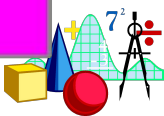
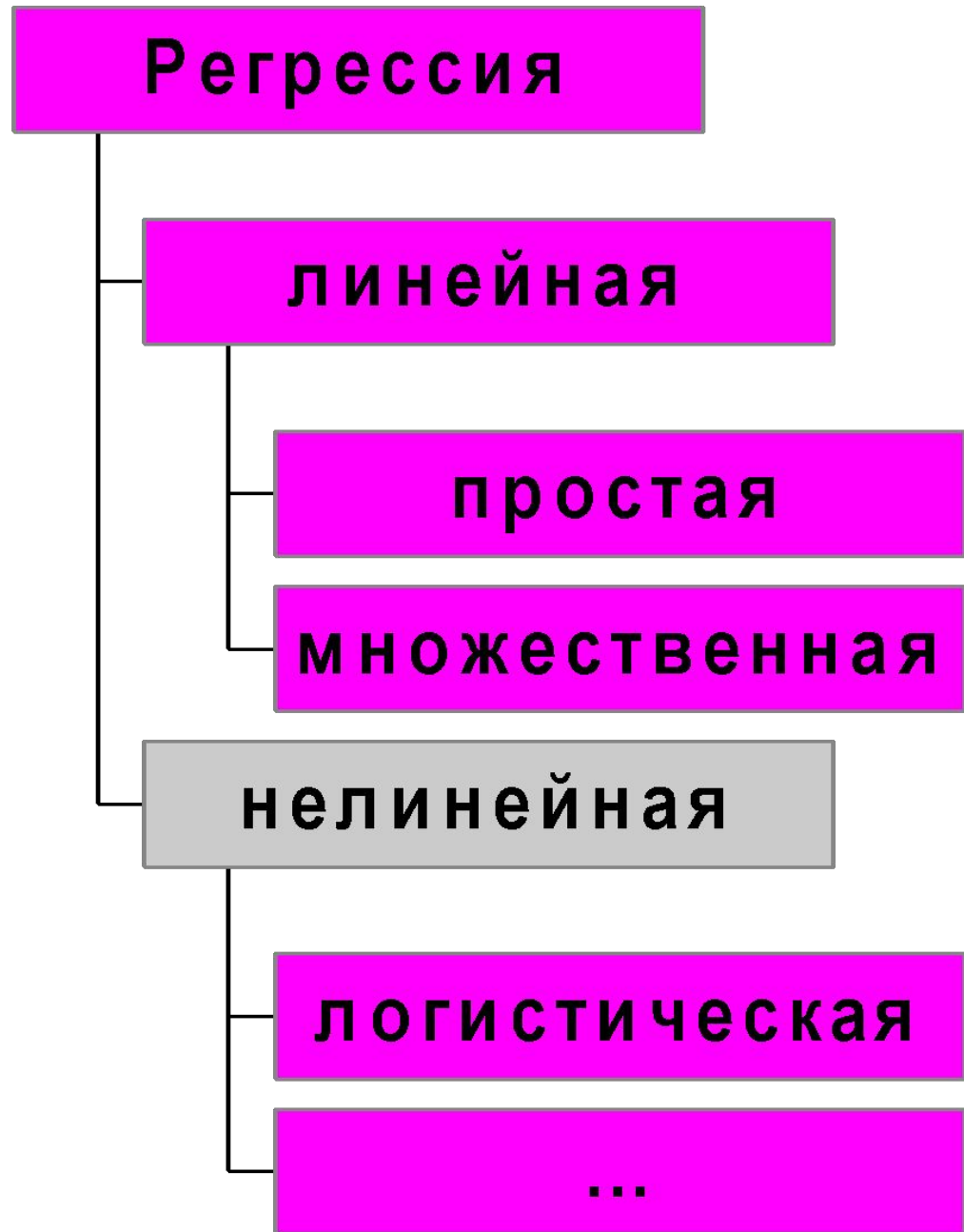


Нелинейная регрессия

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

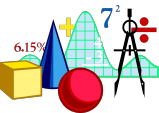


Может быть так, что зависимость между переменными нелинейная. Тогда применяем нелинейную регрессию



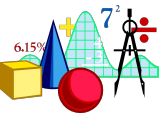


Бинарная логистическая регрессия
позволяет исследовать зависимость
дихотомических зависимых
переменных от независимых
переменных, имеющих любой вид
шкалы





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.





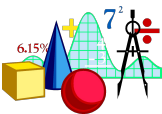
Математическая модель

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

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

p – вероятность наступления события, x – независимые переменные

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



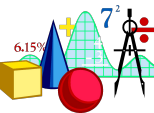


Математическая модель

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

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

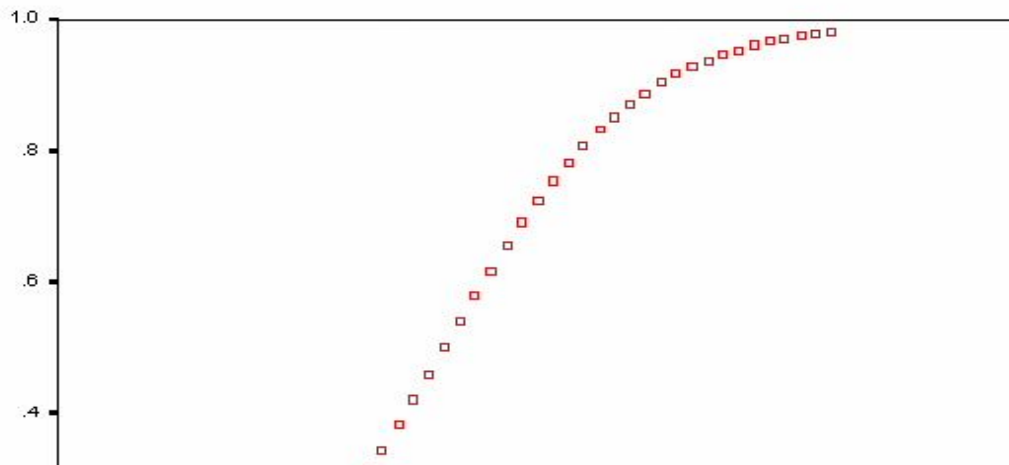
Наша задача, как всегда, - оценить коэффициенты b_i





