример

- Рассмотрим пример из медицины (Breast cancer survival.sta)
- Оценим шанс на выживание пациентов разного возраста с опухолью различных размеров (две независимые переменные)



Пример

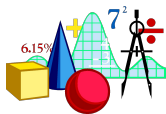
Age – Age (years)

Pathsize - Pathologic Tumor Size (cm)

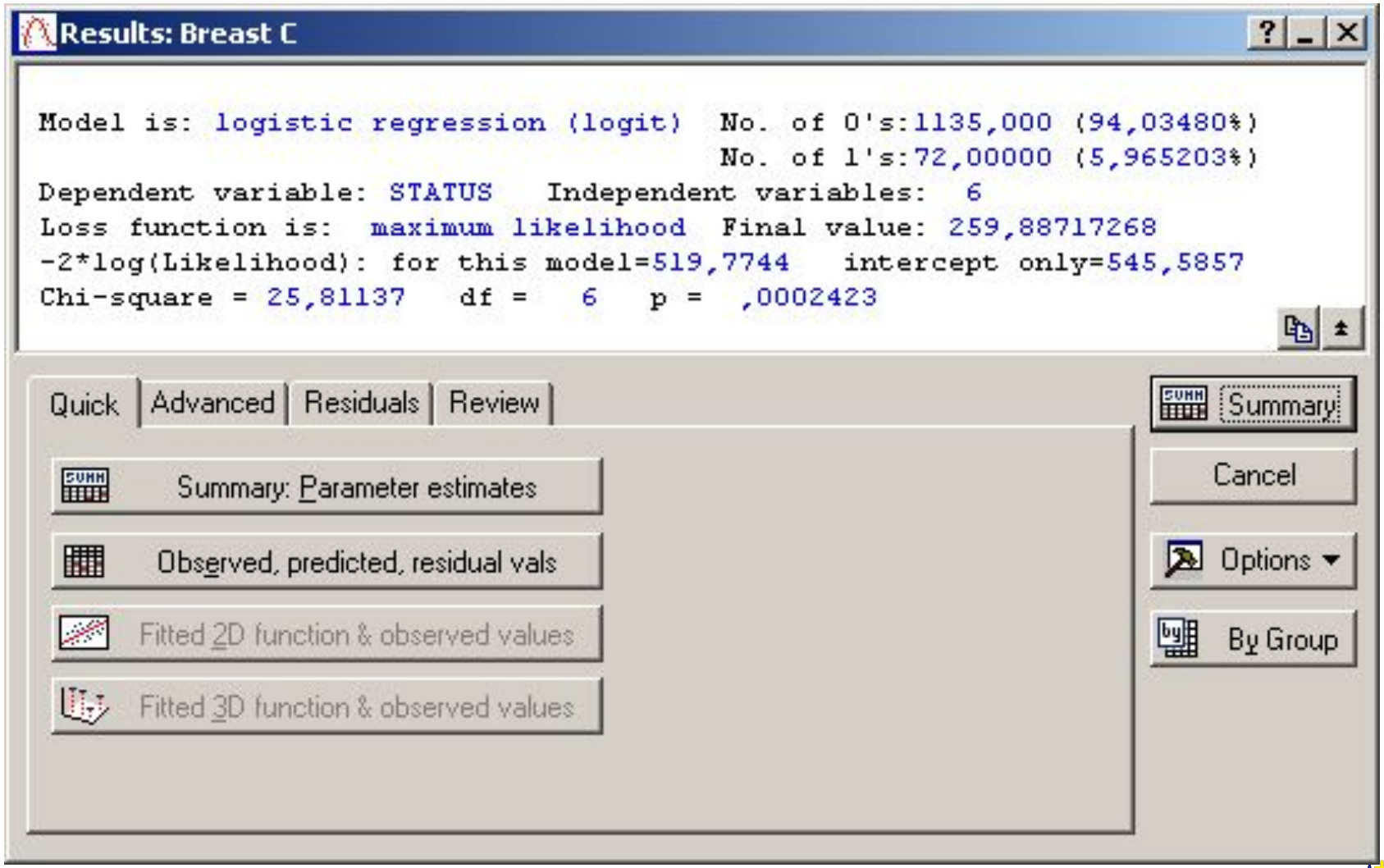
Lnpos - Positive Axillary Lymph Nodes

…

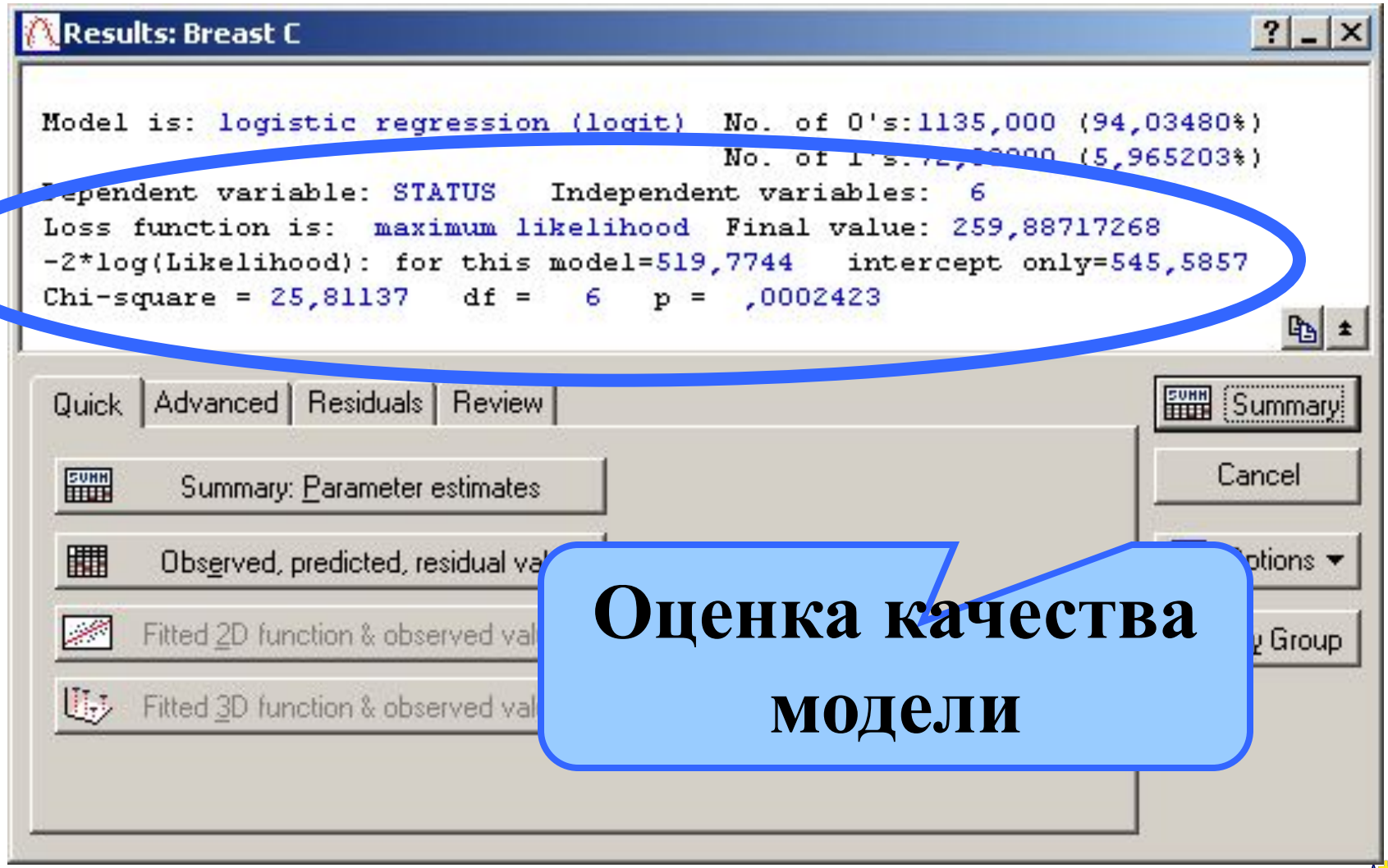
Status – Censored/Died



Результаты


Results: Breast C
Model is: logistic regression (logit) No. of 0's:1135,000 (94,03480%)
No. of 1's:72,00000 (5,965203%)
Dependent variable: STATUS Independent variables: 6
Loss function is: maximum likelihood Final value: 259,88717268
-2*log(Likelihood): for this model=519,7744 intercept only=545,5857
Chi-square = 25,81137 df = 6 p = ,0002423
Quick
Advanced
Residuals
Review
Summary: Parameter estimates
Observed, predicted, residual vals
Fitted 2D function & observed values
Fitted 3D function & observed values
Summary
Cancel
Options
By Group

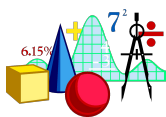
Результаты


Results: Breast C
Model is: logistic regression (logit) No. of 0's:1135,000 (94,03480%)
No. of 1's:72,00000 (5,965203%)
Dependent variable: STATUS Independent variables: 6
Loss function is: maximum likelihood Final value: 259,88717268
-2*log(Likelihood): for this model=519,7744 intercept only=545,5857
Chi-square = 25,81137 df = 6 p = ,0002423
Quick
Advanced
Residuals
Review
Summary
Summary: Parameter estimates
Cancel
Observed, predicted, residual val
Options
Fitted 2D function & observed val
Group
Fitted 3D function & observed val
Оценка качества модели

Качество модели

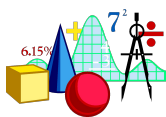
- Качество приближения регрессионной модели оценивается при помощи функции подобия. Мерой правдоподобия служит отрицательное удвоение значения логарифма этой функции - -2LL.

- В качестве начального значения для -2LL принимается значение, которое получается для регрессионной модели, содержащей только константу.

