Модели бинарных откликов

В предыдущем разделе мы провели регрессионный анализ в предположении, что отклик − переменная *Тест* является непрерывной случайной величиной, имеющей нормальное распределение. На самом деле *результаты тестирования*, как и другие количественные переменные в таблице исходных данных − *пульс*, *IQ*, *давление*, *уровень тревожности*, *длительность подготовки*, принимают целочисленные значения. Так как эти величины принимают большое количество различных значений, мы в праве сделать допущение о непрерывности. В отличие от указанных величин переменная *Зачет* принимает всего два значения − *зачет*, *незачет*. Такие переменные называются бинарными. Они принимают два значения, которые могут обозначаться по-разному, например, 0 и 1; или *да* и *нет* и т.д. В программе *STATISTICA* предусмотрена возможность построения уравнения регрессии, если отклик является бинарной величиной, а предикторы − непрерывными переменными.

Так как технически достаточно сложно смоделировать бинарную функцию от непрерывных аргументов, задачу регрессии формулируют иначе. Вместо предсказания бинарной переменной предсказывают непрерывную переменную со значениями на отрезке [0, 1]. Если при этом переменная принимает значение большее либо равное 0,5, то полагают что она равна 1, в противном случае − 0. В программе *STATISTICA* реализованы логит и пробит модели [6].

Посредством уравнения регрессии

$$Y = exp(b_0 + b_1X_1 + ... + b_nX_n)/\{1 + exp(b_0 + b_1X_1 + ... + b_nX_n)\}$$

в логит модели значения Y вне зависимости от коэффициентов регрессии и значений X всегда будут принадлежать отрезку [0, 1].

В пробит регрессии предполагается, что Y – это некоторая функция распределения. Как известно из теории вероятностей, функция распределения принимает значения только из интервала [0, 1] .

Построим логит регрессионную модель зависимости отклика *Группа риска* от 21 предиктора - *финансового показателя предприятия* на платформе пакета STATISTICA по модельной базе данных.

Вероятностно-статистическая модель логит-регрессии имеет вид:

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

$$z = b_0 + b_1\mathrm{R}_3 + b_2\mathrm{R}_4 + b_3\mathrm{R}_5 + b_4\mathrm{R}_6 + b_5\,\mathrm{L}_1 + b_6\mathrm{L}_2 + b_7\mathrm{L}_3 + b_8\mathrm{P}_1 + b_9\mathrm{P}_2 + b_{10}\frac{1}{\mathrm{A}_1} + b_{11}\mathrm{A}_2 + b_{12}\frac{1}{\mathrm{A}_3} +$$

$$+ b_{13}\mathrm{A}_4 + b_{14}\frac{1}{\mathrm{A}_5} + b_{15}\,\mathrm{A}_6 + b_{16}\,\mathrm{F}_1 + b_{17}\,\mathrm{F}_2 + b_{18}\,\mathrm{F}_3 + b_{19}\mathrm{F}_4 + b_{20}\frac{1}{\mathrm{F}_8} + b_{21}\mathrm{F}_{11}$$

Для построения модели необходимо вычислить коэффициенты $b_0,...,b_{21}$.

Бинарную переменную «уровень риска», принимающую значение «кредит» или «отказ» получаем по правилу:

если $Y \in [0,0.5]$, то *высокий риск*,

если $Y \in (0.5,1]$, то *низкий риск.*

Классификация предприятий по двум группам уровней риска банкротства *высокий риск* и *низкий риск* соответствует решению финансового учреждения дать кредит или отказать в кредите.

Щелкнем по кнопке *Анализ*, в выпадающем меню выберем процедуру *Углубленные методы анализа*, во вновь появившемся меню − процедуру *Нелинейное оценивание* (рис.1). В появившемся диалоге *Нелинейное оценивание* щелкнем по процедуре *Логит регрессия* (рис.2).

