

Программа регрессионного анализа «Regresia»

**Выполнили: Р. Бондарь,
М. Гончаров,
В. Никитин**

Рига 2001

Ведение

Способность компьютеров перерабатывать большие массивы информации открывают широкие возможности для применения сложных статистических методов прогнозирования, в частности, регрессионного анализа. Используя подобные методы регрессии, можно осуществить анализ и прогноз, например, объема продаж в зависимости от таких переменных, как число новых семей, уровень доходов или размеры процентных ставок на бирже. Оценки такого рода используются, например, для исчисления ликвидности, для формулирования долгосрочной финансовой политики и для предсказания тенденций изменения курсов акций.

На фоне всеобщей заинтересованности в повышении качества прогнозов особенно заметным становится разрыв между часто используемыми обычными методами прогнозирования и более эффективными и точными статистическими методами, возможности применения которых открываются благодаря творческому подходу к использованию компьютеров. Благодаря статистическому подходу, использующему математический анализ тенденций и взаимосвязей, качество прогнозов может быть существенно улучшено.

1. Основные положения регрессионного анализа

Одним из методов прогнозирования с использованием компьютеров, получивших широкое распространение является регрессионный анализ, представляющий собой попытку объективно определить степень движения во времени одной переменной величины (например, "сбыта" или "прибылей") по отношению к другим переменным (таким, как "доходы", "численность населения", "новое строительство" и т.п.).

Главным достоинством этого метода является точность измерения статистических соотношений и определения их надежности. Кроме того, регрессионный анализ позволяет привлекать значительно большее количество данных, чем это возможно при использовании любых интуитивных или немашинных методов.

Хотя регрессионный анализ пока еще остается за пределами обычной практики прогнозирования в корпорациях, многие крупные компании в последние годы все шире используют этот метод. Как правило, это компании, использующие компьютеры в качестве одного из инструментов управления, а не просто как орудия бухгалтерского учета.

Линейный регрессионный анализ изучает такие случайные эксперименты, на исход которых влияют некоторые неслучайные переменные z_1, z_2, \dots, z_k , называемые факторами. Значения факторов меняются от эксперимента к эксперименту. Обозначим значение j -го фактора в i -ом эксперименте как z_{ij} (i изменяется от 1 до N , а j от 1 до k). Случайную величину, являющуюся результатом i -го опыта, можно представить в виде:

$$X_i = \sum_{j=1}^k z_{ij} b_j + \varepsilon_i,$$

где $\mathbf{b} = (b_1, \dots, b_k)^T$ – вектор коэффициентов регрессии, а

$\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_N)^T$ – вектор ошибок.

Относительно ошибок предполагают, что они не коррелированы и имеют одинаковые дисперсии

$$D[\varepsilon_i] = \sigma^2, \quad i = 1, \dots, N,$$

где σ^2 называют остаточной дисперсией.

В матричном виде модель линейной регрессии представляют в виде:

$$\mathbf{X} = \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad M[\boldsymbol{\varepsilon}] = 0, \quad D[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2 \mathbf{E}_N,$$

где \mathbf{X} – вектор значений наблюдаемых величин,

$\mathbf{Z} = (z_{ij})$ – матрица плана размером $N \times k$,

\mathbf{E}_N – единичная матрица размерности N .

Общим методом оценивания параметров регрессии (вектора \mathbf{b}) в линейной модели вида

$$\mathbf{X} = \mathbf{Z}\mathbf{b} + \boldsymbol{\varepsilon}, \quad M[\boldsymbol{\varepsilon}] = 0, \quad D[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2 \mathbf{E}_N$$

является **метод наименьших квадратов**. Если имеются результаты N экспериментов и число оцениваемых параметров равно k , размерность векторов \mathbf{X} и $\boldsymbol{\varepsilon}$ равна N , размерность матрицы $\mathbf{Z} - N \times k$.