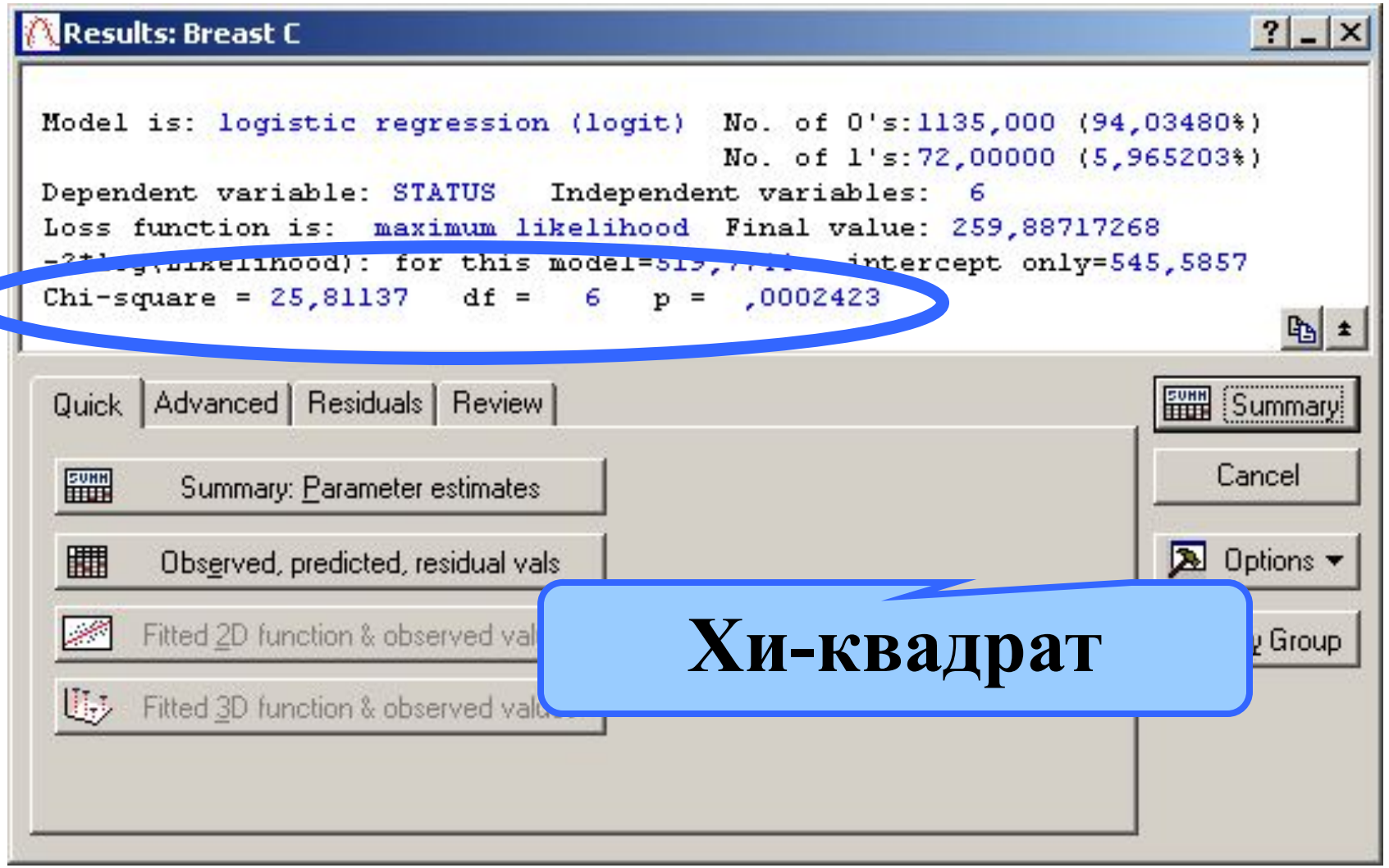


Качество модели

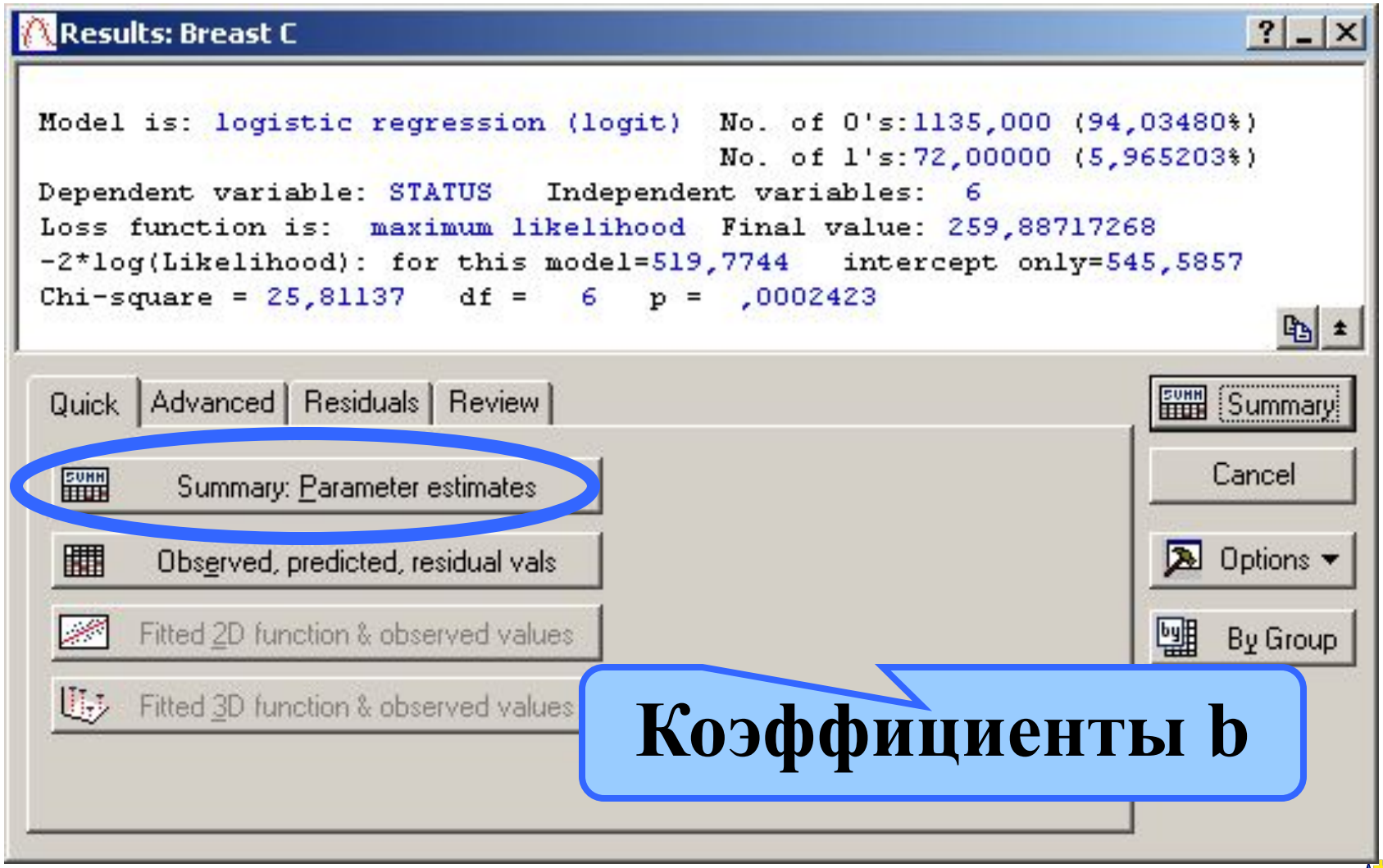
Затем в модель добавляют переменные согласно выбранному методу и вычисляют разность (улучшение качества модели). Разность обозначают как хи-квадрат и вычисляют ее значимость.



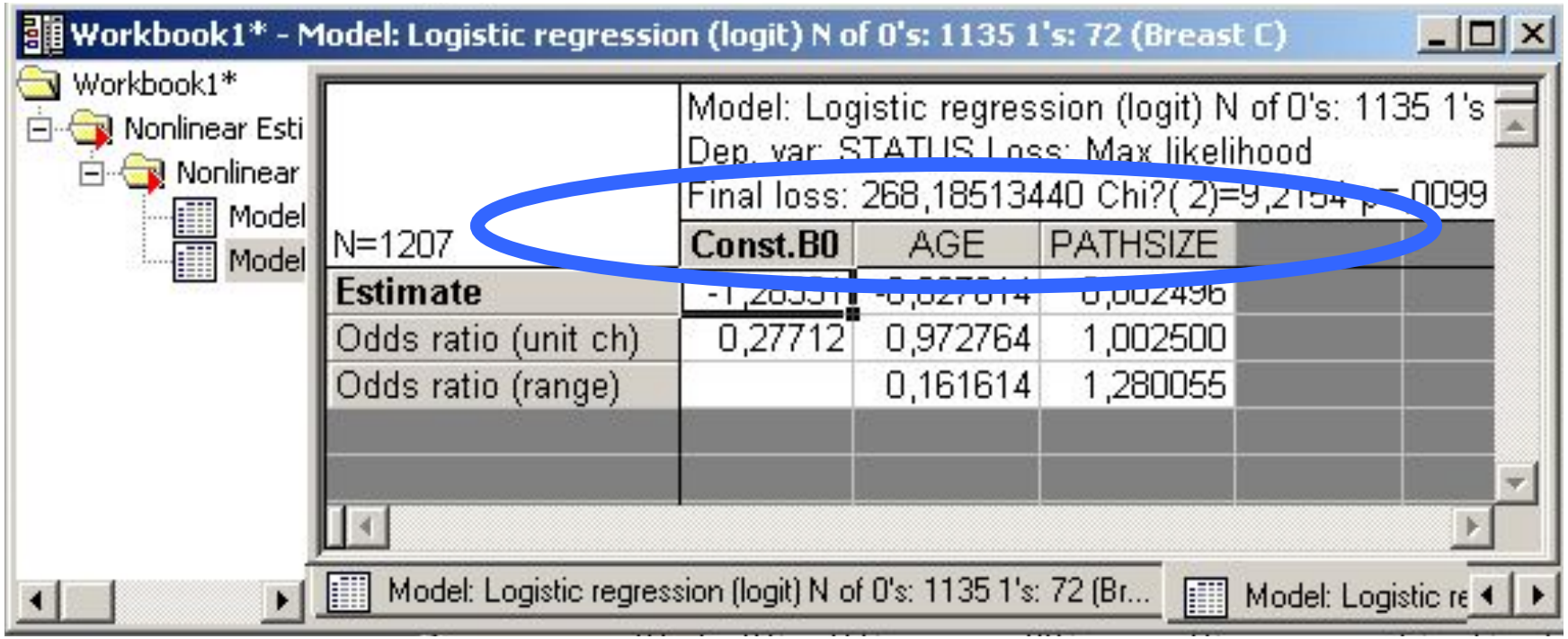
Качество модели



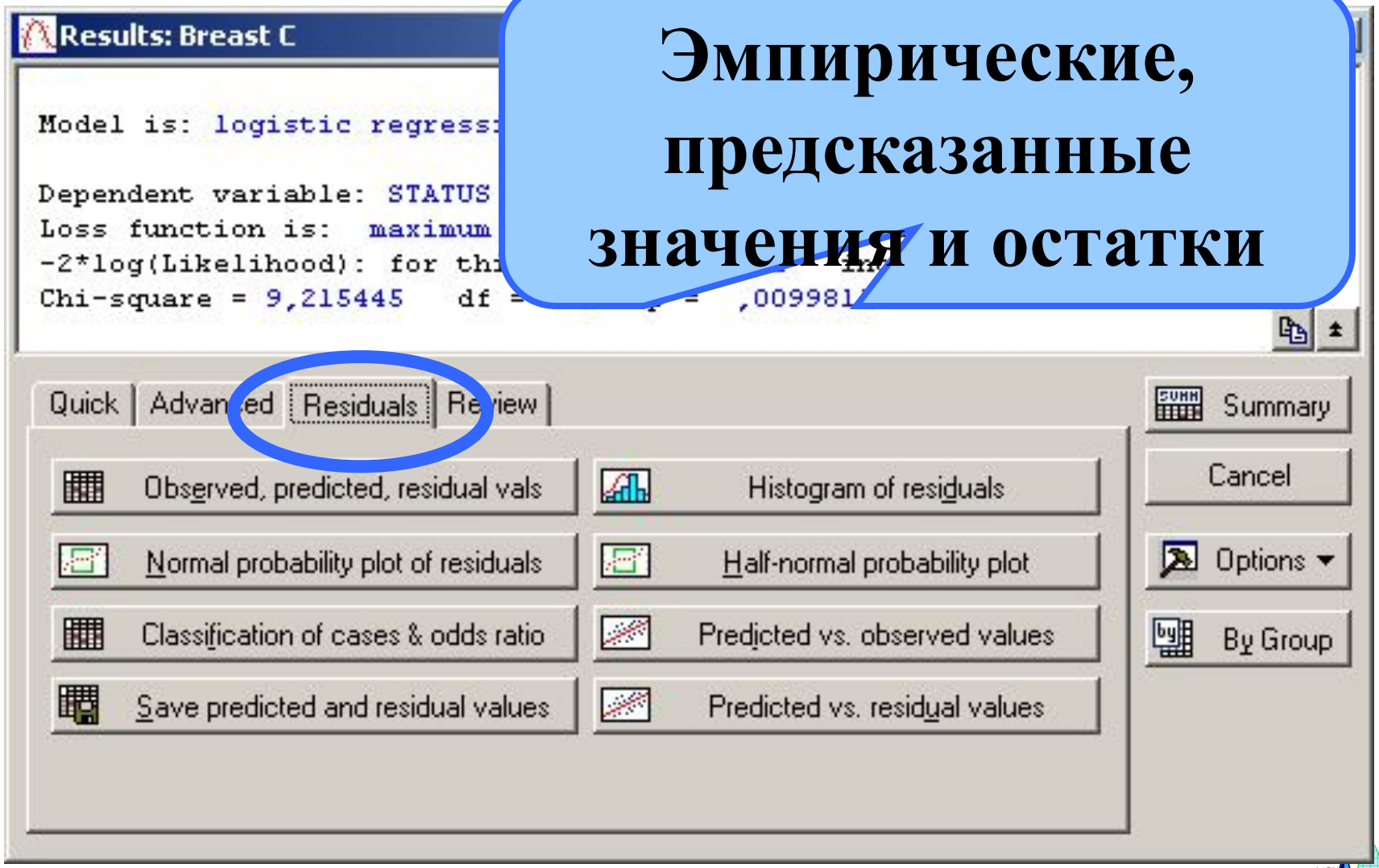
Результаты


Results: Breast C
Model is: logistic regression (logit) No. of 0's:1135,000 (94,03480%)
No. of 1's:72,00000 (5,965203%)
Dependent variable: STATUS Independent variables: 6
Loss function is: maximum likelihood Final value: 259,88717268
-2*log(Likelihood): for this model=519,7744 intercept only=545,5857
Chi-square = 25,81137 df = 6 p = ,0002423
Quick
Advanced
Residuals
Review
Summary
Summary: Parameter estimates
Cancel
Observed, predicted, residual vals
Options
Fitted 2D function & observed values
By Group
Fitted 3D function & observed values
Коэффициенты b