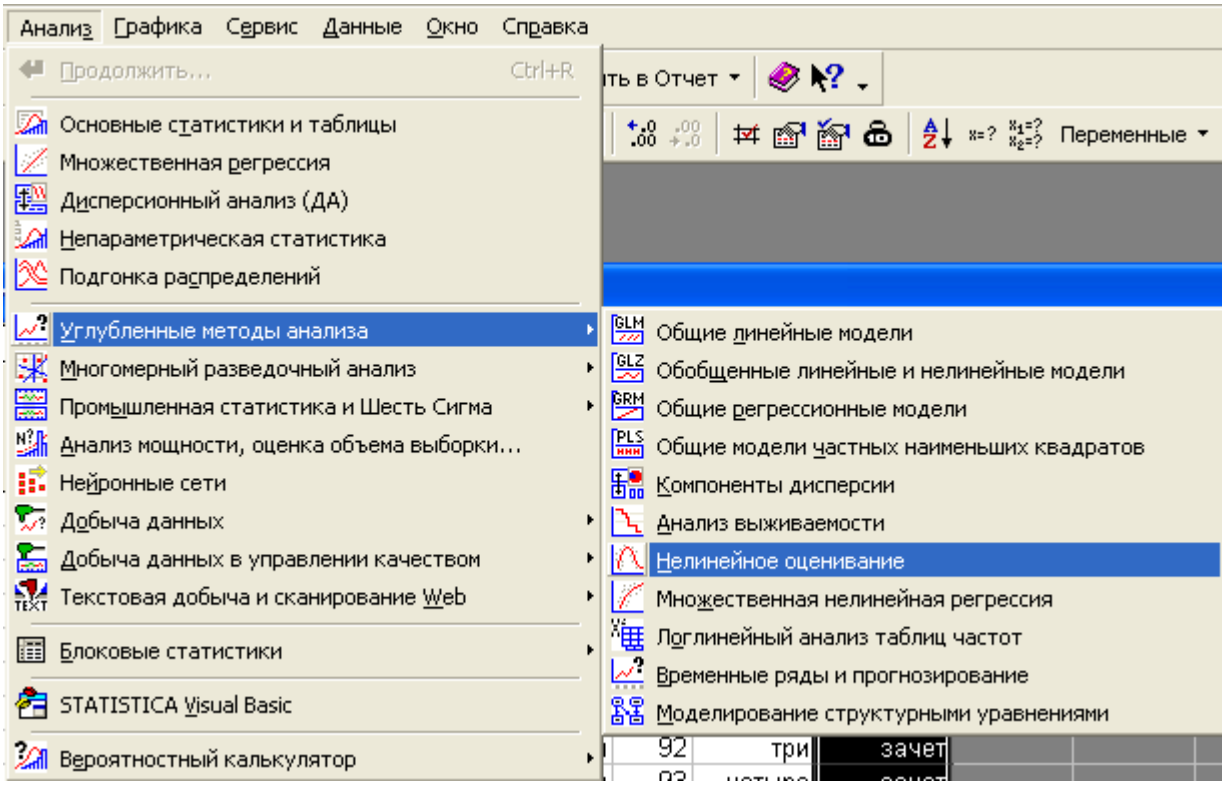

Рис.1

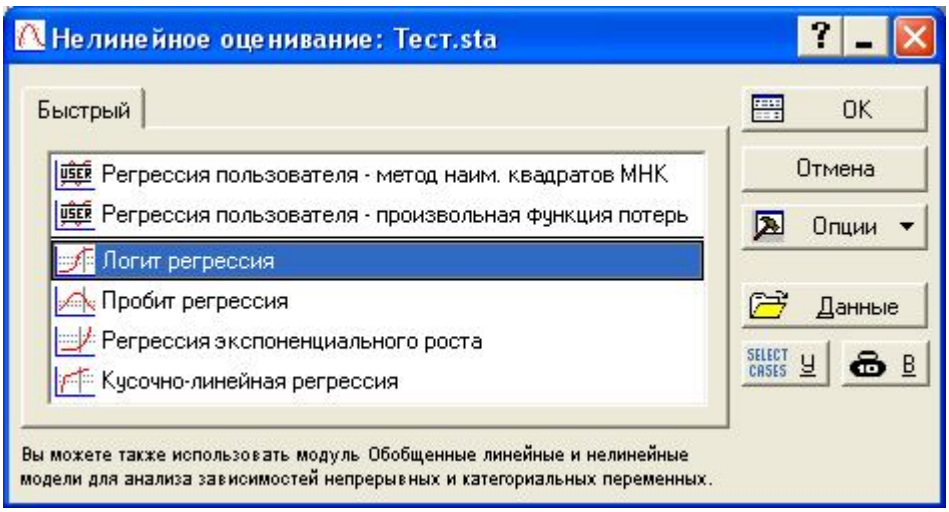

Рис.2

В стартовом диалоге (рис.3) укажем зависимую и независимые переменные и щелкнем по OK. Откроется окно *Оценивание модели*, в котором можно произвести необходимые установки для запуска процедуры оценивания. Из предлагаемых в модуле методов оценивания выберем метод *Розенброка и квази - ньютоновский* (рис.4),