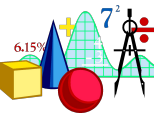
Математическая модель

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



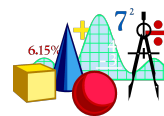
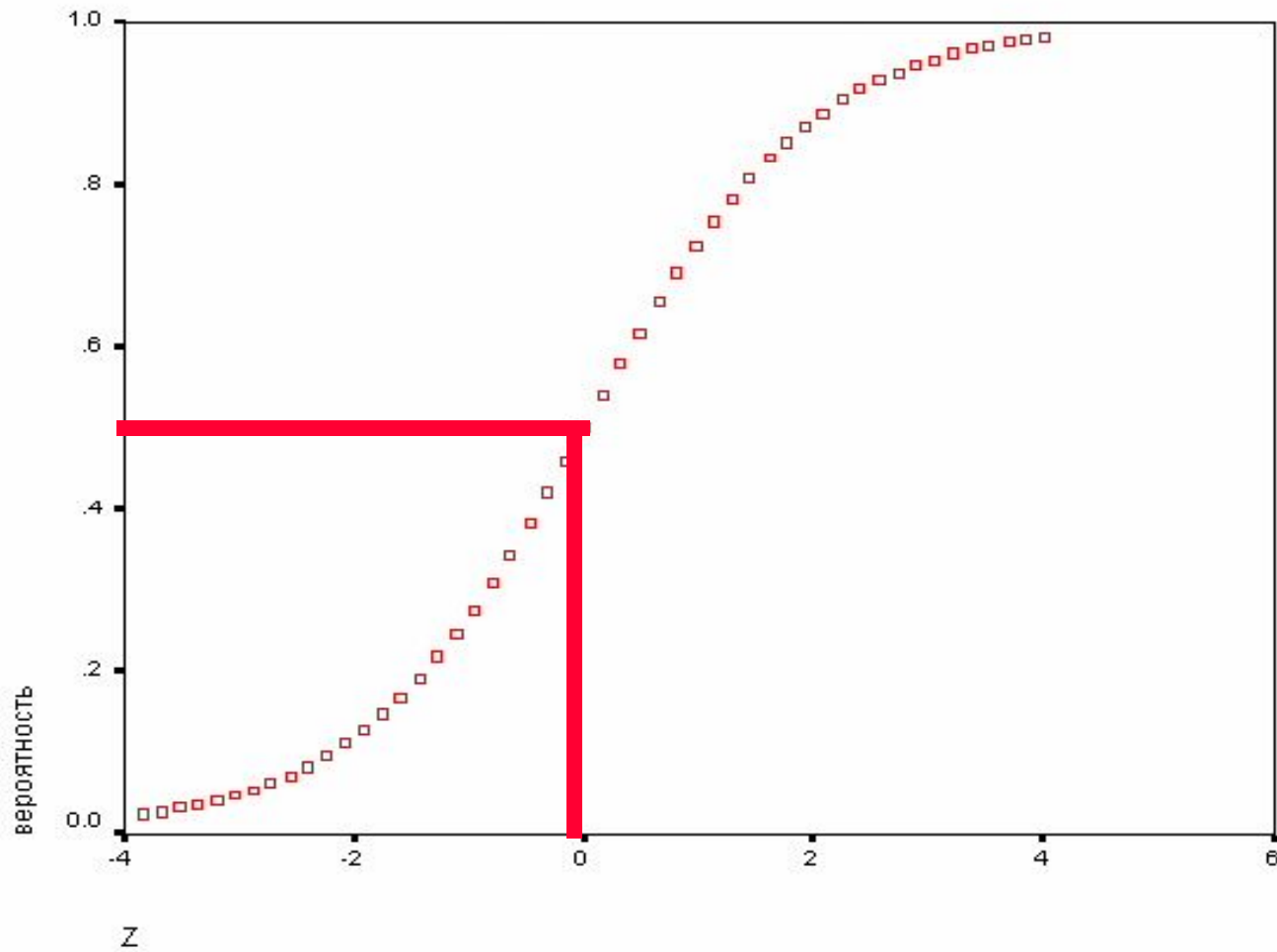
Эта зависимость носит нелинейный характер, причем P не может выходить за пределы диапазона $0 — 1$

Z





Математическая модель

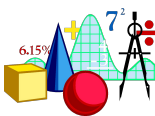




Логистическая регрессия

The screenshot shows the STATISTICA software interface. The 'Statistics' menu is open, and the 'Nonlinear Estimation' option is highlighted. A blue callout box contains the text: **Находится в модуле Nonlinear Estimation**. The menu items are as follows:

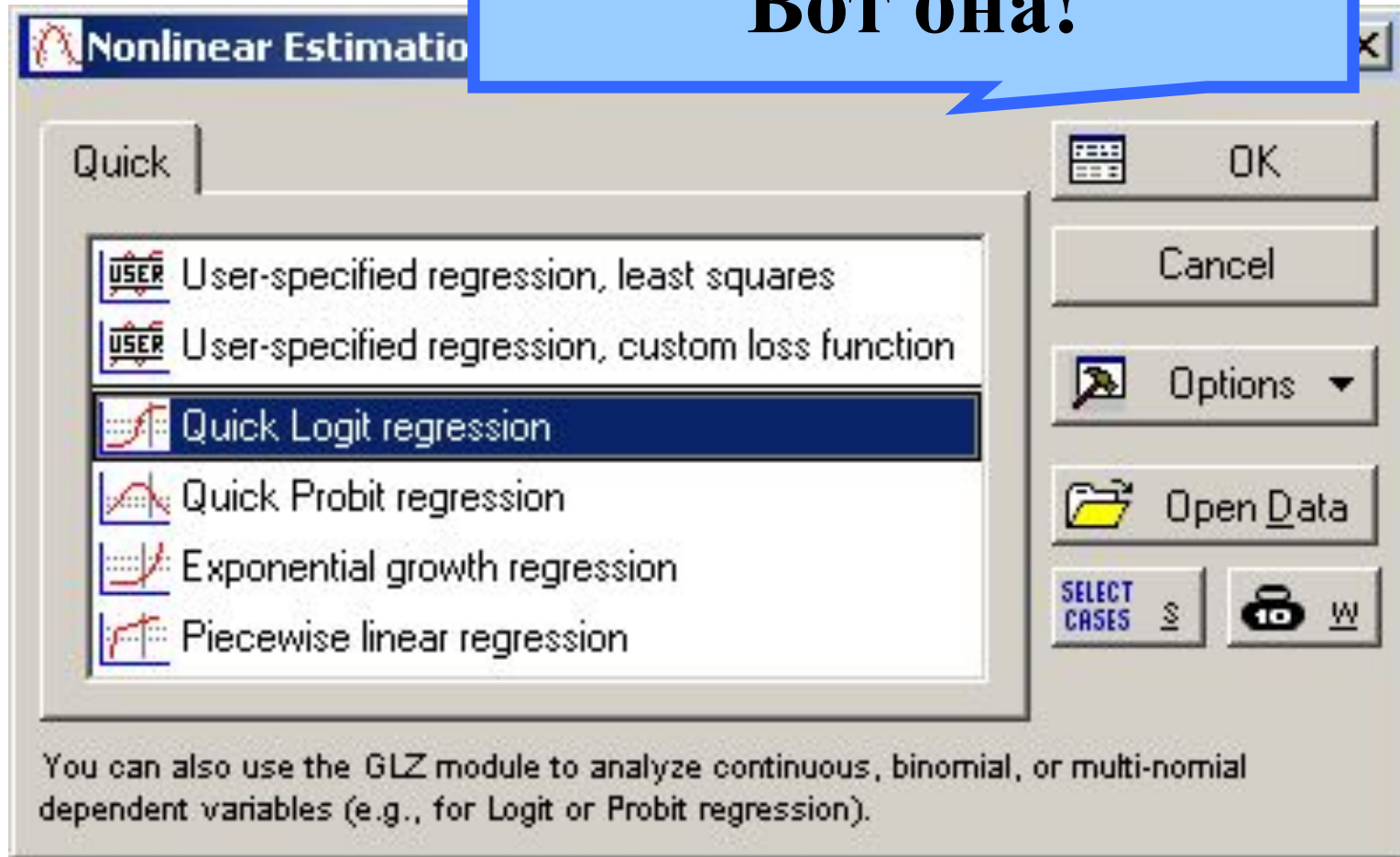
- Resume...
- Basic Statistics/Tables
- Multiple Regression
- ANOVA
- Nonparametrics
- Distribution Fitting
- Advanced Linear/Nonlinear Models**
 - General Linear Models
 - Generalized Linear/Nonlinear Models
 - General Regression Models
 - General Partial Least Squares Models
 - NIPALS Algorithm (PCA/PLS)
 - Variance Components
 - Survival Analysis
 - Nonlinear Estimation**
 - Fixed Nonlinear Regression
- Multivariate Exploratory Techniques
- Industrial Statistics & Six Sigma
- Power Analysis
- Automated Neural Networks
- PLS, PCA, Multivariate/Batch SPC
- Variance Estimation and Precision (VEPAC)
- Statistics of Block Data
- STATISTICA Visual Basic
- Batch (ByGroup) Analysis
- Probability Calculator
- Log-Linear Analysis of Frequency Tables
- Time Series/Forecasting
- Structural Equation Modeling





Логистическая регрессия

Вот она!