Регрессионные коэффициенты

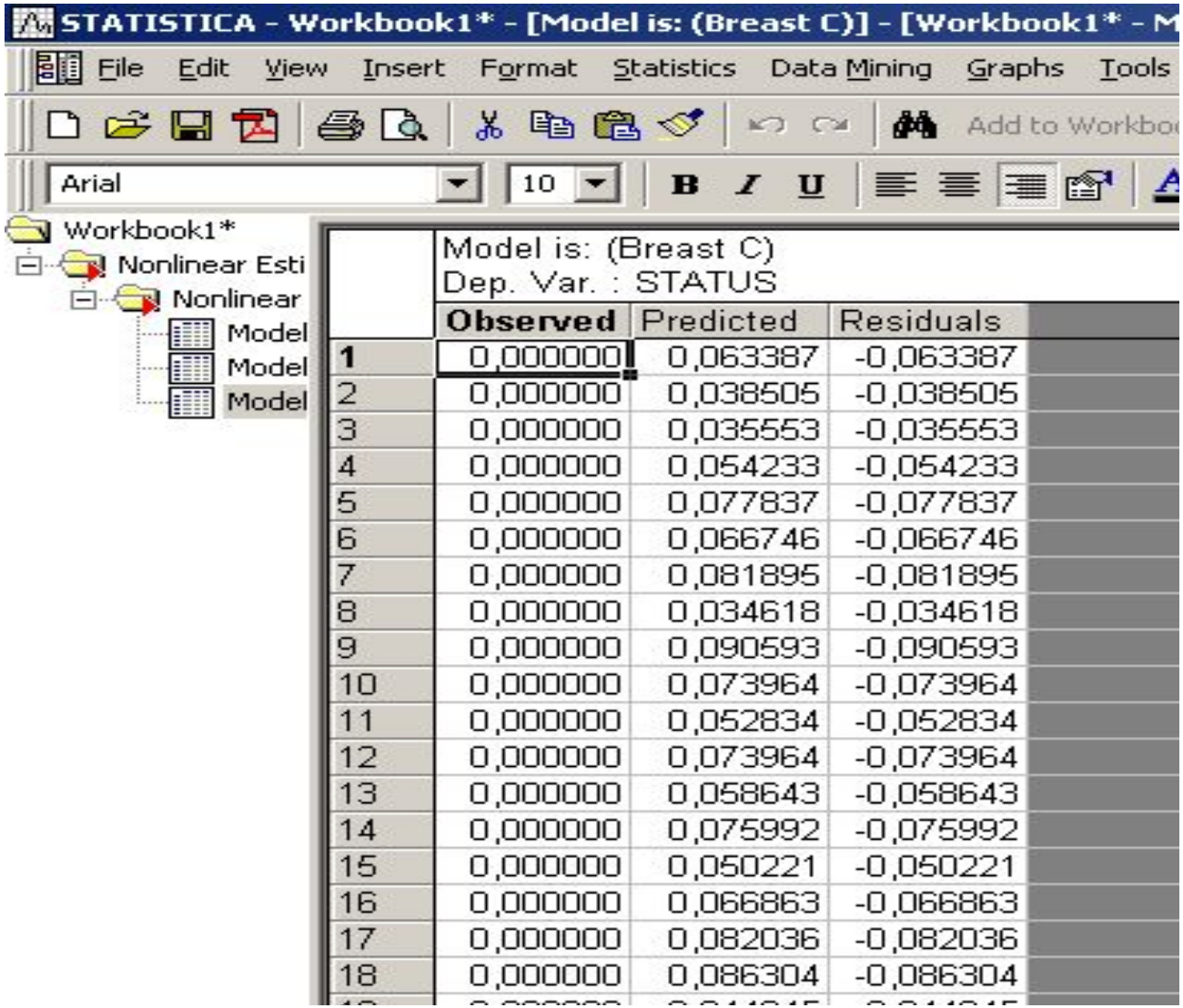


Результаты



Эмпирические, предсказанные значения и остатки

Результаты

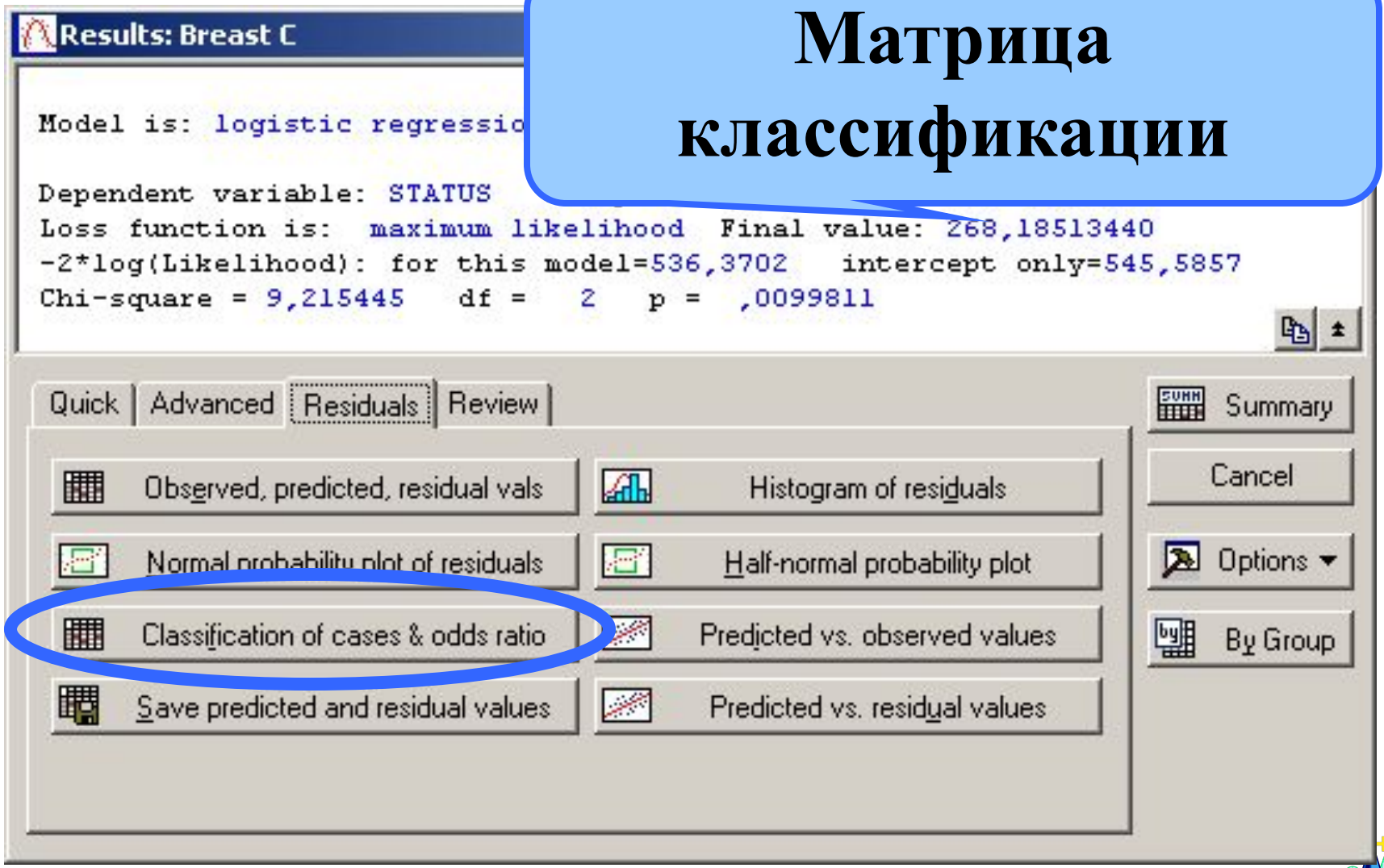


Model is: (Breast C)
Dep. Var. : STATUS

	Observed	Predicted	Residuals
1	0,000000	0,063387	-0,063387
2	0,000000	0,038505	-0,038505
3	0,000000	0,035553	-0,035553
4	0,000000	0,054233	-0,054233
5	0,000000	0,077837	-0,077837
6	0,000000	0,066746	-0,066746
7	0,000000	0,081895	-0,081895
8	0,000000	0,034618	-0,034618
9	0,000000	0,090593	-0,090593
10	0,000000	0,073964	-0,073964
11	0,000000	0,052834	-0,052834
12	0,000000	0,073964	-0,073964
13	0,000000	0,058643	-0,058643
14	0,000000	0,075992	-0,075992
15	0,000000	0,050221	-0,050221
16	0,000000	0,066863	-0,066863
17	0,000000	0,082036	-0,082036
18	0,000000	0,086304	-0,086304

Результаты

Матрица классификации



Результаты

Classification of Cases (breast c.sta)

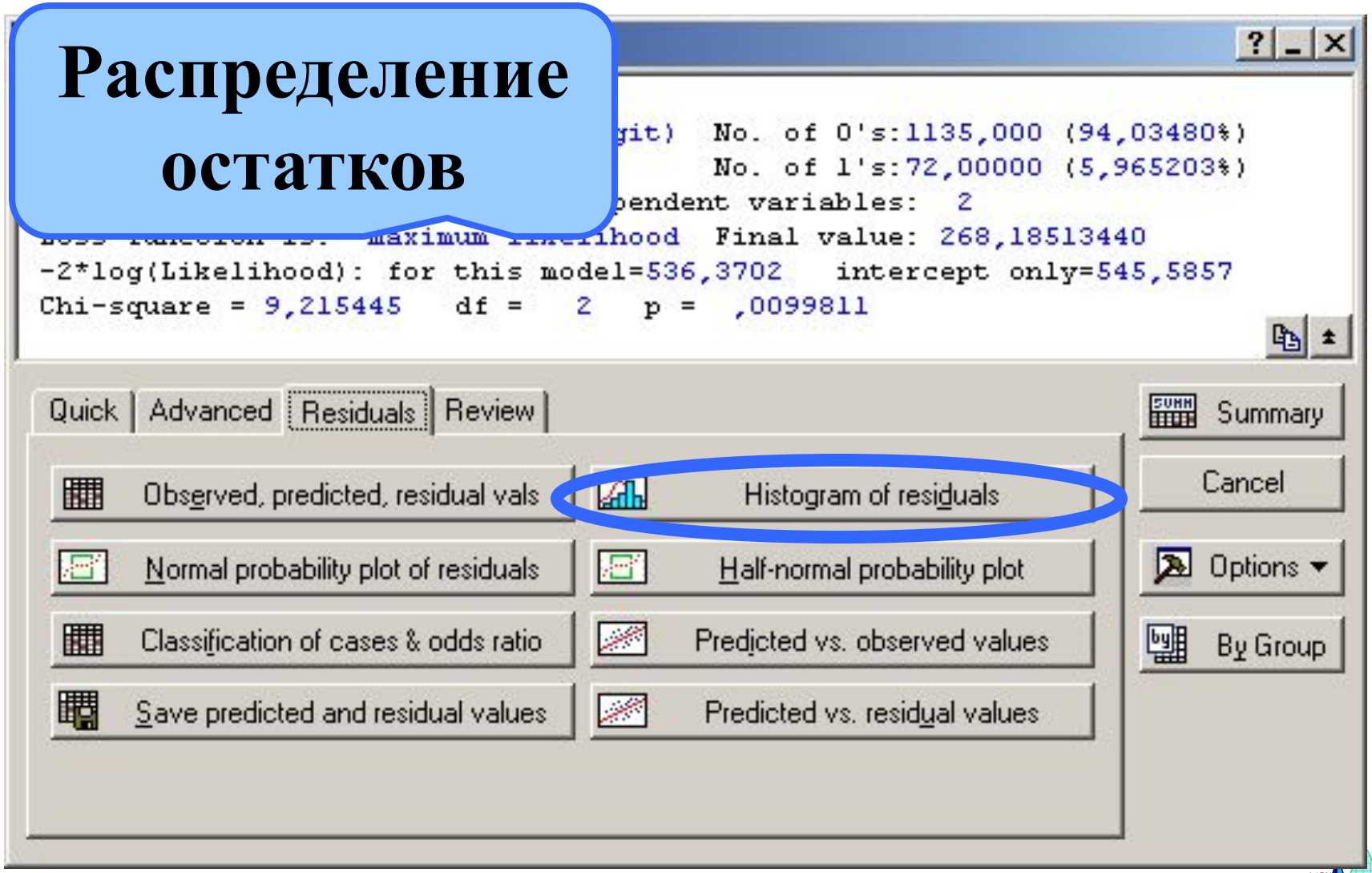
NONLIN. ESTIMAT.