В соответствии с методом наименьших квадратов оценки параметров регрессии находят из условия обращения в минимум квадратичной формы

$$\Psi(\mathbf{b}) = (\mathbf{X} - \mathbf{Z}\mathbf{b})^T (\mathbf{X} - \mathbf{Z}\mathbf{b}),$$

которая представляет собой сумму квадратов разностей между наблюдаемыми значениями и их математическими ожиданиями. Вектор \mathbf{b} , на котором достигается минимум квадратичной формы, называют оценкой **метода наименьших квадратов**. Для нахождения вектора оценок коэффициентов регрессии $\bar{\mathbf{b}}$ ищут решение нормальных уравнений:

$$\mathbf{A}\bar{\mathbf{b}} = \mathbf{Y}, \quad \mathbf{A} = \mathbf{Z}^T \mathbf{Z}, \quad \mathbf{Y} = \mathbf{Z}^T \mathbf{X}.$$

Справедлива следующая теорема.

Теорема. Если матрица \mathbf{A} не вырождена, оценка метода наименьших квадратов единственна и определяется равенством:

$$\bar{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

Если нормальное уравнение имеет несколько решений, то на каждом из них достигается минимум квадратичной формы $\Psi(\mathbf{b})$.

Если матрица \mathbf{A} не вырождена, ковариационная матрица оценок определяется равенством:

$$D[\bar{\mathbf{b}}] = \sigma^2 \mathbf{A}^{-1}.$$

Для оценки остаточной дисперсии σ^2 используют соотношение:

$$\sigma^2 = (\mathbf{Y} - \hat{\mathbf{Y}})^T (\mathbf{Y} - \hat{\mathbf{Y}}).$$

2. Внутренняя функциональность программы REGRESIA

Программа реализована на языке высокого уровня Delphi 5:

Внутренняя структура программы – взаимосвязь нескольких компонентов, в каждом из которых реализованы наборы функций отвечающий индивидуальным качествам объекта.

Компонент работы с матрицами Tmatrix, в нем реализованы функции необходимые для работы с матрицами.

Таблица 1.

Функции класса Tmatrix, доступных из вне класса.

Тип	Наименование	Описание.
constructor	Create(row,col : integer);	Создание матрицы row x col
destructor	Destroy; override;	Уничтожение матрицы.
procedure	revers(matrix: Tmatrix);	Получение обратной матрицы.
procedure	copy(matrix: Tmatrix);	Создание копии матрицы.
procedure	T(result : Tmatrix);	Транспонирование матрицы.
procedure	multi(mtx, result : Tmatrix);	Умножение матрицы на матрицу.
procedure	multiconst(c : Extended);	Умножение матрицы на константу.
procedure	mathadd(result: Tmatrix);	Получение математического дополнения.
procedure	mathaddiction(result: Tmatrix;row,col : integer);	Получение математического дополнения к ячейке row x col.
function	tostring : string;	Получение текстового описание матрицы.
property	cell[x,y : integer] : Extended	Получение значения ячейки [x,y]
property	col : integer;	Получение количества колонок в матрице.
property	row : integer;	Получение количества рядов в матрице.
function	det() : Extended;	Вычисление детерминанта матрицы.

Компонент Tregres вычислительные способности регрессионного анализа.

Таблица 2.

Переменные класса Tregres, доступных извне.

Наименование	Тип	Описание
x	Tmatrix	Матрица X:
Y	Tmatrix	Матрица Y:
y_pre	Tmatrix	Матрица ожидаемых значений
x_new	Tmatrix	Матрица новых данных
y_predicted	Extended	Предсказываемое значение зависимой переменной
B	Tmatrix	Вектор коэффициентов МНК
S	Tmatrix	Вектор коэффициентов Стьюдента
N	integer	Количество степеней свободы
K	integer	Количество степеней свободы
Ysr	Extended	Среднее значение зависимой переменной
Ssr	Extended	Сумма остатков обусловленная регрессией
Sse	Extended	Сумма остатков не обусловленная регрессией
Radj	Extended	Расчетные значения зависимой переменной
sigma	Extended	Среднеквадратическое отклонение
MSSr	Extended	Средняя сумма квадратов обусловленных регрессией
MSSost	Extended	Средняя сумма квадратов остатков
MSSe	Extended	Среднеквадратическая ошибка
F	Extended	Коэффициент Фишера

Таблица 3.