Nonlinear Estimation

Quick

- USER User-specified regression, least squares
- USER User-specified regression, custom loss function
- Quick Logit regression**
- Quick Probit regression
- Exponential growth regression
- Piecewise linear regression

OK

Cancel

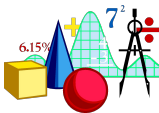
Options ▾

Open Data

SELECT CASES S

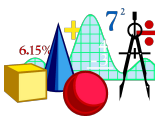
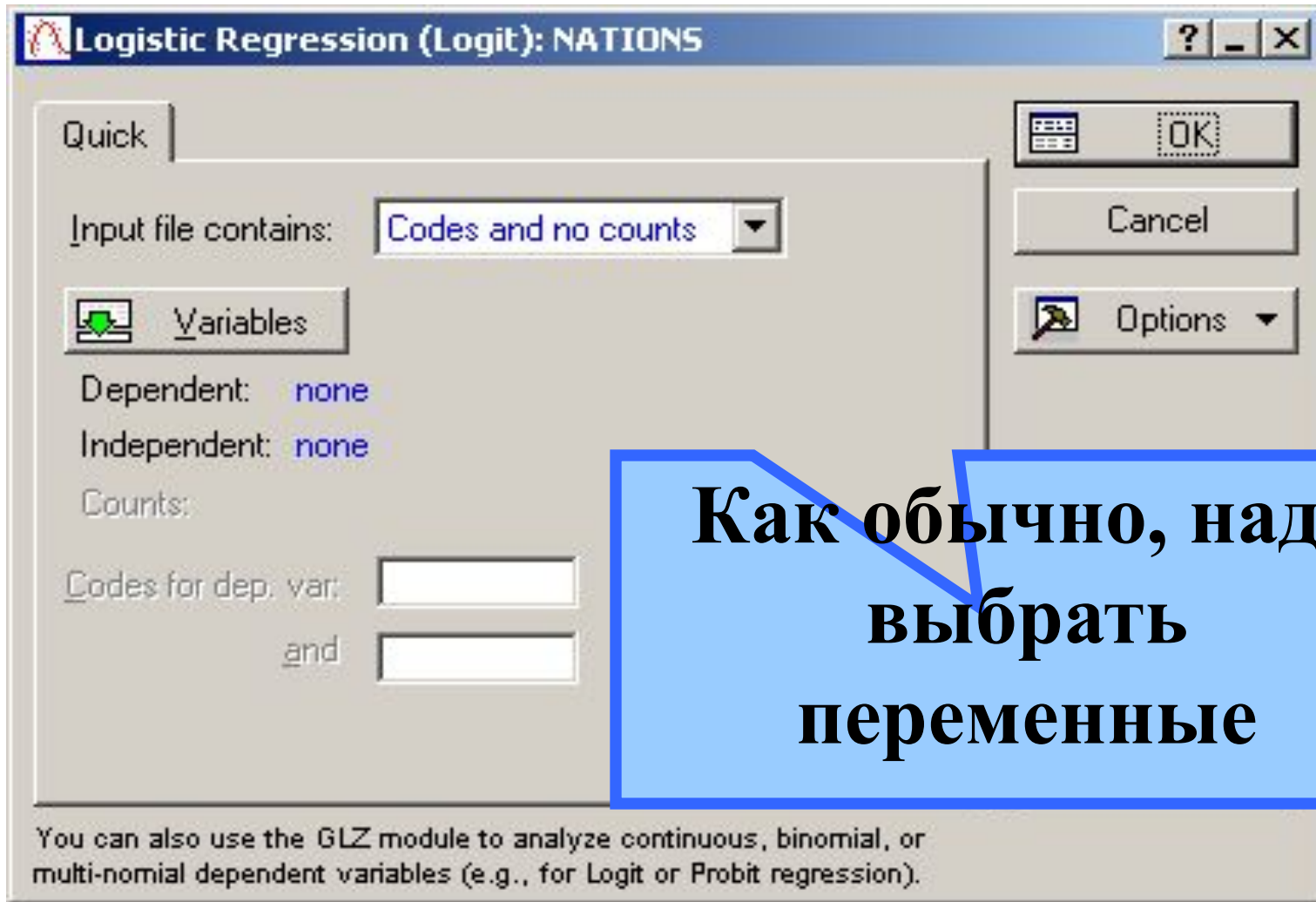
WEIGHTS W

You can also use the GLZ module to analyze continuous, binomial, or multi-nomial dependent variables (e.g., for Logit or Probit regression).





Логистическая регрессия





Пример

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





Пример

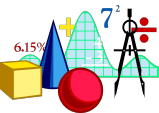
Age – Age (years)

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: 72,0000 (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 values

Fitted 2D function & observed values

Fitted 3D function & observed values

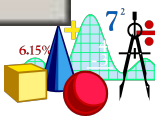
Summary

Cancel

Options

Group

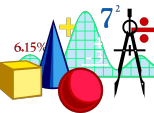
Оценка качества модели





Качество модели

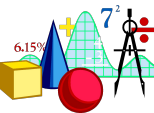
- Качество приближения регрессионной модели оценивается при помощи функции подобия. Мерой правдоподобия служит отрицательное удвоение значения логарифма этой функции - $-2LL$.
- В качестве начального значения для $-2LL$ принимается значение, которое получается для регрессионной модели, содержащей только константу.