Odds ratio: -----

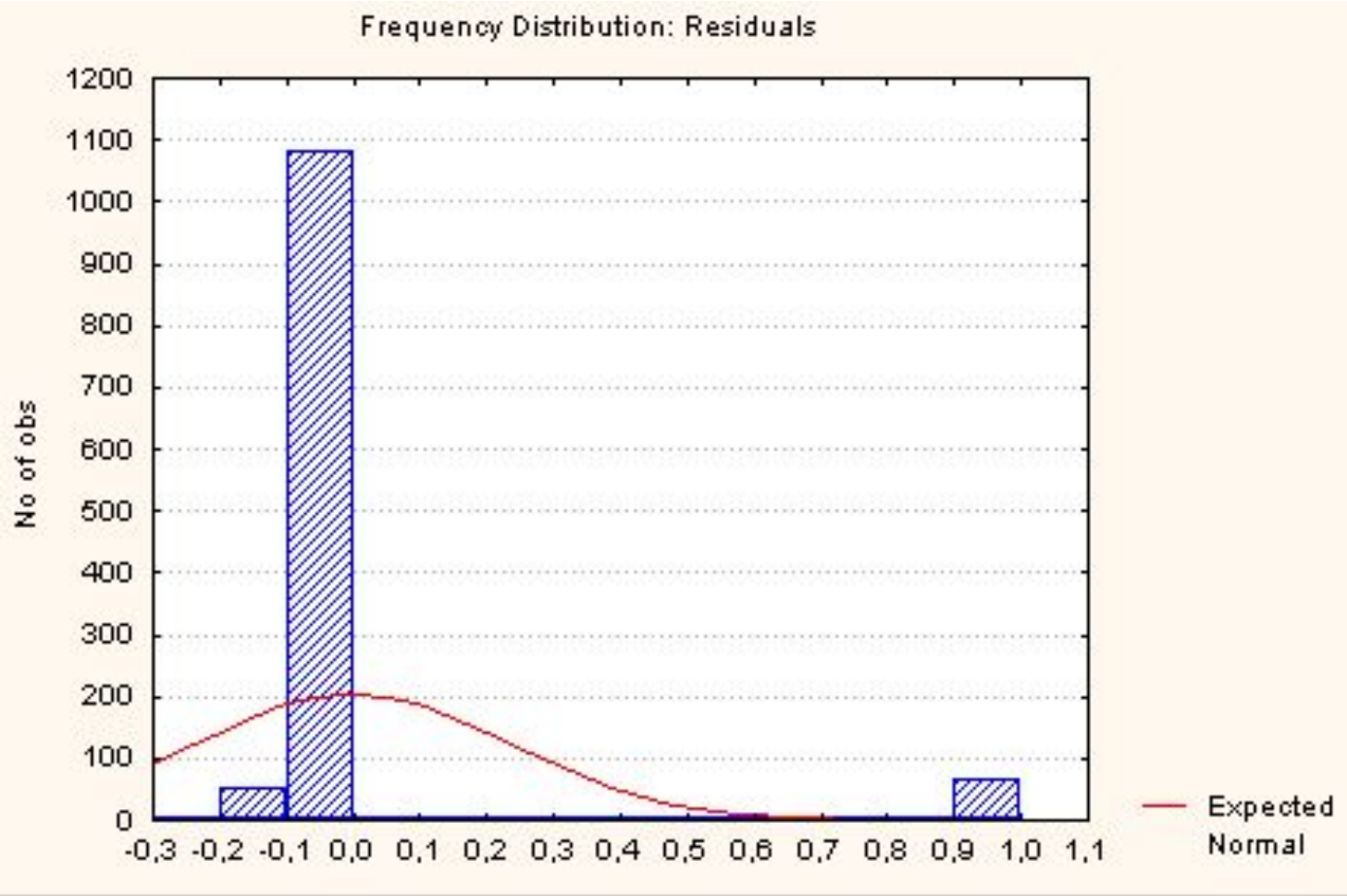
Observed	Pred. 0	Pred. 1	Percent Correct
0	1135	0	100,0000
1	72	0	0,0000

Результаты

Распределение остатков

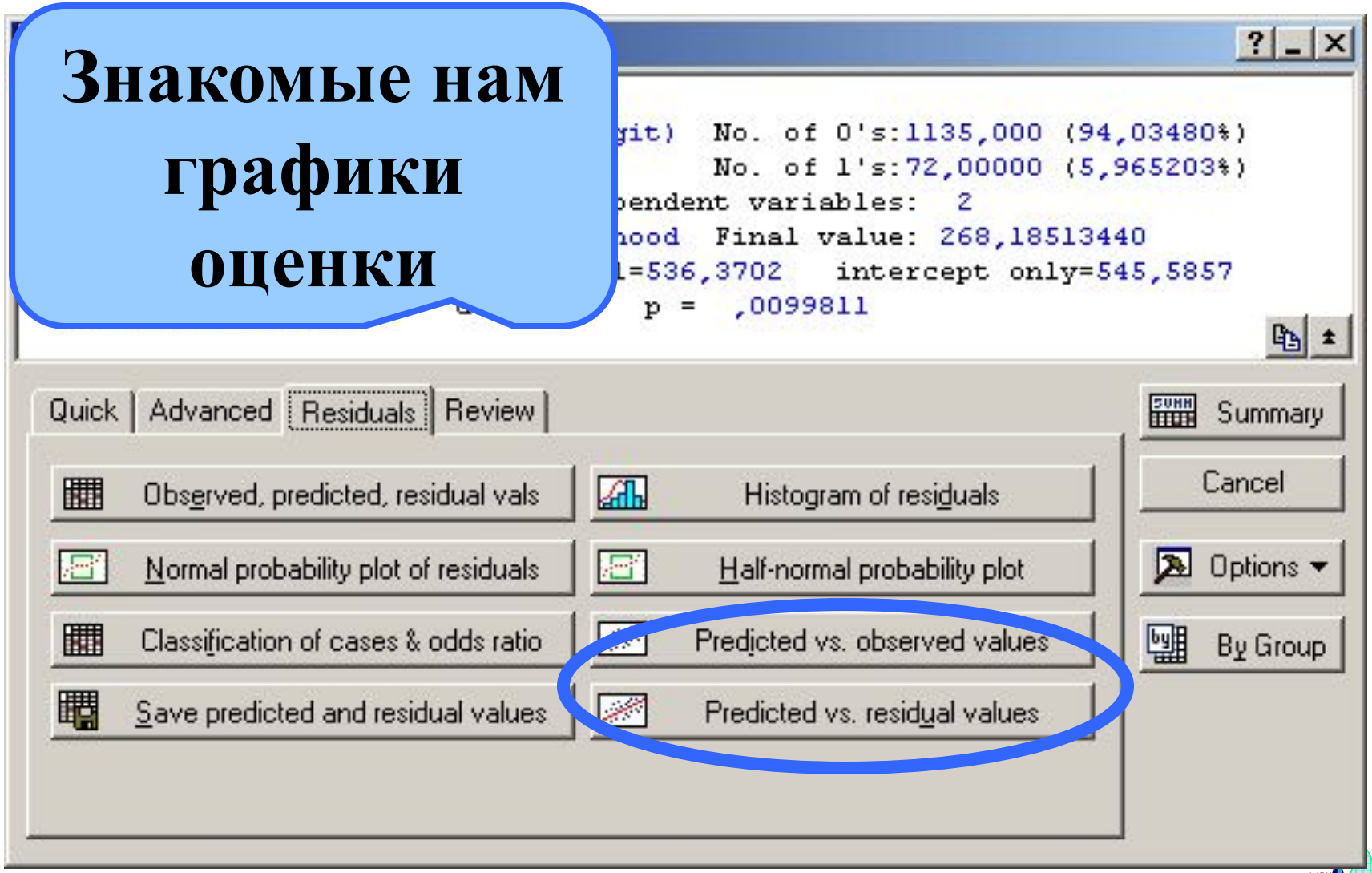


Результаты

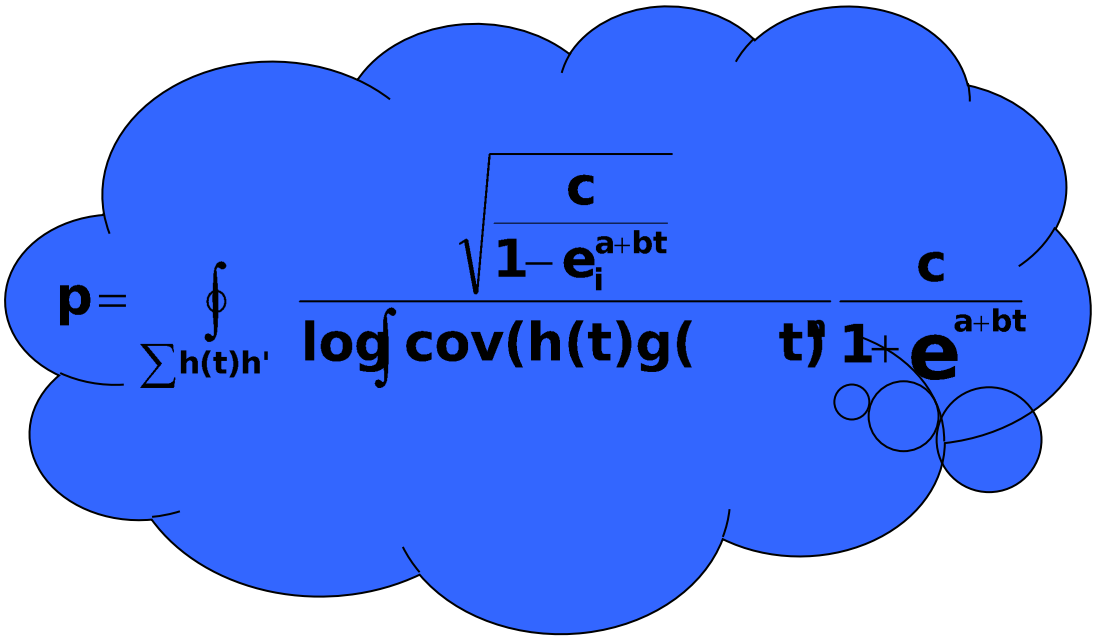

Frequency Distribution: Residuals
No of obs
1200
1100
1000
900
800
700
600
500
400
300
200
100
0
-0,3 -0,2 -0,1 0,0 0,1 0,2 0,3 0,4 0,5 0,6 0,7 0,8 0,9 1,0 1,1
Expected
Normal

Результаты

Знакомые нам графики оценки

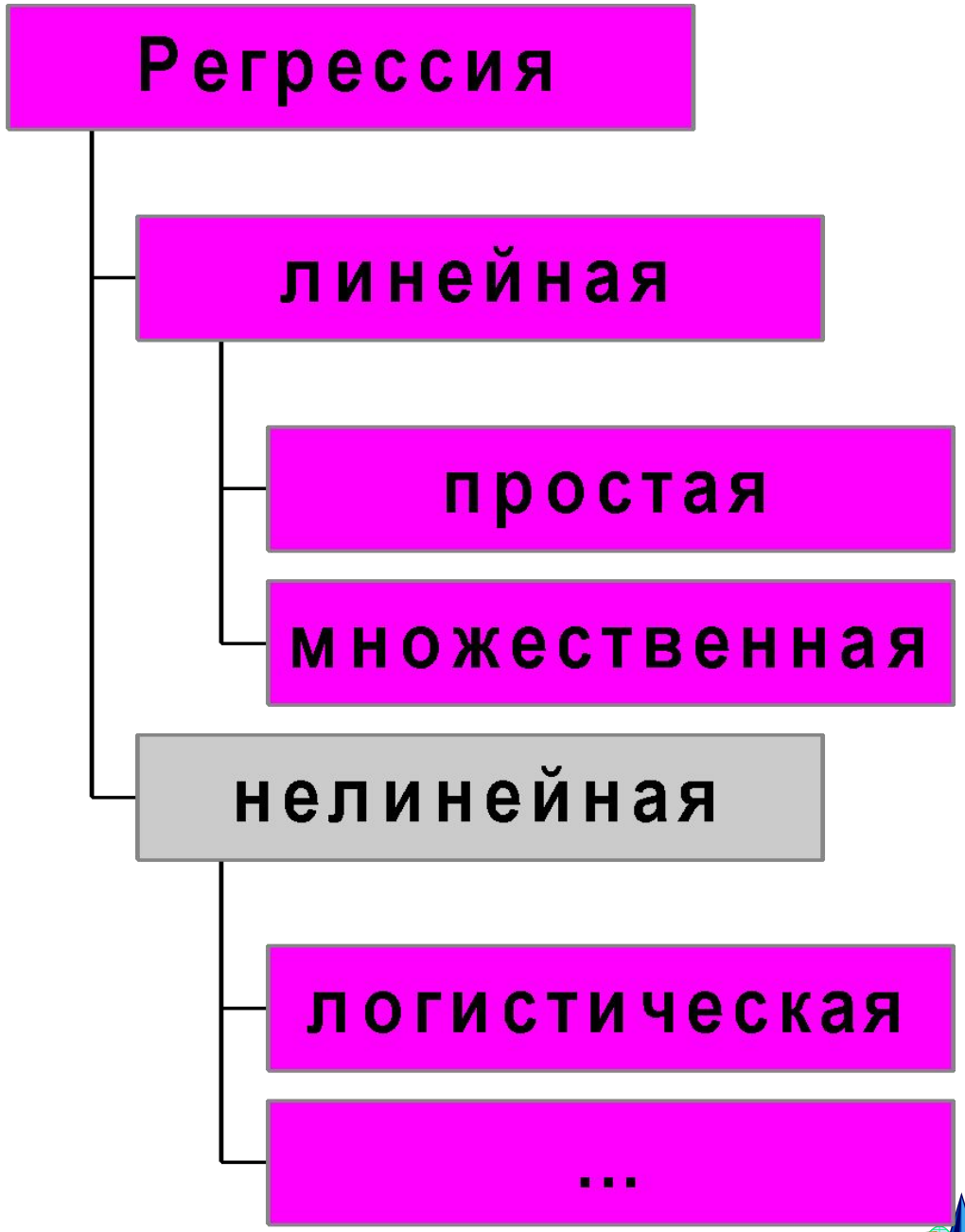


А если у меня такая зависимость, какую я сам придумал ?!



Оценка на экзамене и мотивация так прямо не связаны …

Тогда применяем нелинейную регрессию, а зависимость может быть задана самим пользователем



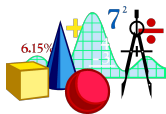
Пример.