Функции класса Tregres, доступных из вне класса.

Тип	Наименование	Описание.
constructor	Create	Конструктор объекта.
destructor	Destroy; override;	Деструктор объекта.
procedure	count(x1,y1 : Tmatrix);	Количество элементов.
function	pridict(x1 : Tmatrix) : Extended;	Функция предсказания.
property	onlog : TlogEvent;	Логирование внутренней информации.

3. Описание интерфейса REGRESIA

Данный программный продукт предназначен для анализа данных, эконометрических и статистических расчетов. В программе реализованы методы и модели линейной регрессии.

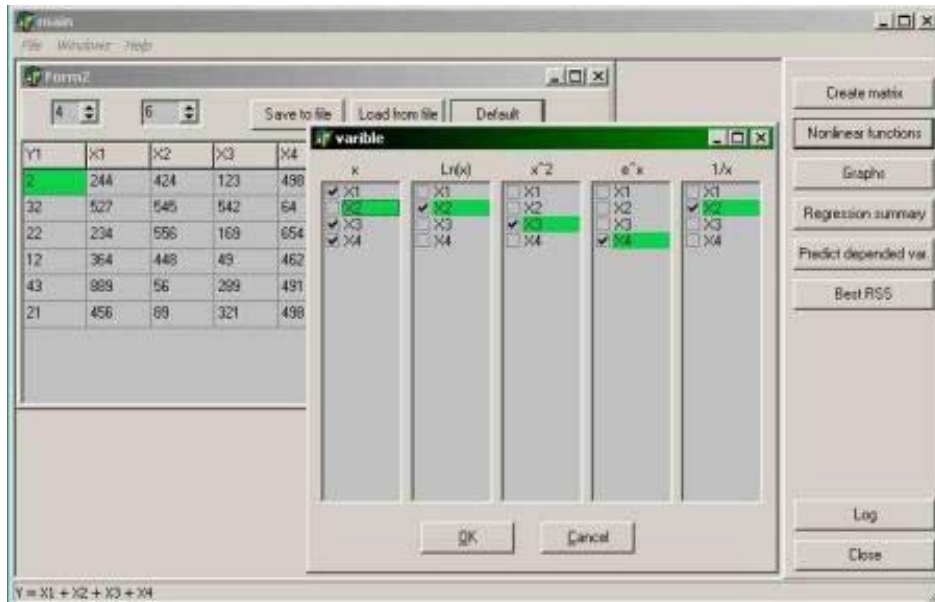


Рис. 1. Окно выбора переменных для динамического изменения модели

Также, программа содержит многие другие необходимые функции: описательная статистика (среднее, дисперсия и т.п.), различные графики (рис. 4, рис. 5), таблица корреляций переменных, быстрое построение графиков функций по формулам (рис. 1), импорт данных из текстовых файлов, история команд, сохранение построенной модели на жёстком диске компьютера.



Рис. 2. Выбор переменных для построения графиков

Результаты регрессионного анализа выводятся в табличной форме (рис. 3), что является простым и понятным, для пользователя, видом представления числовых данных.