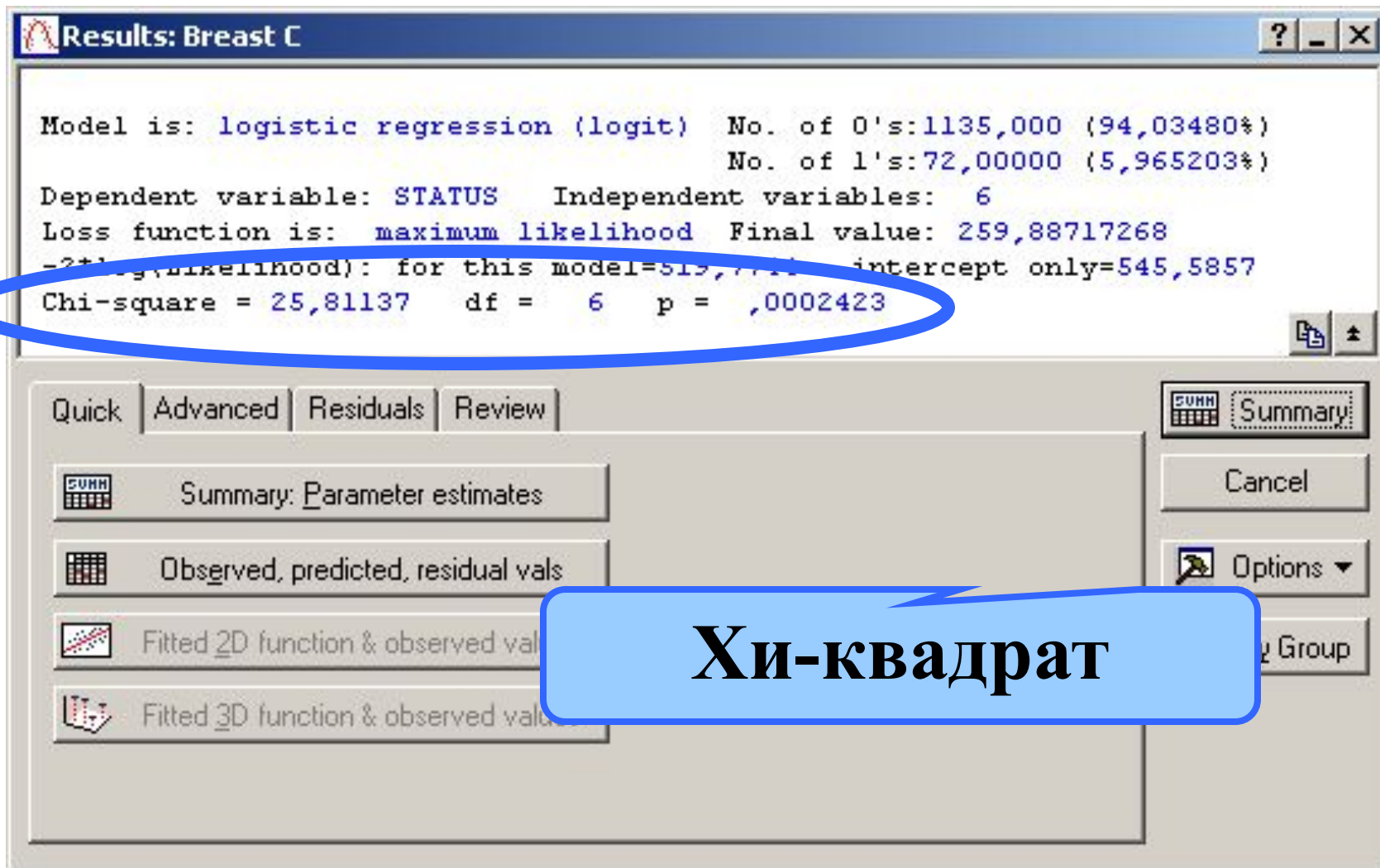


Качество модели

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



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*ln(-likelihood): for this model=519,7741 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 vals

Fitted 3D function & observed vals

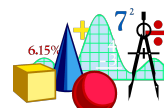
Summary

Cancel

Options

Group

Хи-квадрат





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

Cancel

Options

By Group

Коэффициенты b



Регрессионные коэффициенты

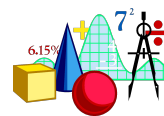
Workbook1* - Model: Logistic regression (logit) N of 0's: 1135 1's: 72 (Breast C)

Model: Logistic regression (logit) N of 0's: 1135 1's
Dep. var: STATUS Loss: Max likelihood
Final loss: 268,18513440 Chi?(2)=9,2154 p= 0099

N=1207

	Const.B0	AGE	PATHSIZE
Estimate	-1,28331	-0,827814	0,002496
Odds ratio (unit ch)	0,27712	0,972764	1,002500
Odds ratio (range)		0,161614	1,280055

Model: Logistic regression (logit) N of 0's: 1135 1's: 72 (Br... Model: Logistic re





Результаты

Эмпирические,
предсказанные
значения и остатки

Results: Breast C

Model is: logistic regression

Dependent variable: STATUS

Loss function is: maximum likelihood

-2*log(Likelihood): for the full model = 9,215445

Chi-square = 9,215445 df = 1 p = ,009981

Quick | **Advanced** | Residuals | Review

Observed, predicted, residual vals | Histogram of residuals

Normal probability plot of residuals | Half-normal probability plot

Classification of cases & odds ratio | Predicted vs. observed values

Save predicted and residual values | Predicted vs. residual values

Summary | Cancel | Options | By Group





Результаты

STATISTICA - Workbook1* - [Model is: (Breast C)] - [Workbook1* - M

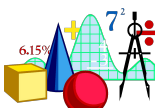
File Edit View Insert Format Statistics Data Mining Graphs Tools

Arial 10 B I U

Workbook1*
Nonlinear Esti
Nonlinear
Model
Model
Model

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

Results: Breast C

Model is: logistic regression

Dependent variable: STATUS

Loss function is: maximum likelihood Final value: 268,18513440

-2*log(Likelihood): for this model=536,3702 intercept only=545,5857

Chi-square = 9,215445 df = 2 p = ,0099811

Quick | Advanced | **Residuals** | Review

Observed, predicted, residual vals | Histogram of residuals

Normal probability plot of residuals | Half-normal probability plot

Classification of cases & odds ratio | Predicted vs. observed values

Save predicted and residual values | Predicted vs. residual values

Summary | Cancel | Options | By Group





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

Output text from the software window:

```
igit) No. of 0's:1135,000 (94,03480%)  
      No. of 1's:72,00000 (5,965203%)  
dependent variables: 2  
Maximum likelihood Final value: 268,18513440  
-2*log(Likelihood): for this model=536,3702 intercept only=545,5857  
Chi-square = 9,215445 df = 2 p = ,0099811
```

Residuals analysis options:

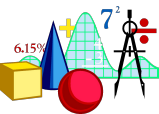
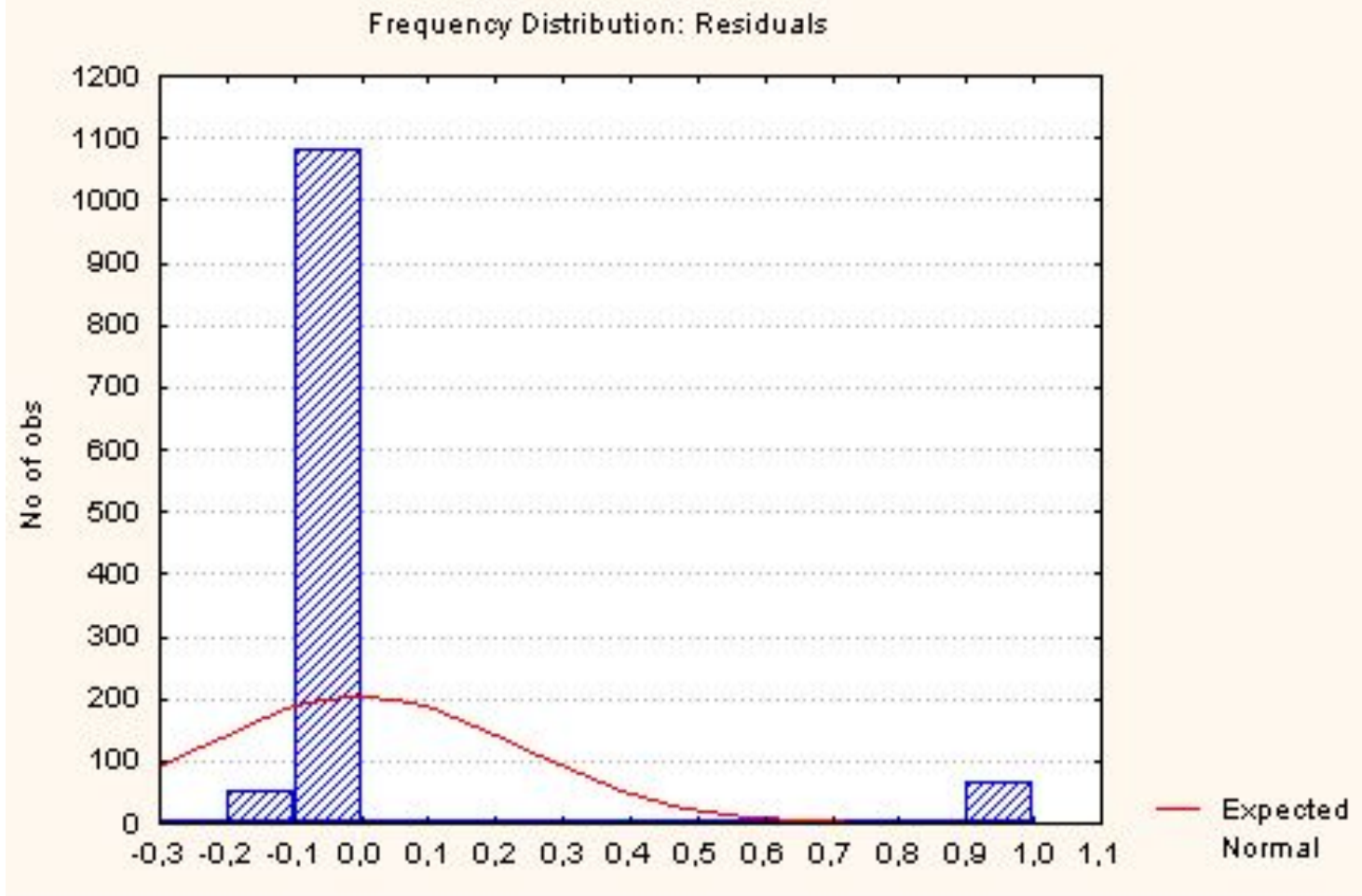
- Observed, predicted, residual vals
- Histogram of residuals** (circled in blue)
- Normal probability plot of residuals
- Half-normal probability plot
- Classification of cases & odds ratio
- Predicted vs. observed values
- Save predicted and residual values
- Predicted vs. residual values

Additional options: Summary, Cancel, Options, By Group





Результаты





Результаты

**Знакомые нам
графики
оценки**

The screenshot shows a software window with the following text:

```
igit) No. of 0's:1135,000 (94,03480%)  
      No. of 1's:72,00000 (5,965203%)  
pendent variables: 2  
hood Final value: 268,18513440  
l=536,3702 intercept only=545,5857  
p = ,0099811
```

Below the text is a dialog box with tabs: Quick, Advanced, Residuals, Review. The 'Residuals' tab is active. It contains several options:

- Observed, predicted, residual vals
- Normal probability plot of residuals
- Classification of cases & odds ratio
- Save predicted and residual values
- Histogram of residuals
- Half-normal probability plot
- Predicted vs. observed values
- Predicted vs. residual values

On the right side of the dialog box, there are buttons for Summary, Cancel, Options, and By Group. The 'Predicted vs. observed values' button is circled in blue.





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

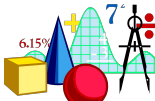




$$p = \frac{\sum_{h(t)h'} \sqrt{\frac{c}{1 - e^{a+bt}}}}{\log \text{cov}(h(t)g(t) \frac{c}{1 + e^{a+bt}})}$$

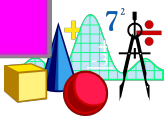
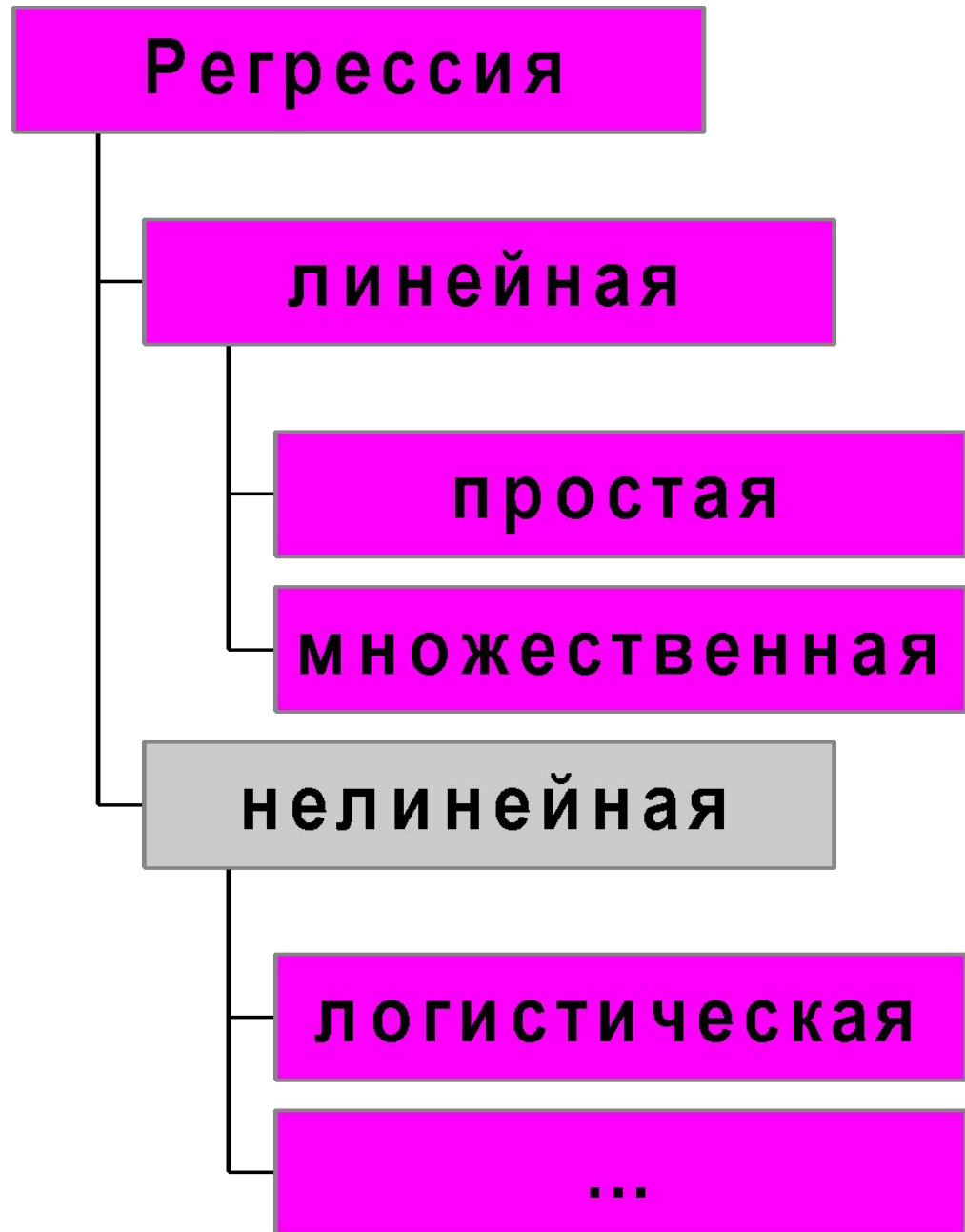


**Оценка на экзамене и
мотивация так прямо не
связаны ...**





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



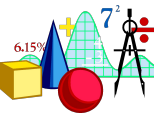
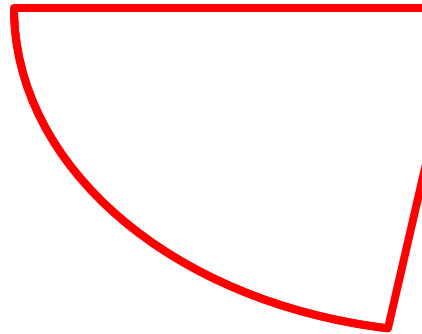


Пример.