Рост населения в США с 1790 по 1960 гг по декадам:

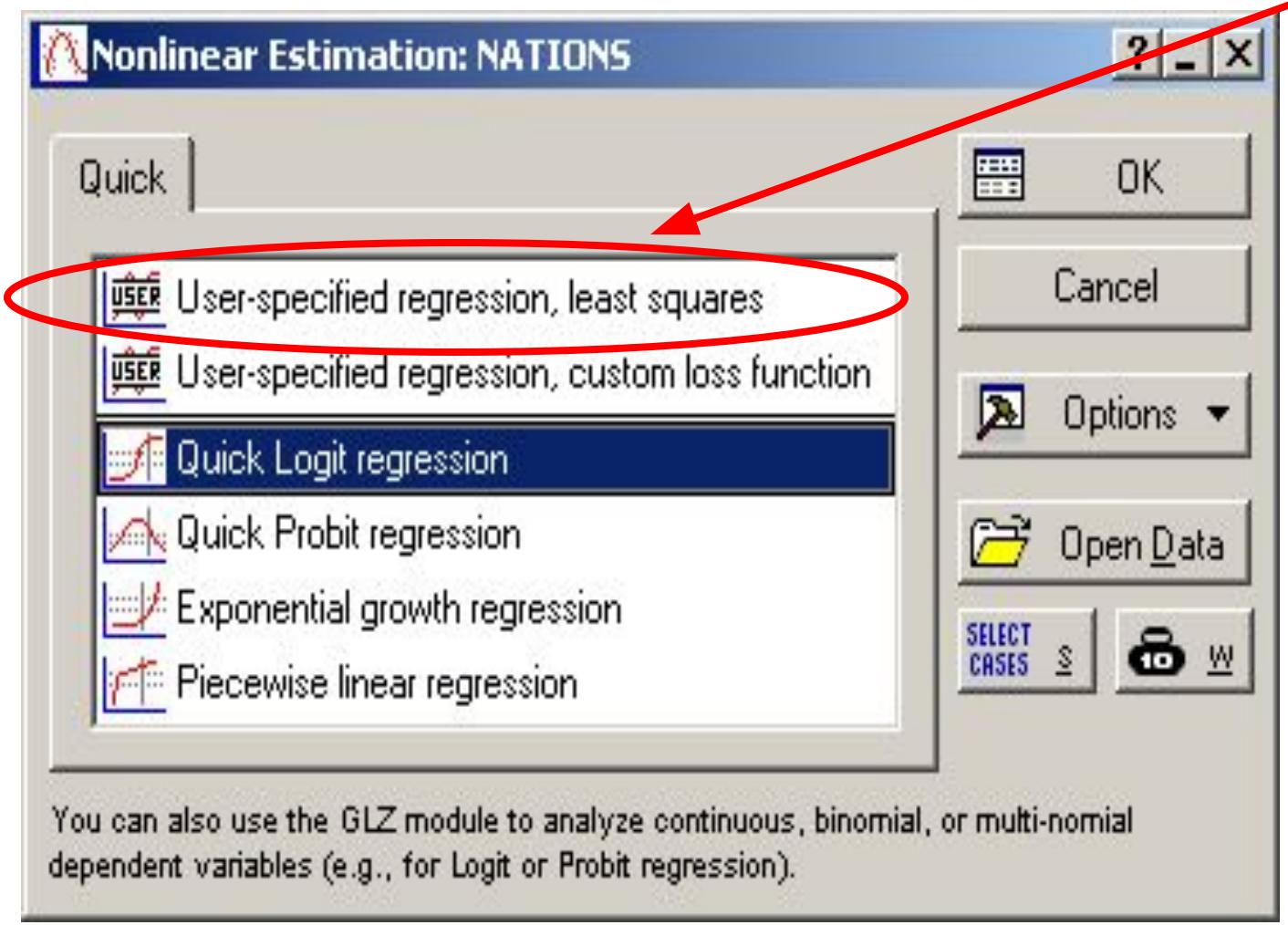
Видно, что зависимость тут скорее не линейная, а экспоненциальная. Демографы знают, что лучше всего зависимость роста населения от времени описывается функцией

$$population = \frac{c}{1 + e^{a+bt}}$$

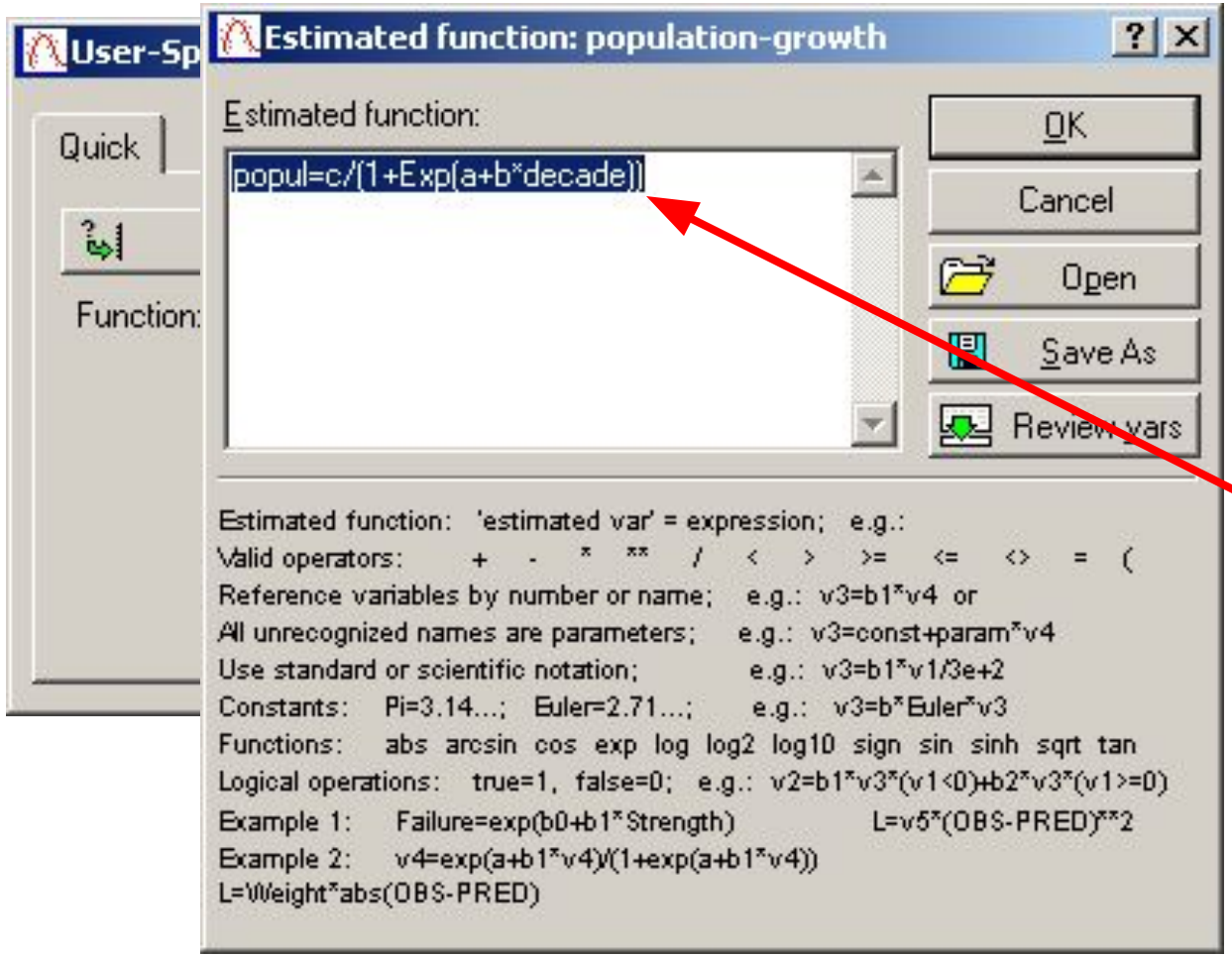
Очевидно, что нашей задачей является определение трех коэффициентов - a, b и c.



