УДК 517.18

РЕГРЕССИОННЫЙ АНАЛИЗ ДАННЫХ В СРЕДЕ WXMAXIMA

А. С. Шульгина, В. П. Колебошина, Б. А. Акишин

Донской государственный технический университет, Ростов-на-Дону, Российская Федерация
anneta.shulgina@mail.ru
vkoleboshina@bk.ru
akiboralex@mail.ru

Исследуются возможности и целесообразность использования некоммерческой системы компьютерной математики Maxima при построении эконометрических моделей и решении задач множественного регрессионного анализа. Отмечается наличие универсальных встроенных функций, позволяющих рассчитывать основные показатели моделей и реализующих оригинальные подходы в вопросах проверки значимости и адекватности. В то же время, открытость программного кода позволяет создавать собственные функции, работающие во взаимодействии с встроенными. В статье реализована классическая проверка существенности моделей по критерию Фишера.

Ключевые слова: система компьютерной математики, эконометрика, регрессия, метод наименьших квадратов, доверительный интервал, проверка статистических гипотез.

UDC 517.18

REGRESSION DATA ANALYSIS IN WXMAXIMA ENVIRONMENT

A.S. Shulgina, V.P. Koleboshina, B.A. Akishin

Don State Technical University, Rostov-on-Don, Russian Federation
anneta.shulgina@mail.ru
vkoleboshina@bk.ru
akiboralex@mail.ru

The authors study the opportunities and appropriateness of using noncommercial system of computer mathematics Maxima at creation of econometric models and solving problems of the multivariate regression analysis. The article notes the existence of the universal built-in functions allowing to count the main indicators of models and realizing original approaches in significance and adequacy checking. At the same time, openness of a program code allows to create eigenfunctions which work in interaction with the built-in ones. For example, classical check of models importance according to Fischer criterion was realized in the article.

Keywords: system of computer mathematics, econometrics, regression, method of least squares, confidence interval, statistical hypothesis testing

В процессе построения содержательных и адекватных эконометрических моделей регрессионного типа возникают, как правило, проблемы отбора факторов при структурной идентификации и плохой обусловленности обрабатываемых матриц при параметрической идентификации [1]. Требуется многократная целенаправленная верификация модели и пересчет ее коэффициентов, что возможно только при компьютерной обработке данных. Системы компьютерной математики, в частности, программа свободного доступа Maxima, содержат ряд встроенных функций, реализующих процедуры регрессионного анализа [3], однако их не всегда бывает достаточно, и требуется дополнительное программирование.

В настоящей работе реализован комбинированный метод расчета дополнительных параметров и статистических характеристик эконометрических моделей в программе Maxima, при

котором их программирование сочетается с использованием встроенных проблемно ориентированных пакетов.

При построении моделей линейной регрессии в Maxima обычно используют функции из пакета **stats**:

- **simple_linear_regression** (X, option) — для парной линейной регрессии,
- **linear_regression** (X, option) — для множественной линейной регрессии,
- **items_inference** — выдает список всех рассчитываемых характеристик,
- **take_inference** — выводит значения конкретных характеристик,

где X — матрица исходных данных, последний столбец которой представляет собой наблюдаемые значения зависимой переменной **y**, а единственной необязательной опцией является надежность статистических выводов (по умолчанию равная 0,95).

Функции пакета **stats** возвращают данные типа **inference_result**. Объекты этого типа, применительно к упомянутым функциям, содержат оценки коэффициентов уравнений регрессии, а также большое количество других рассчитанных параметров, необходимых для анализа статистической значимости полученных моделей и проверки гипотез. Часть этих параметров выводится на экран сразу (по умолчанию), а доступ к остальным обеспечивается при помощи функций **items_inference** и **take_inference**.

Рассмотрим два примера построения и анализа регрессионных моделей. Исходные данные для примеров взяты из пособия [1].