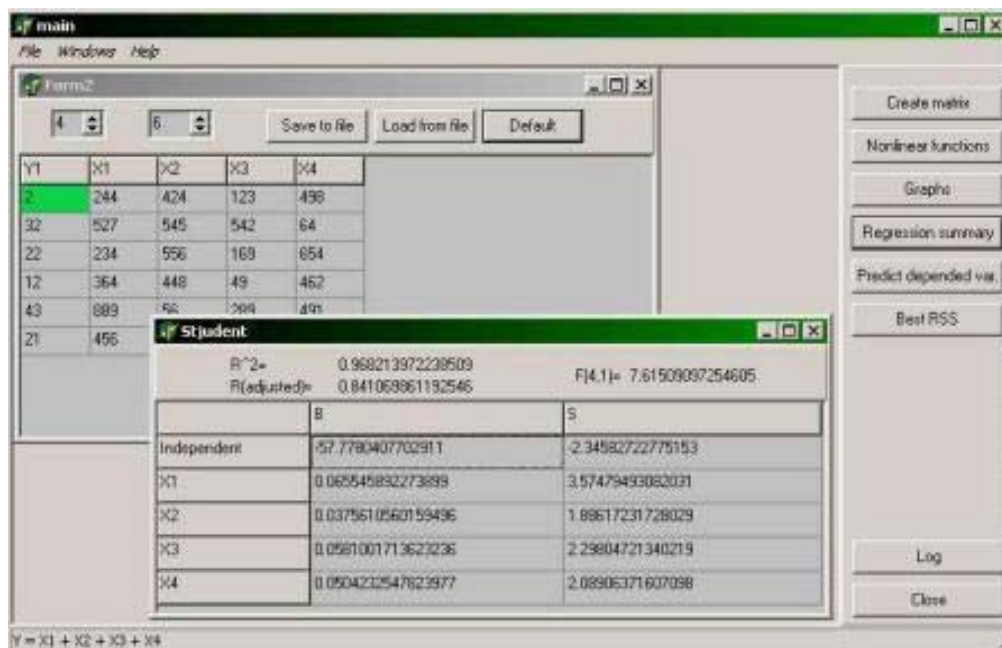


Рис. 3. Таблица результатов регрессионного анализа

Возможно комбинированное использование элементарных функций на независимые переменные. Расчеты сопровождаются графиками, что облегчает понимание их результатов. Это является достаточно важным, при условии сложных математических расчетов.