остальные установки оставим без изменения. Если нажать на OK, появится диалог *Результаты* (рис.5).

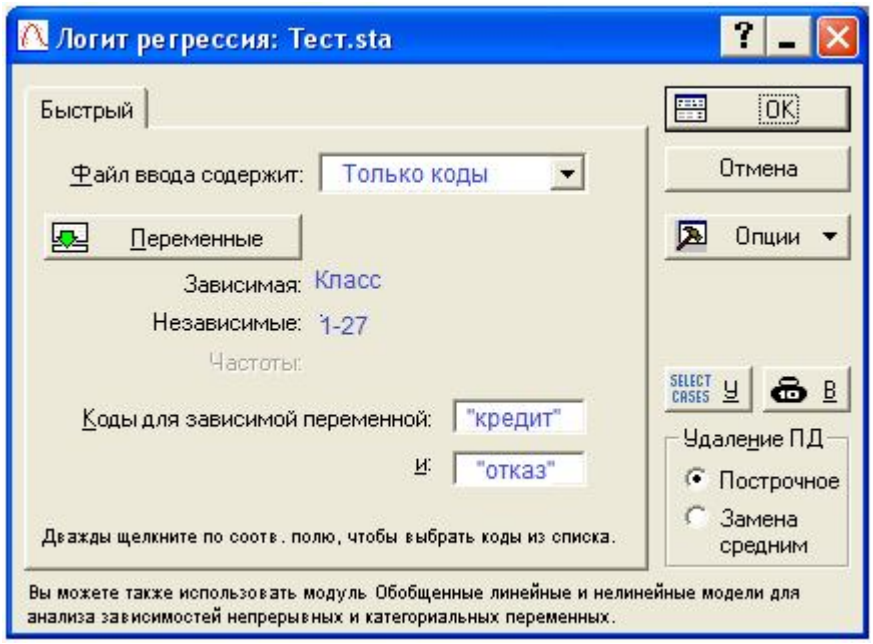

Рис.3

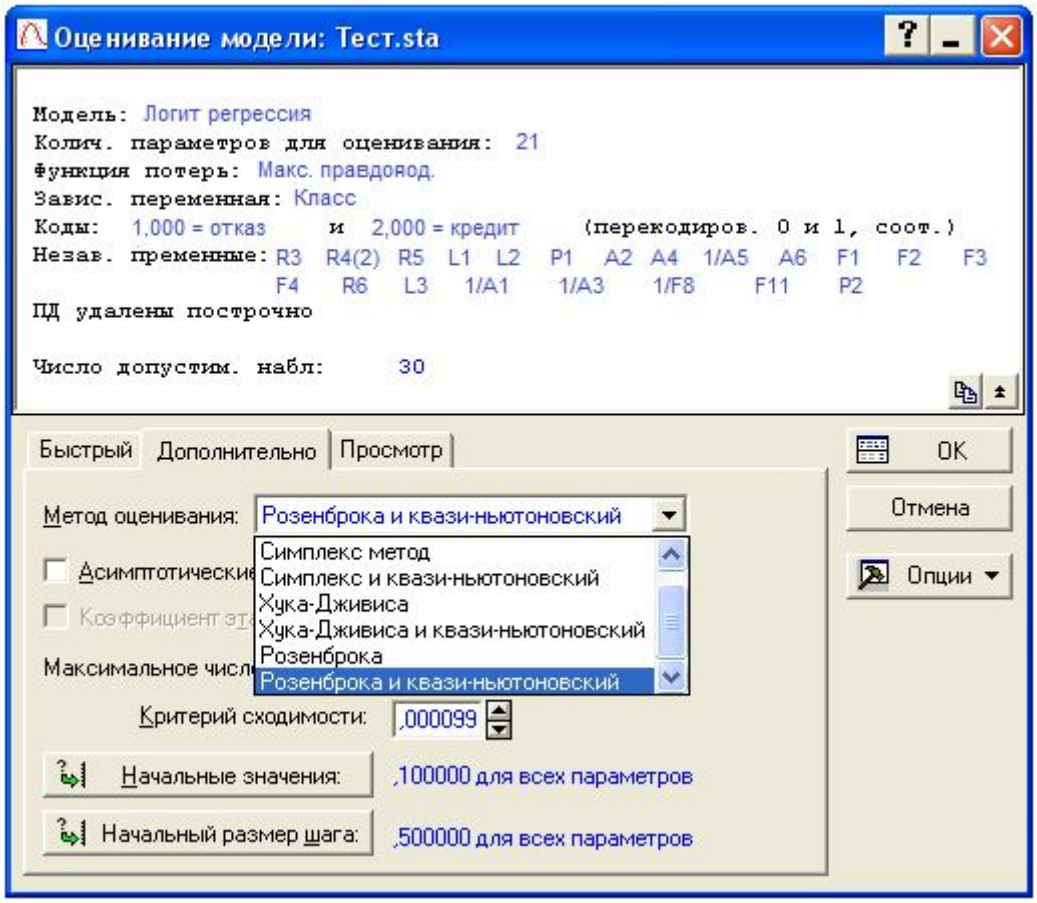

Рис.4

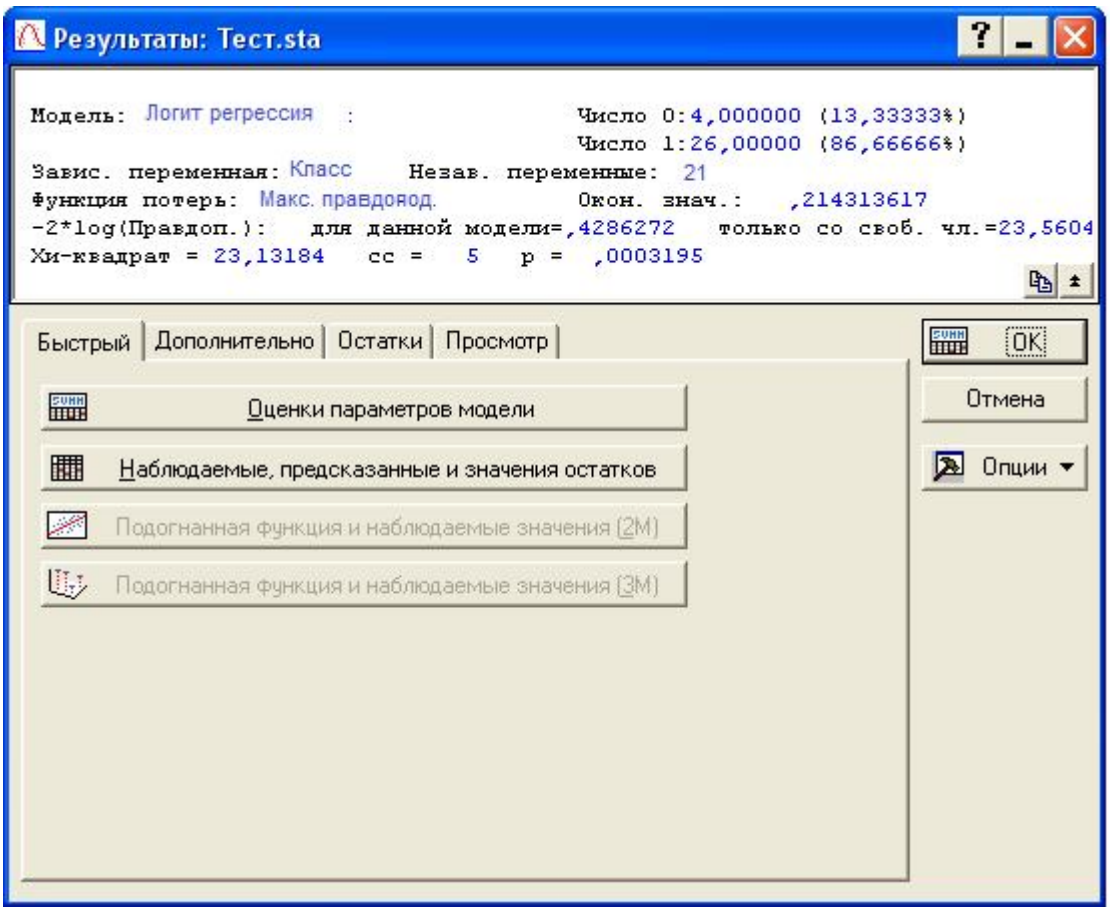

Рис.5

Адекватность построенной модели оценивали по значениям статистик «Итоговые потери», уровень значимости p критерия χ^2. Чем меньше значения статистик, тем модель адекватнее, причем p должно быть меньше, чем 0,05. Вычисленные коэффициенты параметров модели приведены в таблице 1 (переход по кнопке *Оценки параметров модели*). В ней также отображены названия показателей, которые включены в модель. В модель включены показатели для которых уровень значимости p критерия Стьюдента не больше 0,5. Показатель в модели считается статистически значимым, если p меньше, чем 0,05.

Таблица 1 – Таблица значений коэффициентов логит модели

	B0	R3	R4	R5	L1	L2
Оценка	-5,65827	1,291	9,8	0,55090	-2,82475	11
Отн.Шансов (ед.изм)	0,00349	3,635	17440,6	1,73481	0,05932	74422
Отн.Шансов (размах)		1341,960	106919,0	15,21647	0,00057	275925700
	P1	A2	A4	1/A5	A6	
Оценка	1,186	-1,77569	1,641499	0,679449	1	
Отн.Шансов	3,275	0,16937	5,162903	1,972789	3	