Пример 1. Необходимо оценить линейную зависимость средней заработной платы от месячного прожиточного минимума по данным из $n=12$ регионов РФ. Исходные данные приводятся в условных денежных единицах и представлены вложенным списком: первые элементы — значения x, вторые — значения y.

```
(%i2) load("stats")$

      Исходные данные примера 1

(%i3) n:12 $

(%i4) A:  [[78,133], [82,148], [87,134], [79,154], [89,162], [106,195],
           [67,139], [88,158], [73,152], [87,162], [76,159], [115,173]] $

(%i5) M1: simple_linear_regression(A);
                  SIMPLE LINEAR REGRESSION
                  model = 0.9204 x + 76.98
                    correlation = 0.721
                   v_estimation = 157.5
(%o5)           b_conf_int = [0.2972, 1.544]
                hypotheses = H0: b = 0 ,H1: b # 0
                     statistic = 3.291
                distribution = [student_t, 10]
                    p_value = 0.008142
```

Рис.1. Результаты моделирования линейной зависимости по данным примера 1

По умолчанию функция **simple_linear_regression** вывела на экран следующую информацию:

1. уравнение регрессии $\hat{y} = 76{,}98 + 0{,}9204 \cdot x$;
2. коэффициент линейной корреляции $= 0{,}721$ — достаточно высокий;
3. оценка остаточной дисперсии $= 157{,}5$;
4. доверительный интервал для коэффициента b;
5. проверяется нулевая $H_0 : b = 0$ и альтернативная $H_1 : b \neq 0$ гипотезы относительно значимости коэффициента b;
6. statistic= 3,291 значение статистики Стьюдента при 10 степенях свободы;
7. **p_value** — значение плотности вероятности распределения Стьюдента для проверки гипотезы о статистической значимости коэффициента b. В рассматриваемом случае $p_v = 0{,}008142$, что значительно меньше уровня значимости $\alpha = 0{,}05$, поэтому нулевая гипотеза H_0 отвергается и коэффициент b можно считать статистически значимым.

Отобразим результаты моделирования на графике.

```
(%i6) wxplot2d([[discrete,A], take_inference(model,M1)],
[x,60,130],[style,[points],[lines,2]],[color,blue,red],
[xlabel,"Прожиточный минимум"],[ylabel,"Заработная плата"])$
```

(%t6)

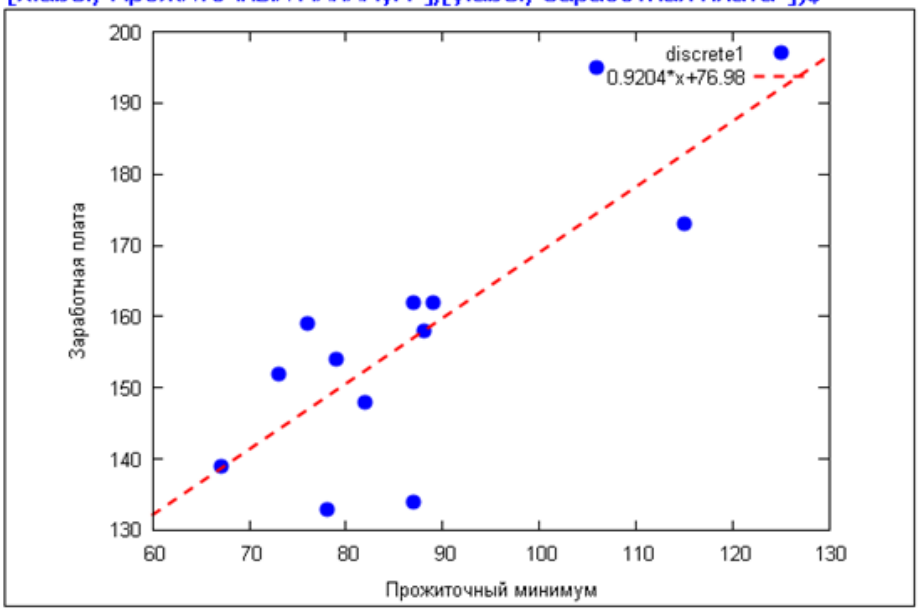

Рис.2. Графическое отображение результатов моделирования

Функция **items_inference** выдает полный перечень всех рассчитанных функцией **simple_linear_regression** параметров. Кроме отображенных по умолчанию, здесь присутствуют: средние значения, дисперсии, коэффициент детерминации, различные доверительные интервалы и остатки.