Для построения уравнений нелинейной регрессии служит модуль Nonlinear Estimation

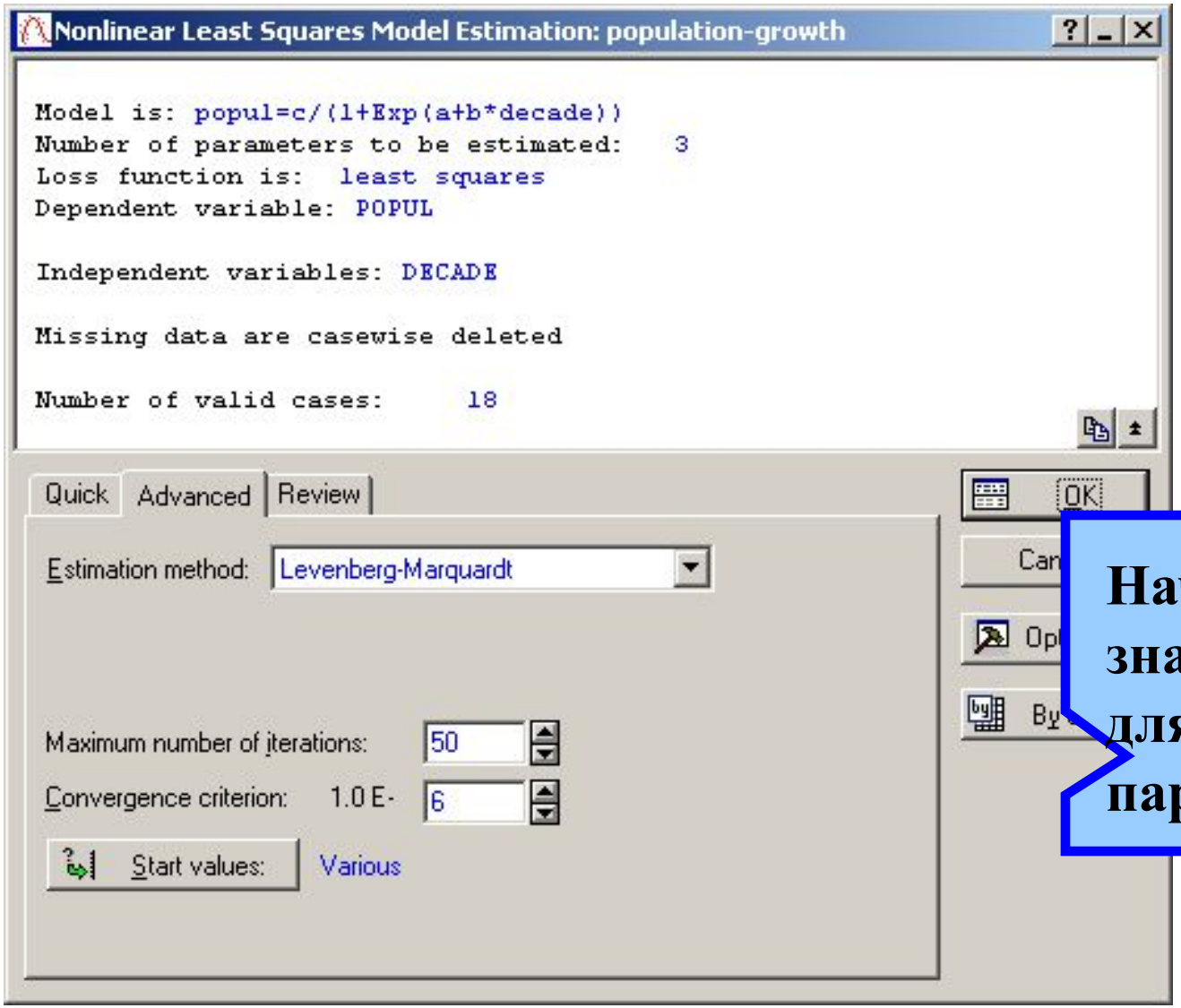


Для построения уравнений нелинейной регрессии служит модуль Nonlinear Estimation



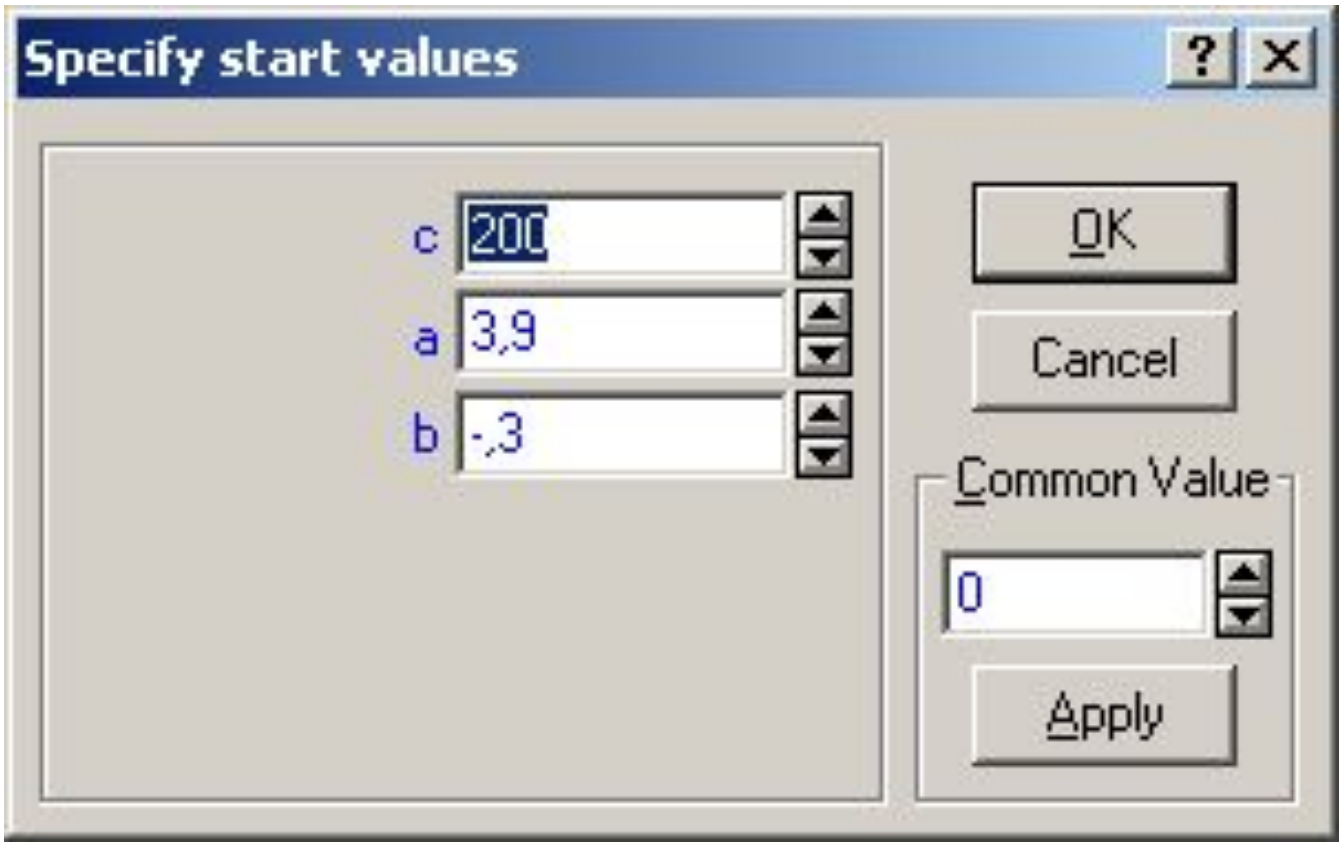
Тут набираем формулу, которая, по нашему мнению, хорошо описывает полученную зависимость

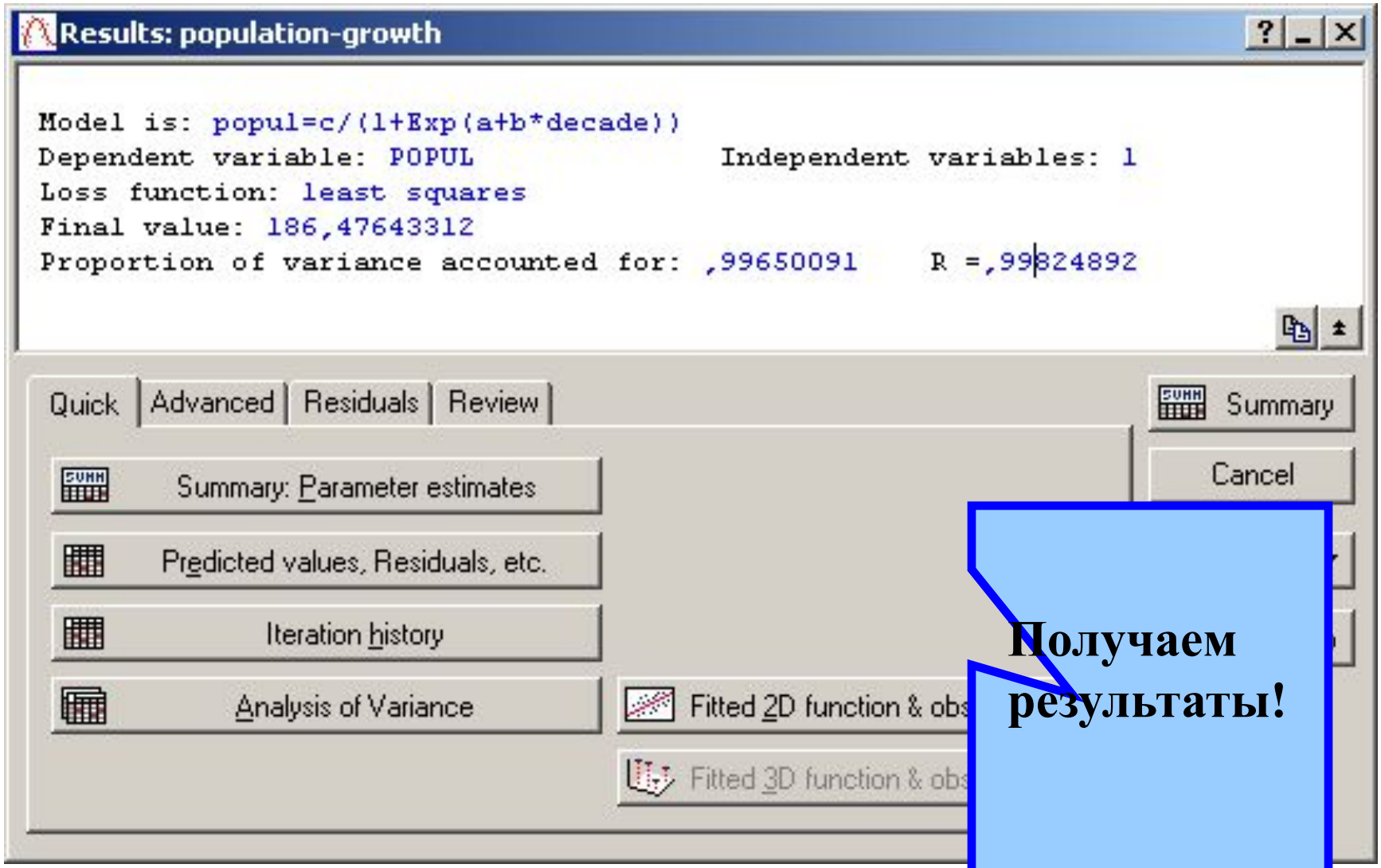
Маленькие (?) хитрости



Начальные значения для параметров

Маленькие (?) хитрости




Results: population-growth
Model is: popul=c/(1+Exp(a+b*decade))
Dependent variable: POPUL
Independent variables: 1
Loss function: least squares
Final value: 186,47643312
Proportion of variance accounted for: ,99650091
R =,99824892
Quick
Advanced
Residuals
Review
Summary
Cancel
Summary: Parameter estimates
Predicted values, Residuals, etc.
Iteration history
Analysis of Variance
Fitted 2D function & obs
Fitted 3D function & obs
Получаем результаты!

Оценка параметров

Model: POPUL=c/(1+exp(a+(b*DECADE))) (new.sta)

Continue..

Dep. var: POPUL Loss: (OBS-PRED)**2
Final loss: 186,47643305 R=,99825 Variance explained: 99,650%

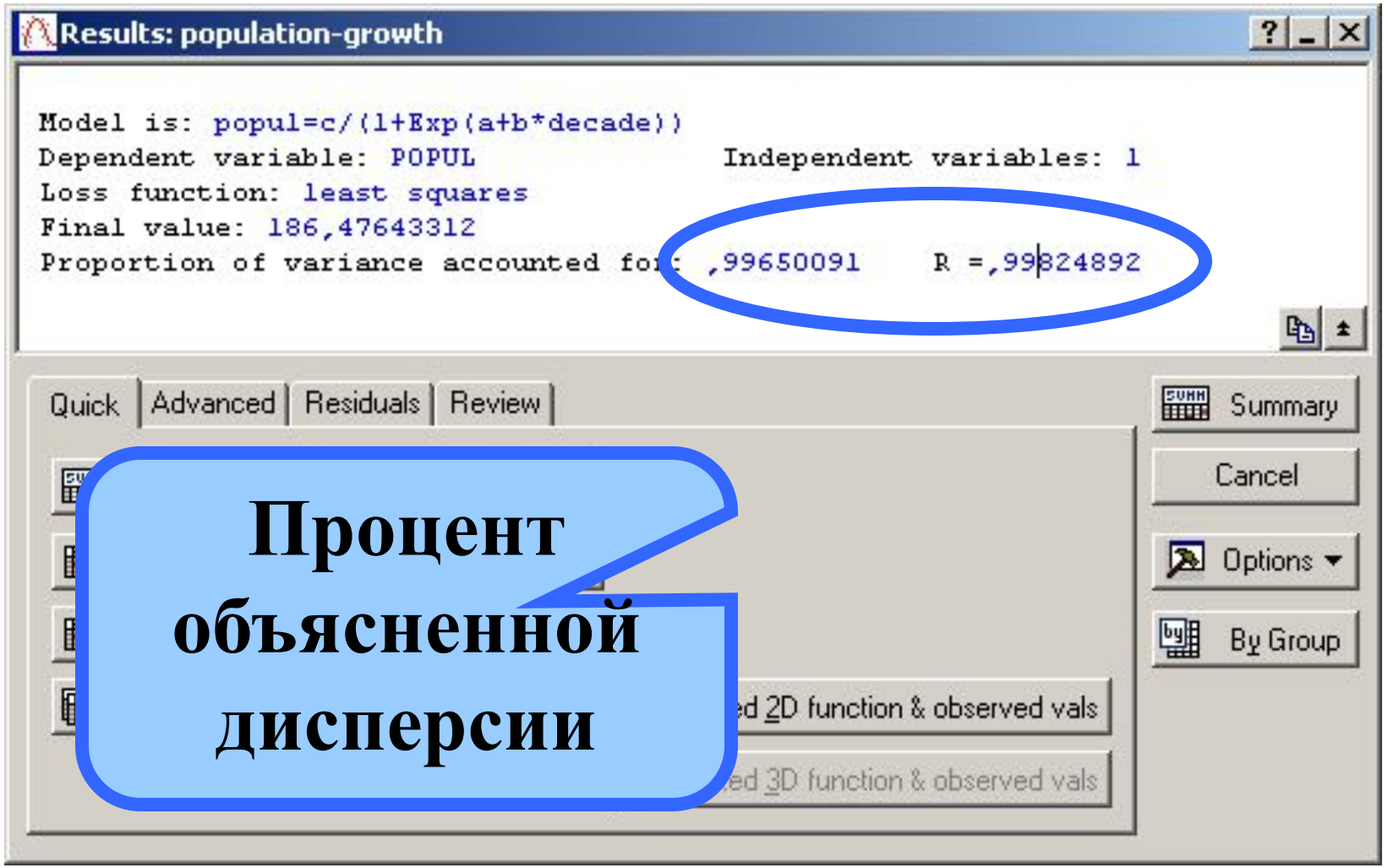
N=18	C	A	B
Estimate	243,9955	3,888804	-,278852

- **Теперь, подставив коэффициенты в исходную формулу**

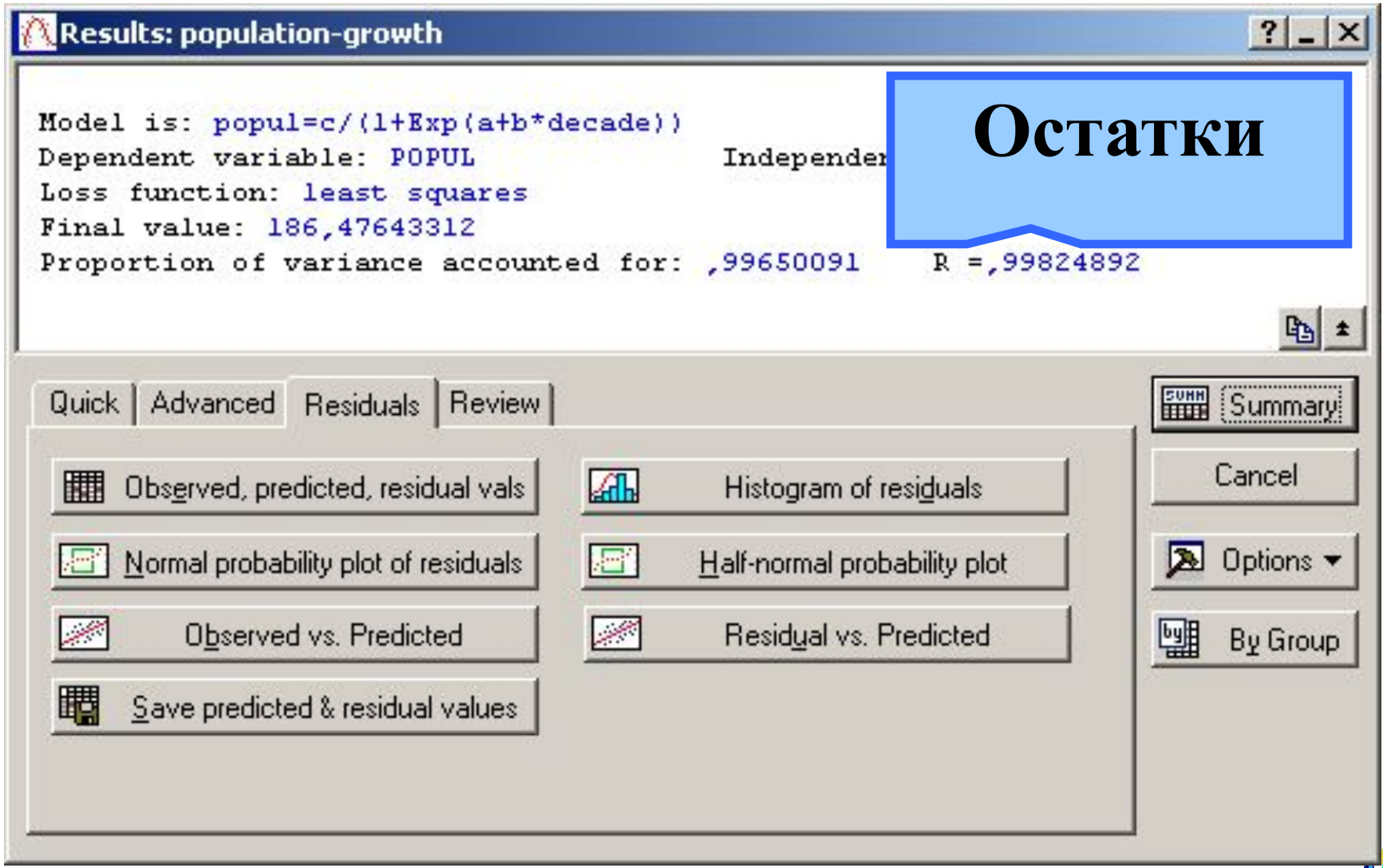
$$population = \frac{244}{1 + e^{3,89 - 0,28t}},$$