(ед.изм)						
Отн.Шансов (размах)	3233,755	0,00000		5,588205	178088600	
	F1	F2	F3	F4	R6	L3
Оценка	-0,054929	-2,25009	2,85586	3,365	0,3246	-1,26420
Отн.Шансов (ед.изм)	0,946553	0,10539	17,38941	28,928	1,3835	0,28247
Отн.Шансов (размах)	0,548683	0,00585		2631,071	985,3386	0,04958
	1/A1	1/A3	1/F8	F11	P2	
Оценка	-117,151	-7,61051	2,10	0,170077	2,2	
Отн.Шансов (ед.изм)	0,000	0,00050	8,14	1,185396	8,8	
Отн.Шансов (размах)	0,000	0,00000	69586,52	2,016432	127560,2	

Адекватность модели также определяется количеством ошибок классификации наблюдений, осуществленной в соответствии с построенной моделью. В таблице 2 в строках указана исходная классификация наблюдений (предприятий), в столбцах – предсказанная по модели. Так 3604 предприятий со значением уровня риска банкротства кредит предсказаны как кредит и 7 – ошибочно предсказаны как отказ. В тоже время 31 предприятие со значениями уровня риска банкротства отказ ошибочно предсказаны моделью как кредит и 2270 – правильно предсказаны как отказ.