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

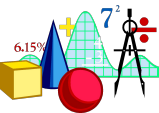
Видно, что зависимость тут скорее не линейная, а экспоненциальная. Демографы знают, что лучше всего зависимость роста населения от времени описывается функцией

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





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





Для построения уравнений нелинейной регрессии служит модуль Nonlinear Estimation

Nonlinear Estimation: NATIONS

Quick

- User-specified regression, least squares
- User-specified regression, custom loss function
- Quick Logit regression
- Quick Probit regression
- Exponential growth regression
- Piecewise linear regression

OK

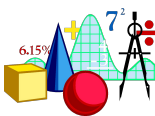
Cancel

Options

Open Data

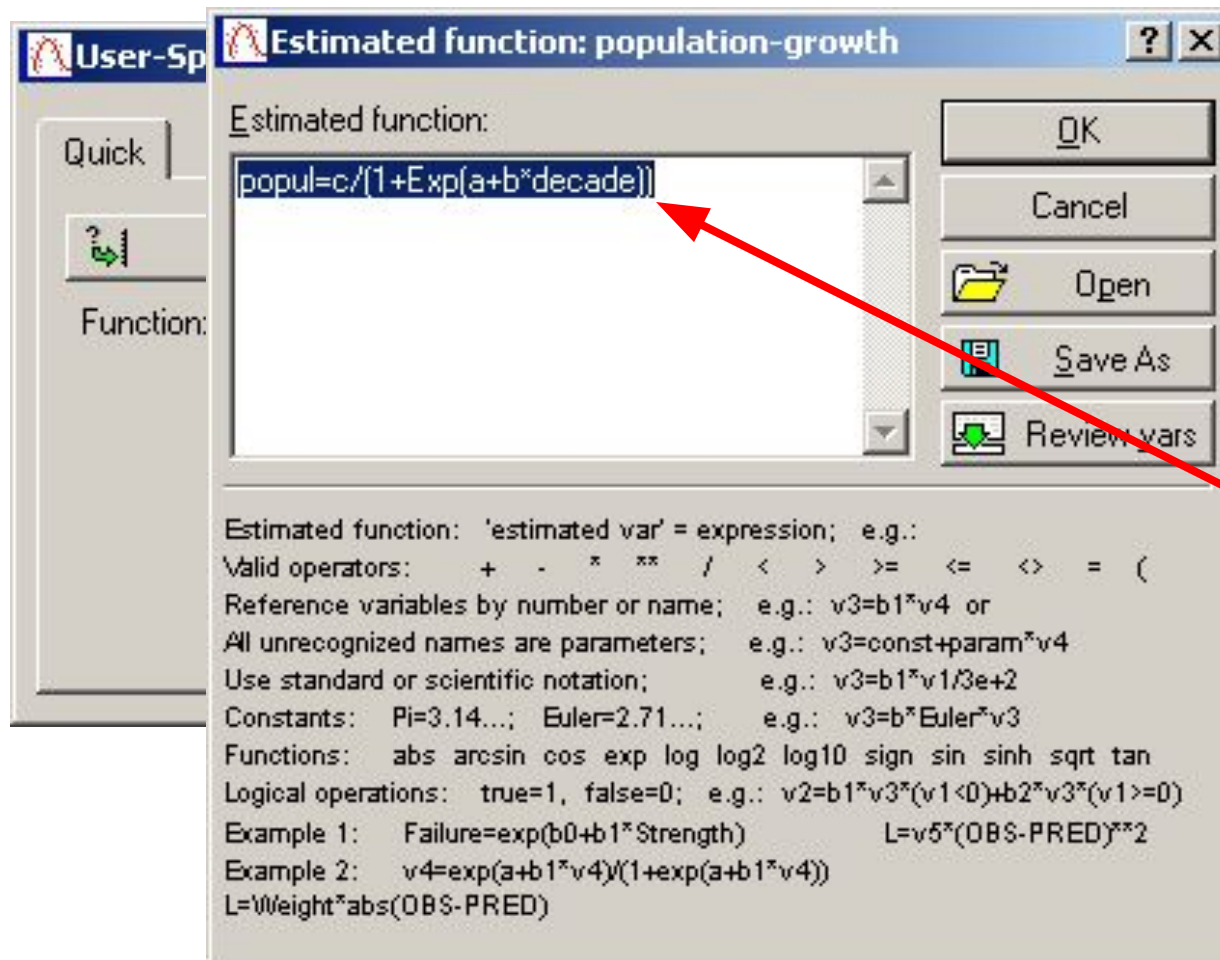
SELECT CASES

You can also use the GLZ module to analyze continuous, binomial, or multi-nomial dependent variables (e.g., for Logit or Probit regression).

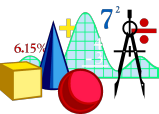




Для построения уравнений нелинейной регрессии служит модуль Nonlinear Estimation



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



Маленькие (?) хитрости

Nonlinear Least Squares Model Estimation: population-growth

Model is: $popul=c/(1+Exp(a+b*decade))$
Number of parameters to be estimated: 3
Loss function is: least squares
Dependent variable: POPUL