мы можем оценить население США в будущем - через 19, 20, 1000 лет…

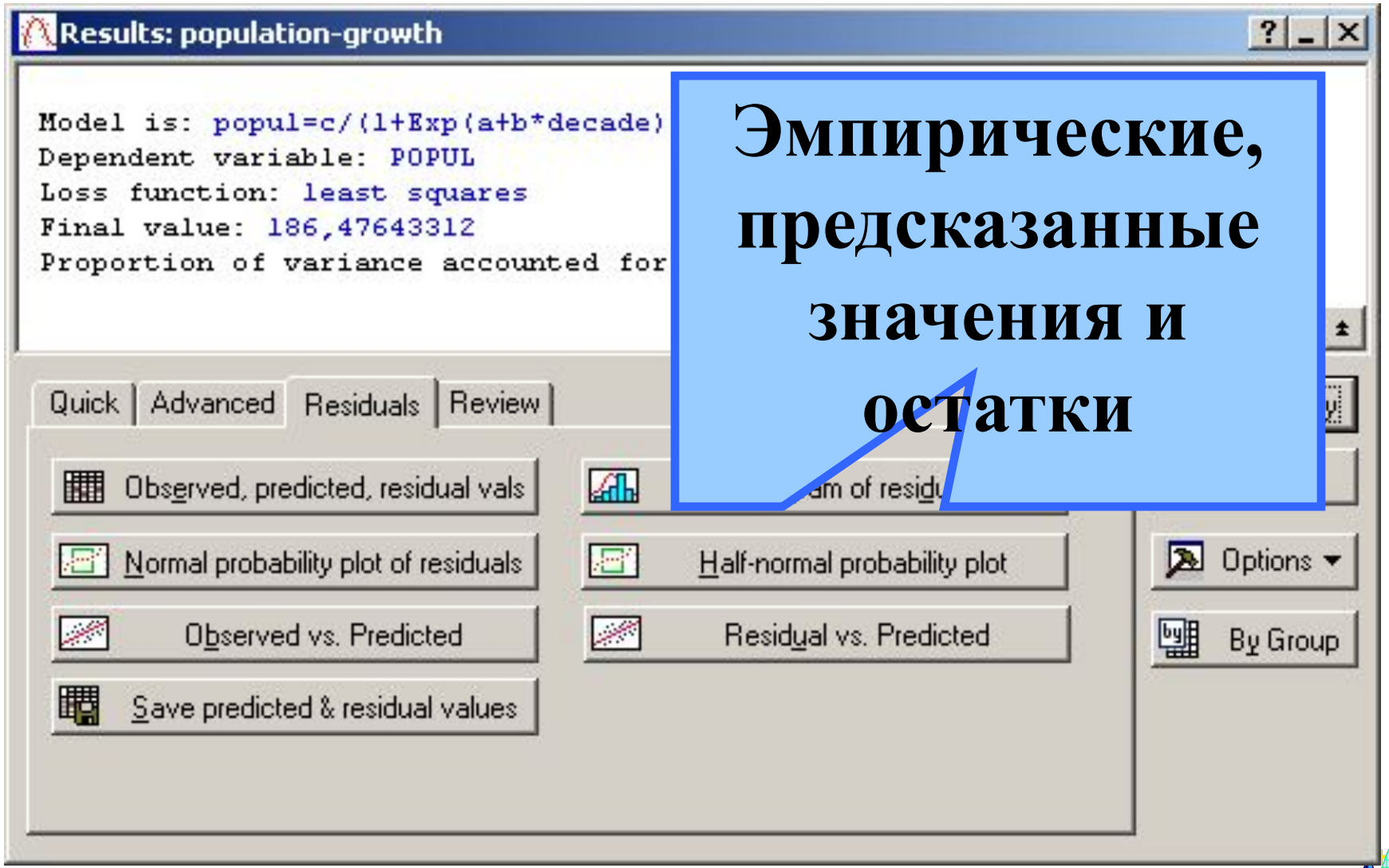
Оценка модели



Оценка модели

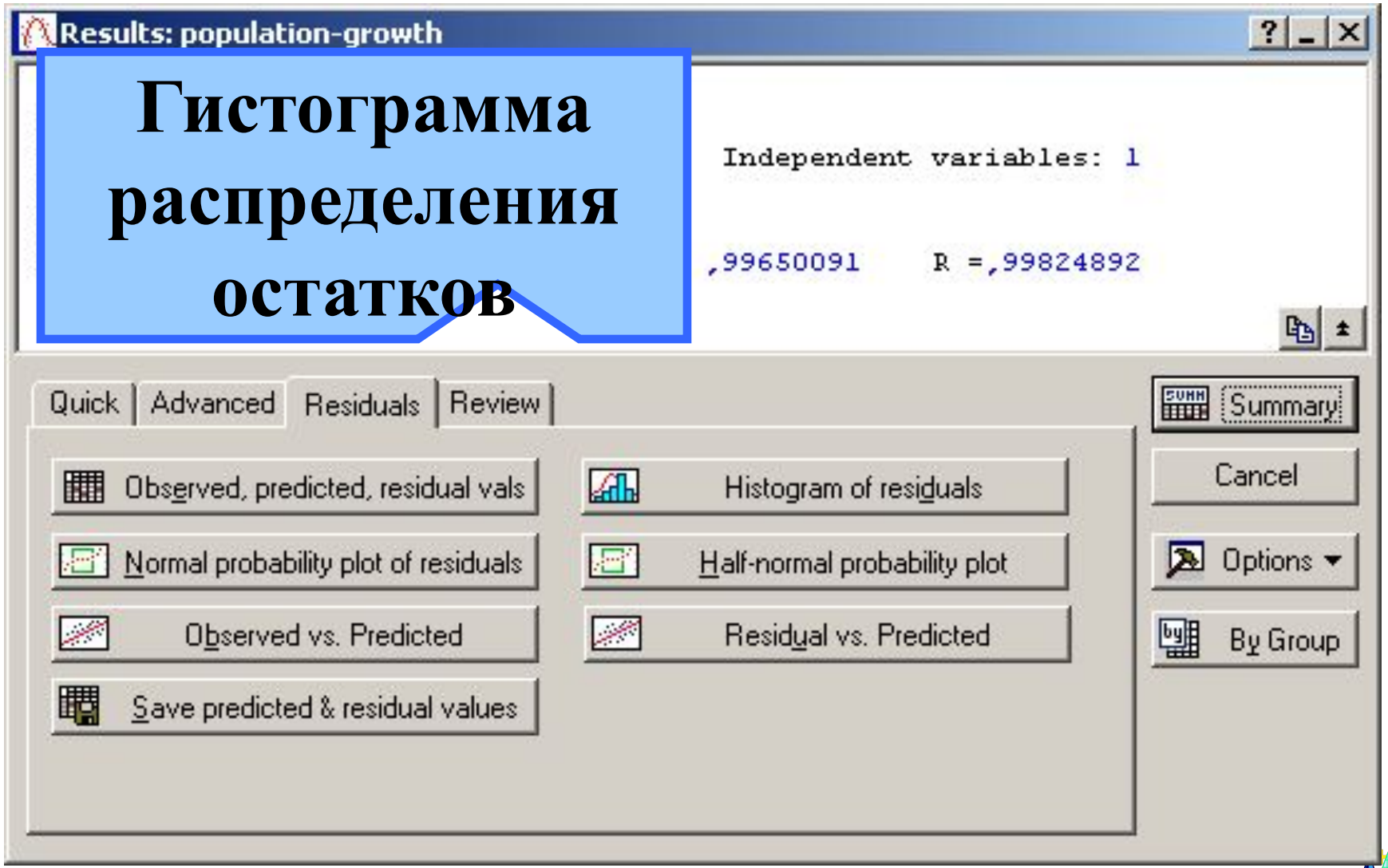


Оценка модели



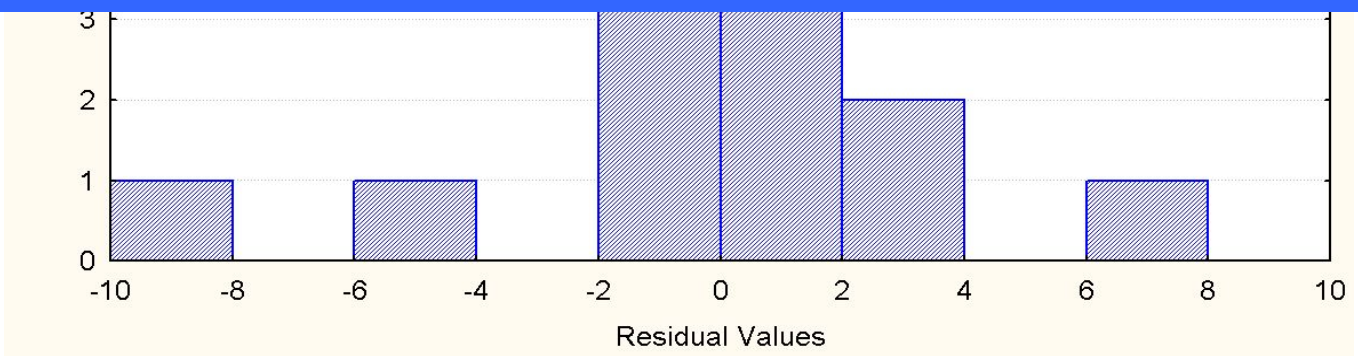
Эмпирические, предсказанные значения и остатки