Рассчитываемые параметры

```
(%i7) items_inference(M1);
(%o7) [model,means,variances,correlation,adc,a_estimation,a_conf_int,b_estimation,
b_conf_int,hypotheses,statistic,distribution,p_value,v_estimation,v_conf_int,
cond_mean_conf_int,new_pred_conf_int,residuals]
```

Рис.3. Перечень рассчитываемых параметров линейного уравнения регрессии

Используя рассчитанный коэффициент детерминации **adc**, проверим адекватность модели по критерию Фишера и сделаем интервальный прогноз:

F-статистику вычислим по известной формуле через **adc**, а критическое значение Fkp- с помощью функции **quantile_f** (квантиль распределения Фишера) из пакета **distrib**.

```
        Коэффициент детерминации

(%i8) R2: take_inference('adc, M1);
(%o8) 0.4719

        F - статистика

(%i9) F: R2/(1-R2)*(n-2);
(%o9) 8.935

        F критическое

(%i10) load (distrib)$
       Fkp: quantile_f(0.95, 1, n-2);
(%o11) 4.965
```

Рис.4. Результаты проверки критерия Фишера

Поскольку $F > Fkp$, то регрессию можно считать значимой в смысле критерия Фишера.

Используем полученную модель для прогноза: увеличим средний прожиточный минимум на 10%, подставим это значение в уравнение регрессии (точечный прогноз средней заработной платы) и выведем доверительный интервал прогноза.

```
(%i12) x_y: take_inference('means, M1);
(%o12) [85.58,155.8]

        Прогноз

(%i13) x: x_y[1]*1.1;
(%o13) 94.14

(%i14) y: take_inference(model,M1), x=x;
(%o14) 163.6

(%i15)  take_inference(new_pred_conf_int, M1), x=x;
(%o15) [134.0,193.2]
```

Рис.5. Доверительный прогноз на линейной модели

Таким образом, полученная регрессионная математическая модель показывает: если увеличить средний месячный прожиточный минимум до 94,14 условных денежных единиц, то средняя заработная плата будет находиться в 95% доверительном интервале (134; 193,2) (его рассчитывает сама функция **simple_linear_regression)**.

Пример 2. Построить математическую модель для расчета объемов реализации одного из продуктов фирмы (зависимая переменная y). В качестве исходных объясняющих факторов выбраны: x_1 — время, x_2 — расходы на рекламу, x_3 — цена товара, x_4 — средняя цена у конкурентов, x_5 — индекс потребительских расходов.

Имеющиеся статистические данные подготовлены в текстовом файле. Подключаем пакет **numericalio** и загружаем данные с помощью функции **read_matrix** [2]:

```
(%i16) load(numericalio)$ N: 16 $

(%i18) data: read_matrix("F:\\example_2.txt");
```

(%o18)

Y	X1	X2	X3	X4	X5
126	1	4	15	17	100
137	2	4.8	14.8	17.3	98.4
148	3	3.8	15.2	16.8	101.2
191	4	8.7	15.5	16.2	103.5
274	5	8.2	15.5	16	104.1
370	6	9.7	15	18	107
432	7	14.7	18.1	20.2	107.4
445	8	18.7	13	15.8	108.5
367	9	19.8	15.8	18.2	108.3
367	10	10.6	16.9	16.8	109.2
321	11	8.6	16.3	17	110.1
307	12	6.5	16.1	18.3	110.7
331	13	12.6	15.4	15.4	110.3
345	14	6.5	15.7	16.2	111.8
364	15	5.8	16	17.7	112.3
384	16	5.7	15.1	16.2	112.9

Рис.6. Импорт исходных данных для примера 2

Анализ матрицы парных корреляций для примера 2 показывает, что в первом приближении можно остановиться на линейной регрессии, зависящей от двух факторов x_2 и x_5.

```
(%i19) load(descriptive)$

(%i20) X: submatrix (1,data) $

       Матрица парных корреляций

(%i21) cor(X), numer;
```