Independent variables: DECADE

Missing data are casewise deleted

Number of valid cases: 18

Quick | **Advanced** | Review

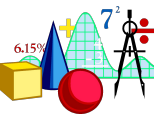
Estimation method: Levenberg-Marquardt

Maximum number of iterations: 50

Convergence criterion: 1.0 E- 6

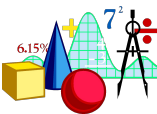
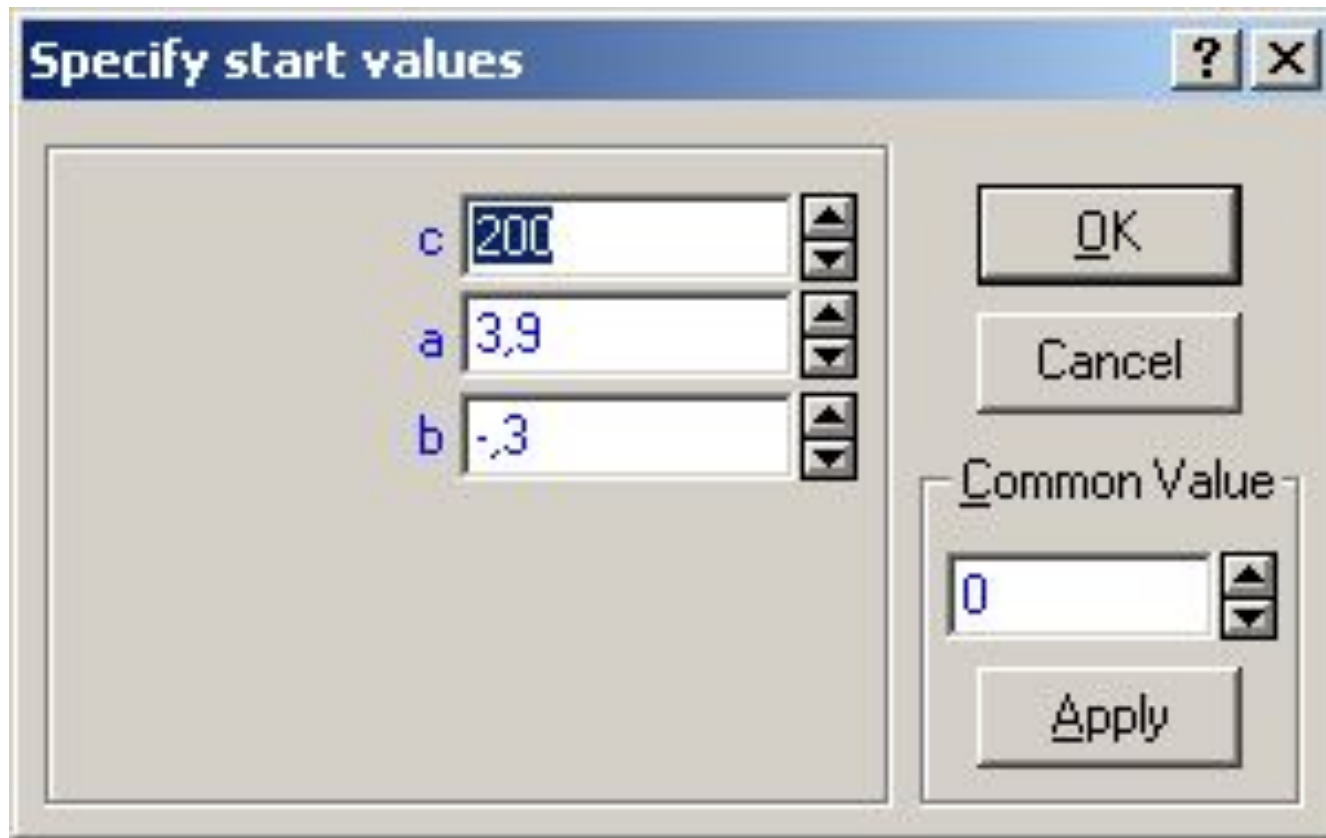
Start values: Various

Callout text: **Начальные значения для параметров**





Маленькие (?) хитрости





Results: population-growth

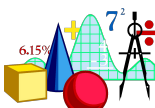
Model is: $\text{popul} = c / (1 + \text{Exp}(a + b * \text{decade}))$
Dependent variable: POPUL Independent variables: 1
Loss function: least squares
Final value: 186,47643312
Proportion of variance accounted for: ,99650091 R = ,99824892

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!**



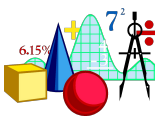


Оценка параметров

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%

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

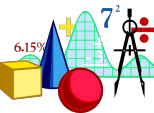




- **Теперь, подставив коэффициенты в исходную формулу**

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

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





Оценка модели

Results: population-growth

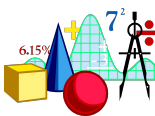
```
Model is: popul=c/(1+Exp(a+b*decade))
Dependent variable: POPUL           Independent variables: 1
Loss function: least squares
Final value: 186,47643312
Proportion of variance accounted for: ,99650091    R = ,99824892
```

Quick | Advanced | Residuals | Review

Summary
Cancel
Options
By Group

ed 2D function & observed vals
ed 3D function & observed vals

**Процент
объясненной
дисперсии**





Оценка модели

Results: population-growth

Model is: $popul = c / (1 + \exp(a + b * decade))$
Dependent variable: POPUL Independent variable: decade
Loss function: least squares
Final value: 186,47643312
Proportion of variance accounted for: ,99650091 R = ,99824892

Остатки

Quick | **Advanced** | Residuals | Review

Summary

Observed, predicted, residual vals Histogram of residuals
Normal probability plot of residuals Half-normal probability plot
Observed vs. Predicted Residual vs. Predicted
Save predicted & residual values

Cancel
Options
By Group





Оценка модели

Results: population-growth

Model is: $\text{popul} = c / (1 + \text{Exp}(a + b * \text{decade}))$
Dependent variable: POPUL
Loss function: least squares
Final value: 186,47643312
Proportion of variance accounted for

Quick | Advanced | Residuals | Review

Observed, predicted, residual vals | Histogram of residuals

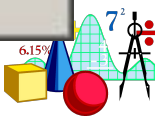
Normal probability plot of residuals | Half-normal probability plot