Оценка модели

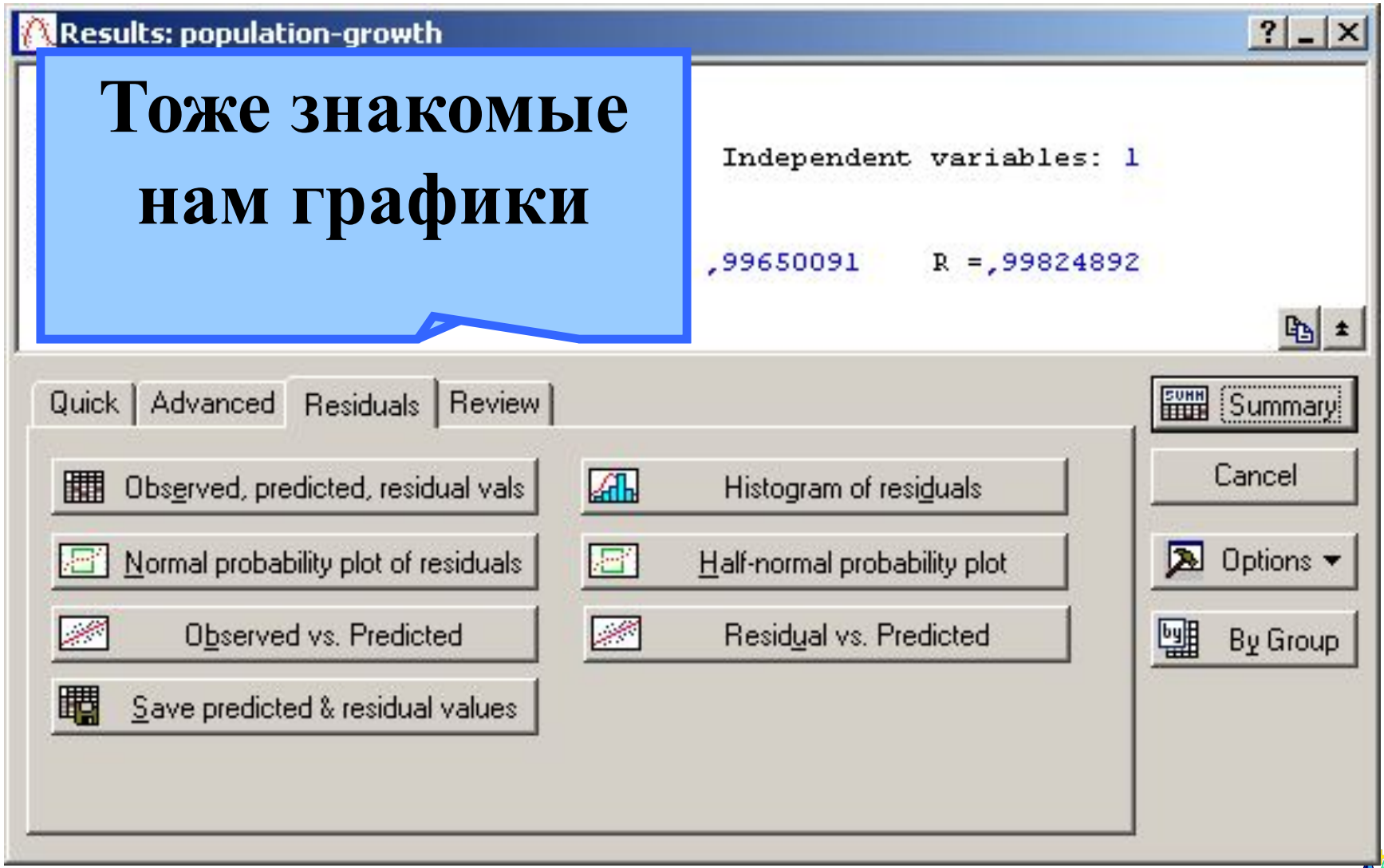


Оценка модели

Распределение должно быть как можно ближе к нормальному

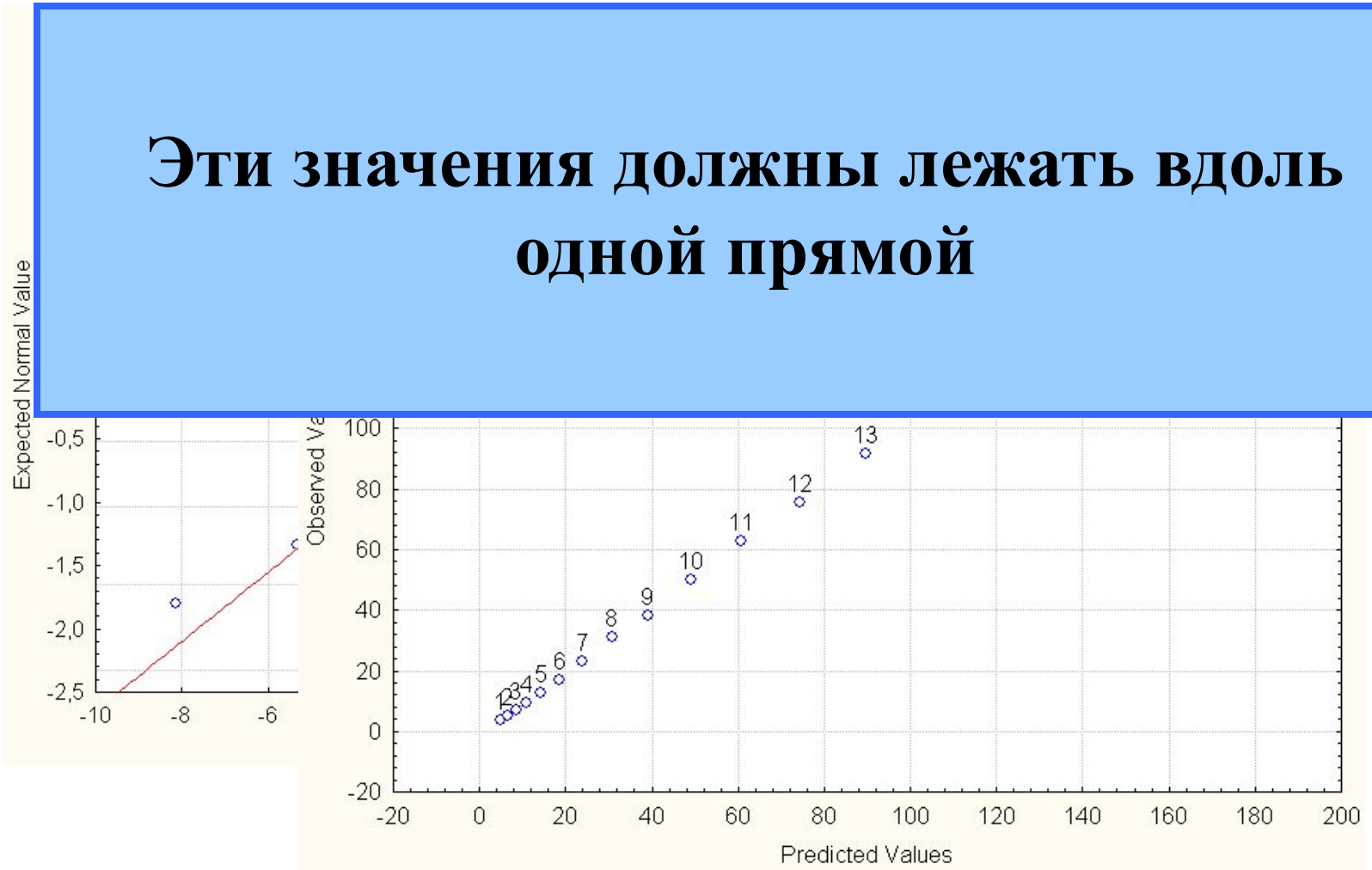


Оценка модели



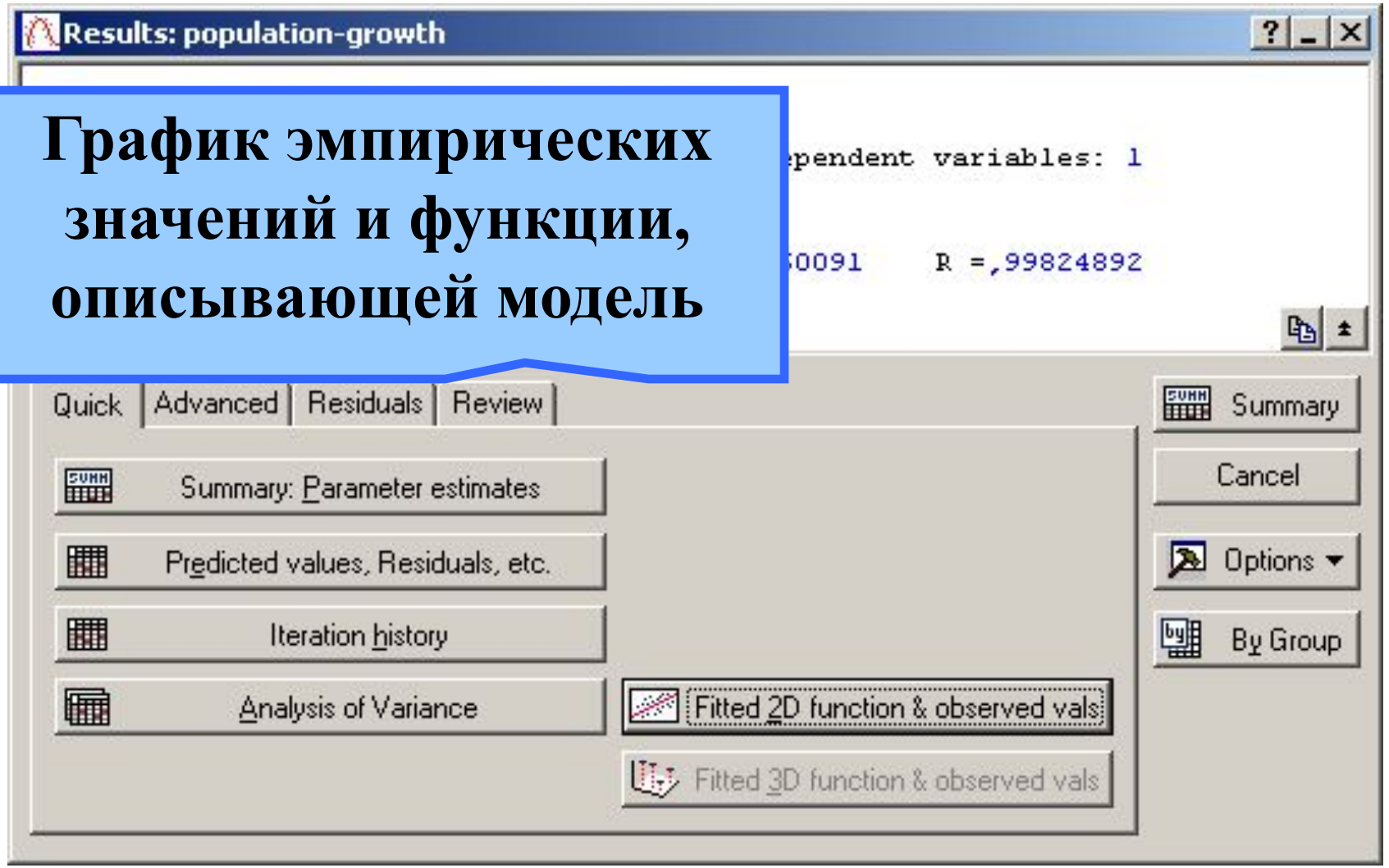
Оценка модели

Эти значения должны лежать вдоль одной прямой

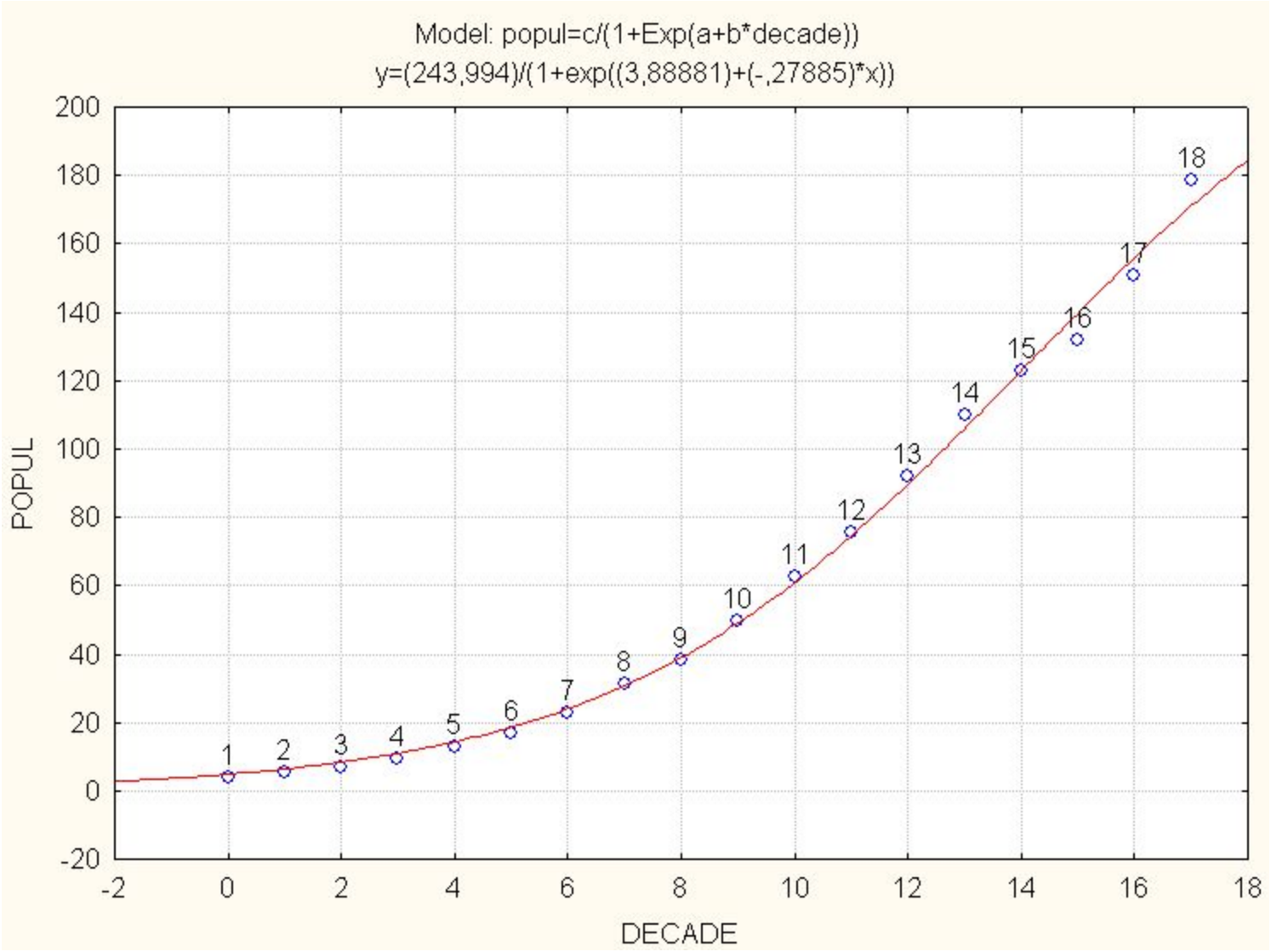


Оценка модели

График эмпирических значений и функции, описывающей модель



Оценка модели



Вот и все!
Задавайте любые зависимости
и проверяйте любые модели!