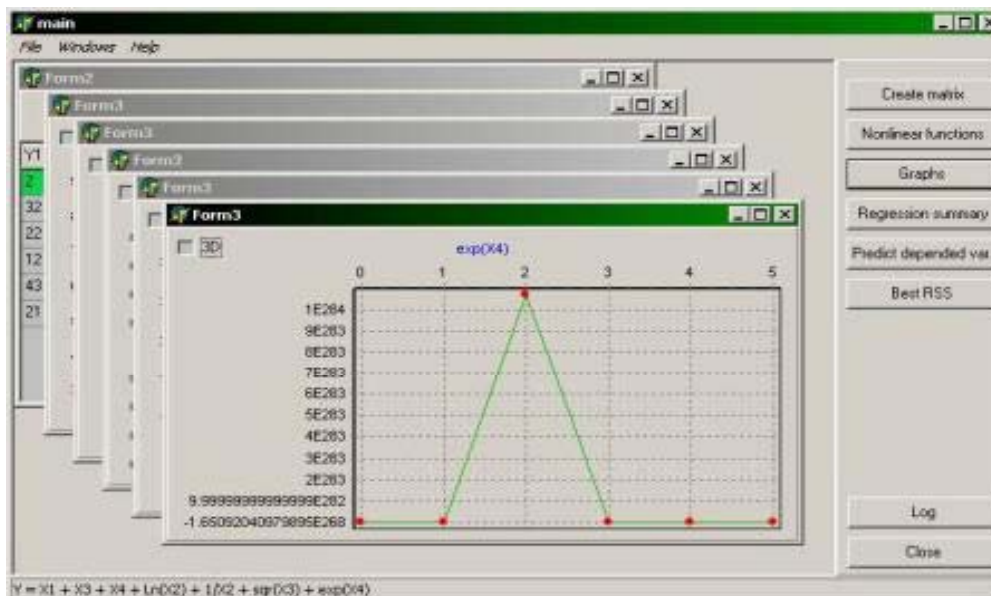
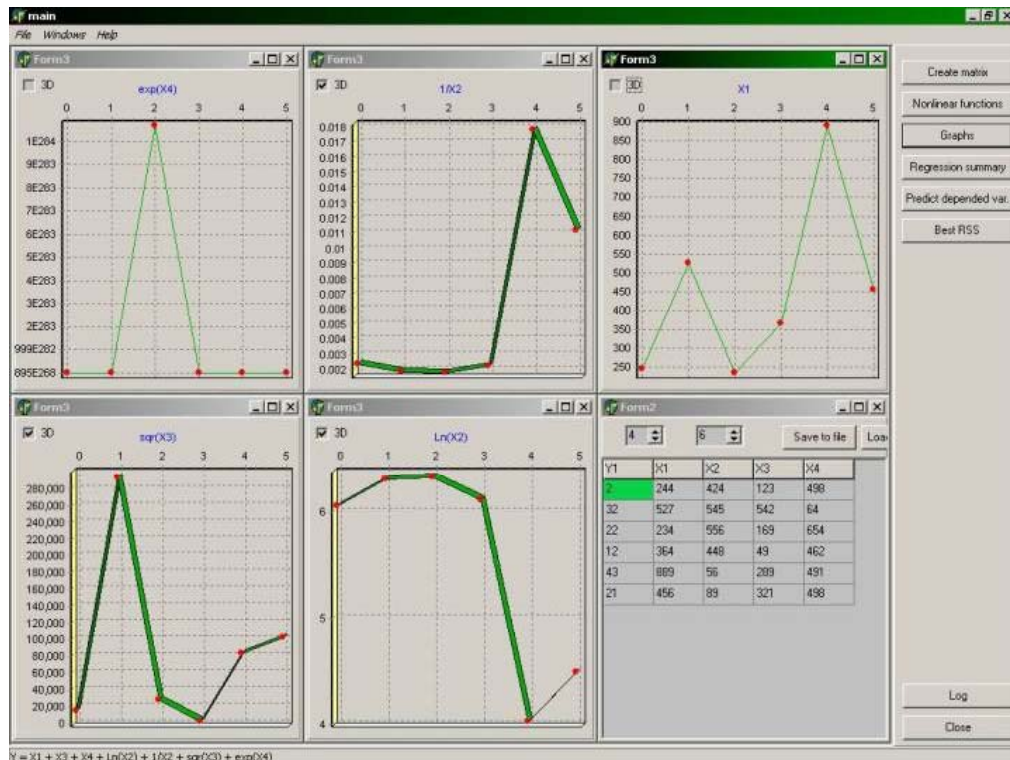


Рис. 4. Построение графиков регрессионной модели

Большое достоинство заключается в том, что версия *бесплатная*. Что играет не маловажную роль, учитывая стоимость программ такого класса.



Она

Рис. 5. Варианты графического представление данных

маленькая и умещается на одной дискете. Таким образом, по своим возможностям в расчете на 1 килобайт далеко превосходит другие аналогичные программы.

4. Заключение

Программа *проста в обращении* и эффективна. Многие операции, на которые в других пакетах уходит уйма времени, в *ней* осуществляются легким движением мыши или нажатием нескольких клавиш. Инсталляция производится обычным копированием файла в любую папку на вашем компьютере, что позволяет не беспокоиться о таких вещах как: переинсталляция программы в случае смены операционной системы или добавление информации в её системный реестр.

Также, программа постоянно модифицируется и совершенствуется, добавляются новые возможности и устраняются неточности.

В следующей версии будут добавлены следующие возможности:

- поддержка внутреннего скрипта, который предоставляет пользователю возможность изменять функциональные возможности программы, дополнять программу своими функциями и приспосабливать программу под необходимости данной задачи.
- Расширение визуальных возможностей представления данных (диаграммы, гистограммы и т.д.)
- Экспорт данных в различные форматы, для интеграции программы с другими приложениями пользователя.
- Автоматическая оптимизация регрессионной модели в соответствии с несколькими методами.
- Построение более качественного, с точки зрения пользователя, интерфейса.