Observed vs. Predicted | Residual vs. Predicted

Save predicted & residual values | Options

By Group

Эмпирические, предсказанные значения и остатки





Оценка модели

Results: population-growth

Independent variables: 1

,99650091 R = ,99824892

Quick | Advanced | **Residuals** | Review

Observed, predicted, residual vals

Normal probability plot of residuals

Observed vs. Predicted

Save predicted & residual values

Histogram of residuals

Half-normal probability plot

Residual vs. Predicted

Summary

Cancel

Options

By Group

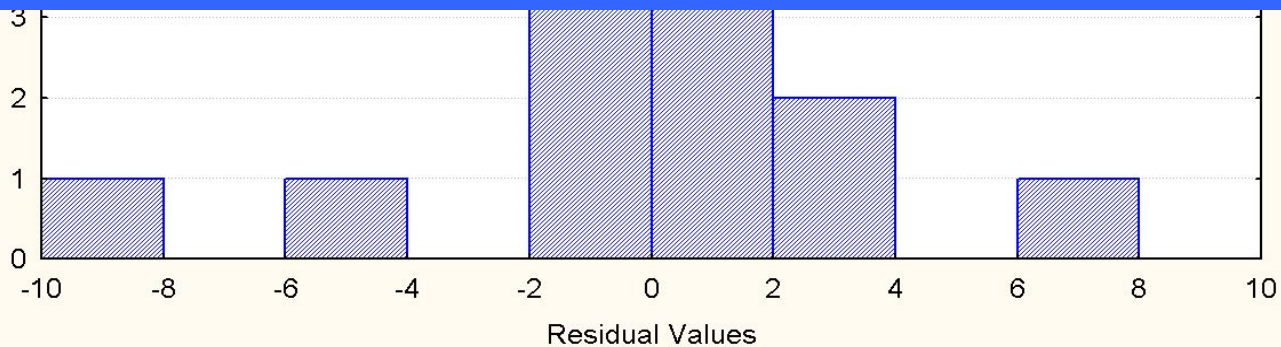
**Гистограмма
распределения
остатков**





Оценка модели

**Распределение должно быть как
можно ближе к нормальному**






Оценка модели

Тоже знакомые
нам графики


Independent variables: 1


,99650091 R = ,99824892


Quick | **Advanced** | Residuals | Review


 Summary

Cancel


 Options ▾


 By Group


 Observed, predicted, residual vals


 Histogram of residuals

 Normal probability plot of residuals

 Half-normal probability plot

 Observed vs. Predicted

 Residual vs. Predicted

 Save predicted & residual values

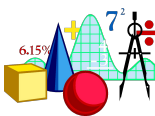
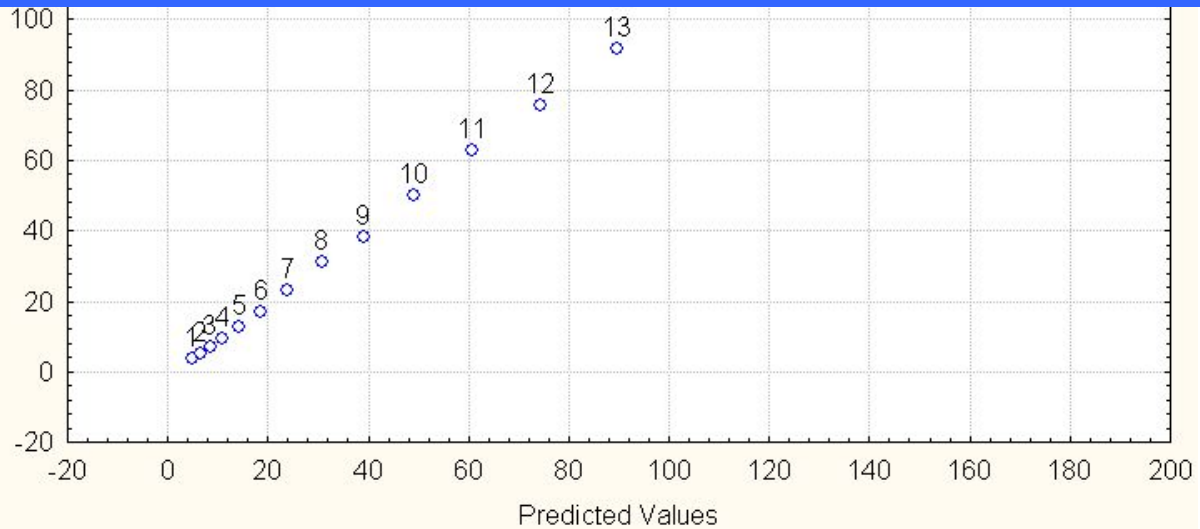
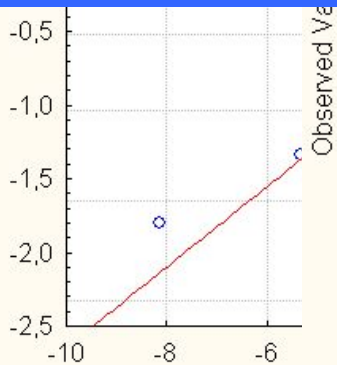




Оценка модели

Эти значения должны лежать вдоль одной прямой

Expected Normal Value





Оценка модели

График эмпирических значений и функции, описывающей модель

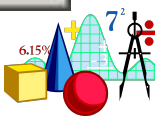
The screenshot shows a software window titled "Results: population-growth". The main content area displays the following text:

```
dependent variables: 1  
0091 R = ,99824892
```

Below the text is a menu with several options:

- Quick | Advanced | Residuals | Review
- Summary: Parameter estimates
- Predicted values, Residuals, etc.
- Iteration history
- Analysis of Variance
- Fitted 2D function & observed vals (highlighted)
- Fitted 3D function & observed vals

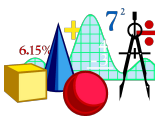
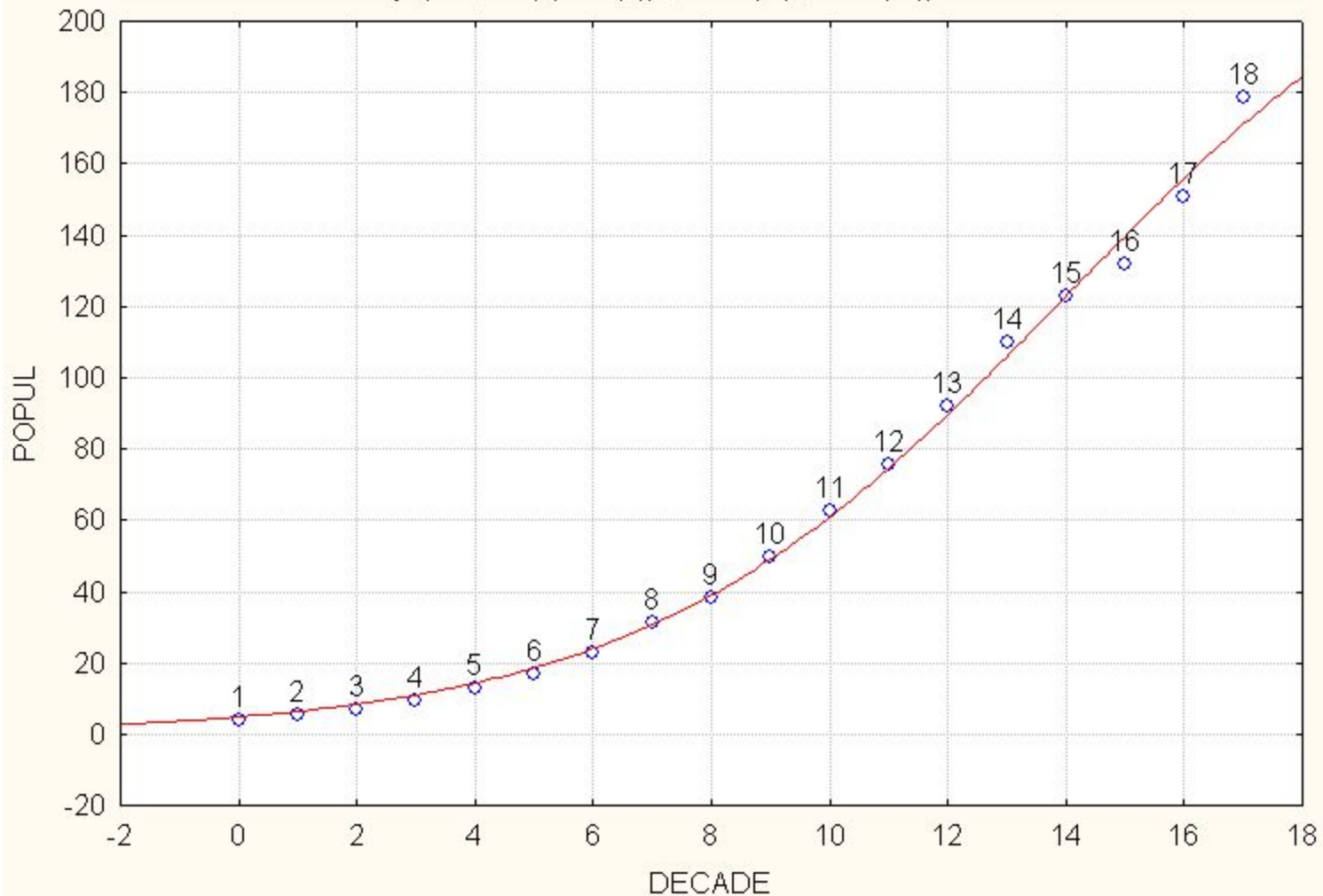
On the right side of the window, there are additional buttons: Summary, Cancel, Options, and By Group.





Оценка модели

Model: $\text{popul} = c / (1 + \text{Exp}(a + b * \text{decade}))$
 $y = (243,994) / (1 + \exp((3,88881) + (-,27885) * x))$





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