Проценты безошибочных классификаций составили соответственно 98,6 и 99,8%. Общий процент верно классифицированных моделью предприятий равен 99,36%. Если эти проценты считать статистическими вероятностями, то вероятность, что предприятие вернет заемные денежные средства при его классификации в группу риска банкротства «кредит» равна 0,986. А вероятность возврата заемных денежных средств при классификации предприятия в группу «отказ» равна (1-0,998), т.е. 0,002. Общая вероятность правильного прогноза равна – 0,9936.

Таблица 2 – Таблица количества ошибок классификации наблюдений

Наблюдаемые	Предсказанные отказ	Предсказанные кредит	% правильной классификации
отказ	2270	31	98,65276
кредит	7	3604	99,80614

Таким образом, получили логит регрессионную модель, заданную уравнением:

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

где $Z = -5{,}658267+1{,}290519\cdot R_3 +9{,}766556\cdot R_4 + 0{,}5508959\cdot R_5 -2{,}824749\cdot L_1 + 11{,}21751\cdot L_2 + 1{,}186363\cdot P_1 -1{,}775693\cdot A_2 + 1{,}641499\cdot A_4 + 0{,}6794485\cdot 1/A_5 + 0{,}9686623\cdot A_6 - 0{,}05492874\cdot F_1 -2{,}250089\cdot F_2 +2{,}855861\cdot F_3 + 3{,}364808\cdot F_4 + 0{,}3246109\cdot R_6 - 1{,}264199\cdot L_3 - 117{,}1512 \cdot 1/A_1 - 7{,}610508\cdot 1/A_3 + 2{,}096704\cdot 1/F_8 + 0{,}1700767\cdot F_{11} + 2{,}179195\cdot P_2$.