(%o21)

1.0	0.678	0.6459	0.2329	0.2263	0.816
0.678	1.0	0.1065	0.1737	-0.051	0.9602
0.6459	0.1065	1.0	-0.003354	0.204	0.2734
0.2329	0.1737	-0.003354	1.0	0.6978	0.2354
0.2263	-0.051	0.204	0.6978	1.0	0.03078
0.816	0.9602	0.2734	0.2354	0.03078	1.0

Рис.7. Рассчитанная матрица парных корреляций для примера 2

Выделяем список упомянутых переменных и запускаем функцию **linear_regression**. Объект **inference_result**, выдаваемый функцией **linear_regression** по умолчанию включает:

1. оценки коэффициентов b_estimation — уравнение имеет вид: $\hat{y} = -1471 + 9{,}568 \cdot x_2 + 15{,}75 \cdot x_5$
2. t – статистики Стьюдента для проверки значимости коэффициентов,
3. значения **b_p_values** для всех коэффициетов. Они значительно меньше уровня значимости $\alpha = 0{,}05$, поэтому все коэффициенты можно считать статистической значимыми,
4. остаточная дисперсия — случайная величина, имеющая распределение Хи-квадрат и ее доверительный интервал,
5. коэффициент детерминации **adc = 0,8374**

```
(%i22) X2X5Y:makelist([data[k,3], data[k,6], data[k,1]], k, 2, N+1) $

(%i23) res: linear_regression(X2X5Y);
                    LINEAR REGRESSION MODEL
       b_estimation = [ - 1.471 10^3, 9.568, 15.75 ]
       b_statistics = [ - 5.664, 4.223, 6.386 ]
       b_p_values = [ 7.746 10^-5, 9.965 10^-4, 2.396 10^-5 ]
(%o23) b_distribution = [ student_t, 13 ]
       v_estimation = 1.72 10^3
       v_conf_int = [ 904.0, 4.464 10^3 ]
       v_distribution = [ chi2, 13 ]
       adc = 0.8374
```

Рис.8. Результаты моделирования множественной регрессии

Дополнительно рассчитываются также следующие характеристики: ковариационная матрица, доверительные интервалы для всех коэффициентов, остатки, а также значения информационных критериев Байеса (**bic**) и Акаике (**aic**), используемых для сравнения различных моделей.

```
        Список всех параметров

(%i24) items_inference(res);
(%o24) [b_estimation,b_covariances,b_conf_int,b_statistics,b_p_values,
b_distribution,v_estimation,v_conf_int,v_distribution,residuals,adc,aic,bic]

(%i25) take_inference('bic, res);
(%o25) 124.2
```

Рис.9. Расчет критерия Байеса для модели множественной регрессии

Программа Maxima включает также мощный пакет **lsquares** для нелинейного оценивания параметров различных моделей с использованием метода наименьших квадратов [3]. Основная функция пакета **lsquares_estimates (*D, x, e, a*),** где **D** — матрица данных, **x** — имена переменных, **e** — задаваемое нелинейное уравнение, **a** — имена параметров.

Вывод. На примерах программной реализации классических критериев Фишера и Байеса показано, что открытость программного кода системы Maxima позволяет создавать собственные функции, работающие во взаимодействии с встроенными.

Библиографический список

1. Акишин, Б. А. Экономико-математические расчеты на персональном компьютере. Ч. 2. Эконометрические модели. Учебное пособие / Б. А. Акишин, А. В. Галабурдин. — Ростов-на-Дону : РАС ЮРГУЭС, 2008. — 60 с.

2. Решение математических задач с помощью пакета Maxima : Учеб. пособие / Б. А. Акишин [и др.]. — Ростов-на-Дону : Издательский центр ДГТУ, 2015. — 100 с.

3. Maxima 5.38.1 Manul / Документация по текущей версии пакета Maxima. — Режим доступа : http://maxima.sourceforge.net/docs/manual/en/maxima.html (Дата обращения 05.05.2016).