Чтобы по значениям предикторов сделать прогноз бинарного отклика, надо подставить в уравнение числовые значения финансовых показателей предприятия и вычислить *Y*. Округлив полученное значение до целого (естественно это будет либо 0 (отказ), либо 1(кредит)), получим прогнозное значение отклика.

Построенная логит регрессионная модель вполне адекватная, так как итоговые потери, оцененные функцией максимального правдоподобия, малая величина (0,21) и уровень значимости *p* критерия *Хи-квадрат* значительно меньше 0,05 (p = 0,00032). Убедиться в адекватности модели можно также при помощи вкладки *Остатки* диалога *Результаты* (рис.6).

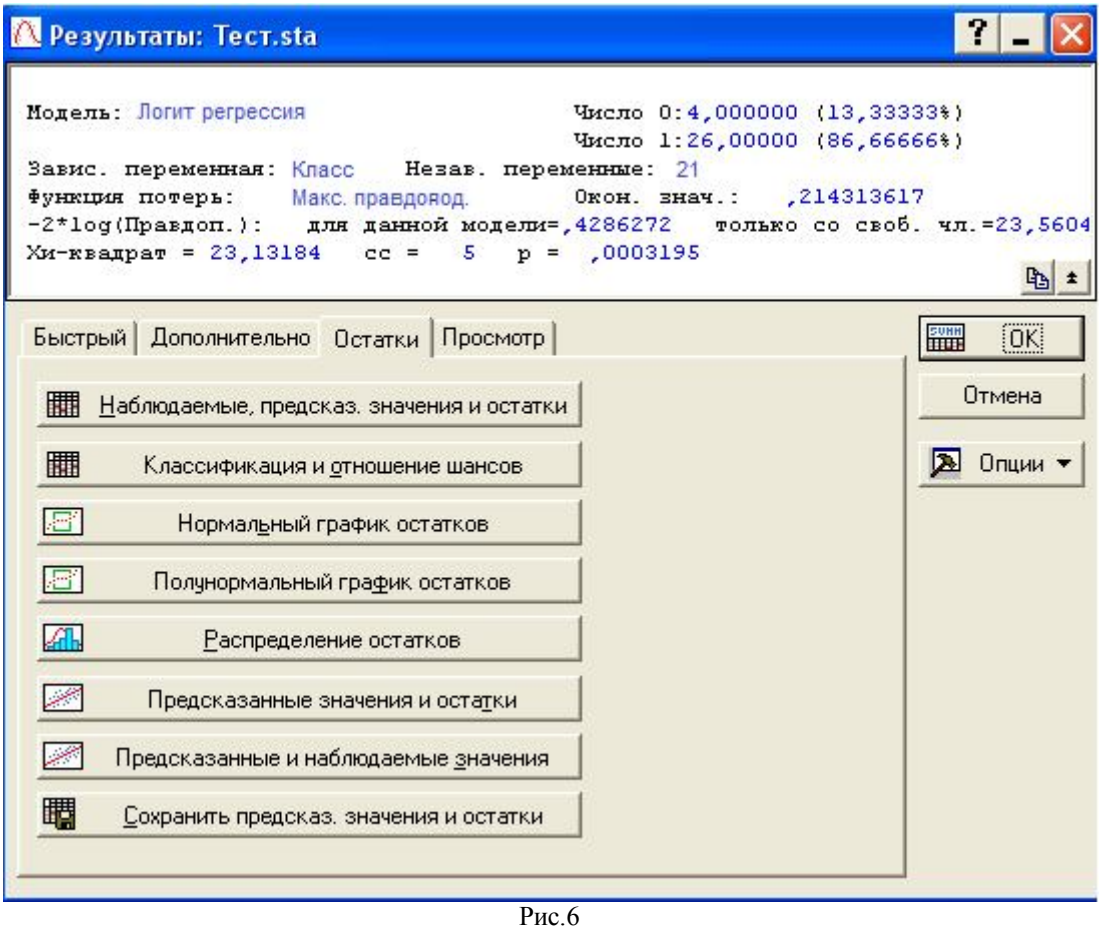

Рис.6

Если нажать на кнопку *Наблюдаемые, предсказанные значения и остатки*, появится таблица в которой указаны исходные значения отклика (*Наблюд*), прогнозные значения (*Предсказанные*) вычисленные посредством логит регрессионной модели, *Остатки* − разница между исходными и прогнозными значениями отклика (рис.7). Если внимательно просмотреть таблицу, легко заметить, что все предсказанные значения после округления совпадают с исходными значениями отклика.

Если воспользоваться кнопкой *Классификация и отношение* шансов, программа представит таблицу, в которой будет указано число верно и неверно классифицированных программой наблюдений (предприятий) по классам – *кредит, отказ*.

	Модель: (Тест.sta) Зав. Пер. : Зачет		
	Наблюд.	Предсказанные	Остатки
1	0,000000	0,000005	-0,000005
2	0,000000	0,029660	-0,029660
3	1,000000	0,899895	0,100105
4	1,000000	1,000000	0,000000
5	1,000000	1,000000	0,000000
6	1,000000	1,000000	0,000000
7	0,000000	0,000627	-0,000627
8	1,000000	1,000000	0,000000
9	1,000000	1,000000	0,000000
10	1,000000	0,984855	0,015145
11	1,000000	1,000000	0,000000
12	1,000000	1,000000	0,000000
13	1,000000	1,000000	0,000000
14	0,000000	0,060900	-0,060900
15	1,000000	1,000000	0,000000
16	1,000000	1,000000	0,000000
17	1,000000	1,000000	0,000000
18	1,000000	0,999999	0,000001
19	1,000000	0,999999	0,000001
20	1,000000	1,000000	0,000000
21	1,000000	1,000000	0,000000
22	1,000000	1,000000	0,000000
23	1,000000	1,000000	0,000000
24	1,000000	1,000000	0,000000
25	1,000000	1,000000	0,000000
26	1,000000	1,000000	0,000000
27	1,000000	1,000000	0,000000
28	1,000000	1,000000	0,000000
29	1,000000	1,000000	0,000000
30	1,000000	1,000000	0,000000

Рис.7

К сожалению, в модуле не предусмотрена кнопка, позволяющая пользователю вычислить функцию отклика для произвольного наблюдения по значениям предикторов.

Еще одним свидетельством адекватности модели, является гистограмма распределения остатков (рис.8), из которой следует, что мы не совершим большой ошибки, считая остатки белым шумом.

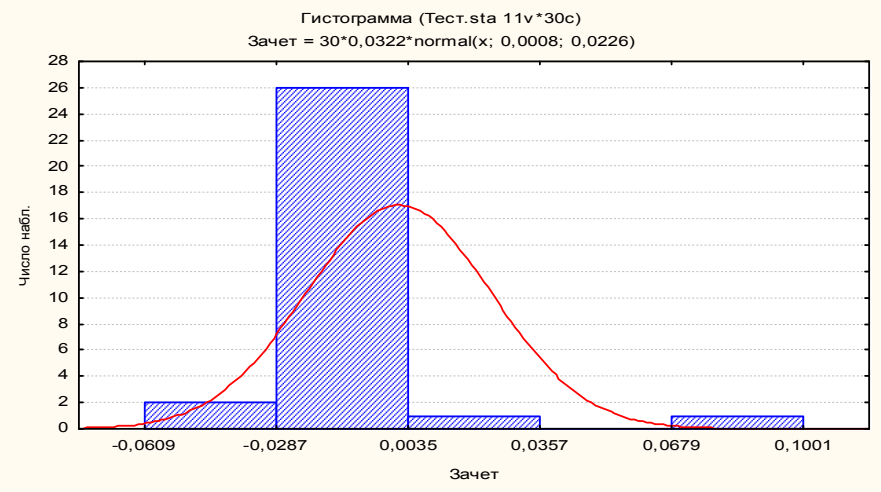

Рис.8